The role of geometry in translocation processes of and on nucleic acids

ribosomal frameshift and dynamic target search

Thomas Schötz

Munich, November 18th, 2010

The role of geometry in translocation processes of and on nucleic acids

ribosomal frameshift and dynamic target search

Thomas Schötz

Dissertation der Fakultät für Physik
der
Ludwig-Maximilians-Universität
München

vorgelegt von
Thomas Schötz
aus Straubing

München, 18. November 2010

1. Gutachter:	Prof. Dr. U. Gerland
2. Gutachter:	Prof. Dr. H. Stark
Tag der mündlichen Prüfung:	26. Januar 2011

"*Loppu hyvin, kaikki hyvin*"

(Finnish proverb)

Contents

List of figures

Zusammenfassung

In der vorliegenden Dissertation wird mit den Werkzeugen der statistischen Physik, insbesondere mit *coarse-graining* ("Vergröberungs-") Methoden, die biologische Fragestellung untersucht, welchen Einfluss die statischen und dynamischen Eigenschaften der räumlichen Struktur von Nukleinsäuren auf Translokationsprozesse ausüben. Dabei kann die Nukleinsäure entweder die Rolle des Translokations*objektes* übernehmen, wie bei der RNA-Translokation durch Ribosomen im Translationsprozess, oder die der Translokations*umgebung*, wie bei der Protein-Translokation bzw. -Zielsuche auf DNA zum Zwecke der Transkriptionsregulation. Hier werden beide Translokationstypen anhand der genannten Beispiele behandelt.

Der erste Teil dieser Dissertation hat daher Faltung und Translokation von RNA zum Thema, insbesondere die Art und Weise, wie bestimmte Tertiärstrukturmotive wie beispielsweise Pseudoknoten das Pausieren des RNA-Moleküls im Ribosom (*ribosomal pausing*) bewirken, das bei vielen RNA-Viren (z.B. HIV) für eine Verschiebung im Leseraster des Ribosoms (*ribosomal frameshift*) notwendig ist.

Ziel ist dabei, ein vergröbertes (*coarse-grained*) dreidimensionales RNA-Modell auf der Ebene einzelner Nukleotidbasen aufzustellen und mit seiner Hilfe den Mechanismus des Pausierens von RNA-Molekülen in Ribosomen über die geometrischen, mechanischen und energetischen Eigenschaften des Modellmoleküls zu klären. Das Modell soll dafür eine Vielzahl von Anforderungen erfüllen: Es soll damit möglich sein, sowohl spontane Faltung als auch kraftinduzierte Auffaltung im Detail zu studieren; die geometrischen (räumlich-helikale Moleküldomänen) und mechanischen Eigenschaften (Torsions- und Biegesteifigkeit der Helizes) der gefalteten Tertiärstruktur trotz fehlender innerer Struktur der Modellbasen korrekt nachzubilden; prinzipiell beliebig komplexe Tertiärstrukturen zu untersuchen, was hinreichend lokale intramolekulare Modellwechselwirkungen erfordert.

Die Aufstellung eines solchen Modells ist gelungen: Das RNA-Molekül wird dabei durch eine Kette über Federn miteinander verbundener Basen-Punktteilchen (*bead-spring*-Modell) repräsentiert, die Faltungseigenschaften durch basenpaar-spezifische Lennard-Jones-Wechselwirkungen realisiert. Helikale Moleküldomänen mit Torsions- und Biegesteifigkeit entstehen durch lokale Basenpaar-Basenpaar-Wechselwirkungen, die von der gegenseitigen Ausrichtung und der momentanen Bindeenergie der

beteiligten Basenpaare abhängen. Aufgrund der Komplexität von Geometrie und Potentialstruktur erwies sich Brownsche Dynamik als Simulationsmethode der Wahl.

Nach dem erfolgreichen Test von Faltungsdynamik sowie der mechanisch-geometrischen Tertiärstruktureigenschaften wurde mit Hilfe detailliert gemessener zeitabhängiger geometrischer und energetischer Größen der Prozess der Molekültranslokation durch eine enge Modellpore mit lokaler Drift in einzelne für die geometrischen Elemente der verwendeten Tertiärstruktur charakteristische Phasen unterteilt (z.B. Auftrennen, Drift, Rekombination). Anhand dieser Phasensequenzen wurden mehrere Translokationsregimes identifiziert (Rückwurf, Pausieren, einzelsträngige Translokation), je nach Porenenergie, Basenpaarenergie und Temperatur. Schlussendlich wurde im relevanten Translokationsregime die Ursache der im Vergleich zum Hairpin festgestellten stark verlangsamten Translokation des Pseudoknotens geklärt: Das Auftrennen von dessen Primärhelix wird durch den hinter der Sekundärhelix gelegenen wenig beweglichen Sequenzabschnitt behindert, der dafür nötige Kraftübertrag geschieht über die noch gefaltete Sekundärhelix, die es beim Hairpin nicht gibt.

Für die Effizienz der Genregulation ist eine möglichst schnelle Suche der Transkriptionsfaktoren nach ihren spezifischen Bindeorten auf der DNA essentiell. Aus diesem Grund wird im zweiten Teil der Dissertation die Translokation von Proteinen auf DNA behandelt. Es ist bekannt, dass Transport solcher Transkriptionsfaktoren auf DNA-Molekülen, der Teils über 1D-Diffusion erfolgt, Teils über Sprünge entlang konformationsabhängiger Sprungkanäle, die u.U. entfernte DNA-Segmente miteinander verbinden, bei asymptotisch schneller DNA-Dynamik ("*annealed*" Limes) superdiffusiv, ohne DNA-Dynamik ("*quenched*" Limes) jedoch paradoxerweise diffusiv ist.

In dieser Arbeit werden die postulierten Korrelationen, welche dieses diffusive Verhalten verursachen, und ihre statistischen Eigenschaften im Detail charakterisiert. Dabei wird mit Hilfe eines *coarse-grained* Modells insbesondere zum ersten Mal eingehend der Mechanismus beleuchtet, der zu quasi-diffusivem Transport auf eingefrorenen DNA-Molekülen führt. Desweiteren wird anhand explizit dynamischer DNA-Moleküle mittels einer Monte-Carlo-Simulation das schrittweise Aufbrechen der Korrelationen demonstriert, das sich in einem nichttrivialen Übergang von Quasi-Diffusion bei großen Zeiten zu Superdiffusion bei kleinen Zeiten äußert.

Es wird also gezeigt, dass das Transportverhalten der Transkriptionsfaktoren von der – gegebenenfalls dynamischen – Topologie des Intersegment-Verbindungsnetzwerks und damit wiederum von der räumlichen Gestalt der DNA-Kette bestimmt wird.

Abstract

In this dissertation the tools of statistical physics, especially coarse-graining methods, are used to examine the biological question what influence the statistical and dynamical properties of the spatial structure of nucleic acids exert on translocation processes. In such processes the nucleic acid can either take the role of the translocated object, as for RNA translocation through ribosomes during the process of translation, or the role of the translocation environment, as for protein translocation and target search on DNA for the purpose of the regulation of transcription. Here, both types of translocation are treated using the mentioned examples.

Therefore, the first part of this dissertation is about folding and translocation of RNA, especially the manner by which certain tertiary structure motifs – like for instance pseudoknots – cause the RNA-molecule to pause within the ribosome (*ribosomal pausing*), an effect needed by many RNA-viruses (HIV, e.g.) to shift the reading frame of the ribosome (*ribosomal frameshift*).

The goal is to create a coarse-grained three-dimensional RNA model on the level of single nucleotide bases and to use it as an instrument to clarify the mechanism of RNA-molecules pausing in ribosomes via the geometrical, mechanical and energetic properties of the model molecule. For this, the model should fulfill a multitude of demands: It should be able to be used to study both spontaneous folding and force-induced unfolding in detail; to emulate the geometrical (spatially-helical molecule domains) and mechanical properties (torsional and bending stiffness of the helices) of the folded tertiary structure correctly, although the model bases lack any inner structure; to study in principle arbitrarily complex tertiary structures, which calls for sufficiently local intra-molecular model interactions.

The modeling has been successful: The RNA molecule is represented by a chain of point-shaped bases inter-connected by springs (bead-spring-model), the folding properties are realized by base-pair-specific Lennard-Jones interactions. Helical molecular domains with torsional and bending stiffness emerge because of local base-pair base-pair interactions, which depend on the directions and the current binding energies of the participating base-pairs. Because of the complexity of the model in terms of geometry and interactions Brownian dynamics was chosen as simulation method.

After having tested the folding dynamics and the mechanical and geometrical properties of the tertiary structure successfully, the process of translocating a molecule through a narrow model pore with local drift was subdivided into single phases (e. g. unzipping, drift, recombination) characteristic for the geometrical elements of the tertiary structure in use, by the means of time-dependent geometrical and energetic variables measured in detail. Using these phase sequences, according to the values of pore energy, base-pair energy and temperature several translocation regimes were identified (rejection, pausing, single-stranded translocation). Finally, the cause for the extremely decelerated observed translocation of the pseudoknot – if compared to the hairpin – was clarified in the relevant translocation regime: The unzipping of the primary helix of the pseudoknot is interfered by the hardly mobile sequence part behind its secondary helix; the still folded secondary helix, which does not exist for hairpins, is able to perform the necessary force transfer.

To make gene regulation efficient it is essential that the target search of transcription factors for their specific binding sites on DNA is as fast as possible. Therefore, the second part of the dissertation is about translocation of proteins on DNA. As is known, if transport of such transcription factors on DNA molecules is partially done by 1D-diffusion and partially by jumps along conformation-dependent jump channels connecting DNA segments possibly distant, it is superdiffusive for asymptotically fast DNA dynamics (“annealed” limit) but paradoxically diffusive if there is no DNA dynamics at all (“quenched” limit).

In this thesis the postulated correlations causing this diffusive behavior and their statistical properties are characterized in detail. In doing so, especially the mechanism leading to quasi-diffusive transport on frozen DNA molecules is treated thoroughly and for the first time. This is done by the means of a coarse-grained model. Furthermore, a Monte-Carlo simulation for explicitly dynamical DNA molecules is used to demonstrate the gradual destruction of the correlations, which manifests in a non-trivial cross-over from quasi-diffusion at long times to superdiffusion at short times.

Therefore, it is shown that the transport behavior of the transcription factors is determined by the – possibly dynamical – topology of the intersegment linking pattern and thus in turn by the spatial configuration of the DNA chain.

1 Introduction

1.1 Translocation processes occurring within the central dogma

The central dogma of molecular biology as introduced by Crick [1] describes the transfer of genetic information from DNA to RNA and from RNA to proteins (Fig. 1.1). Translocation of and on such biopolymers (nucleic acids and proteins) is ubiquitous in the processes of the central dogma.

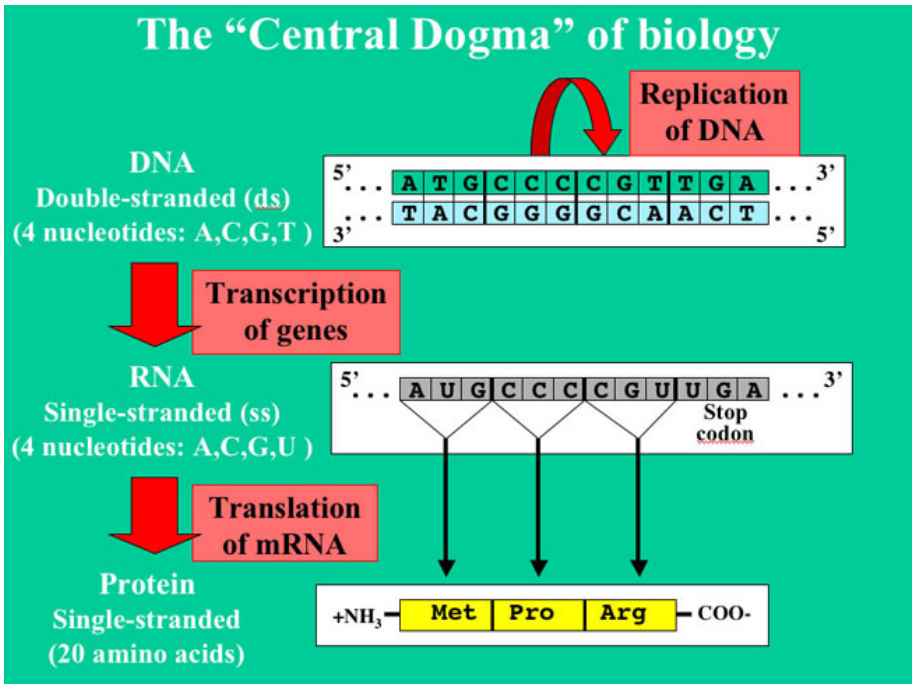


Fig. 1.1: The Central Dogma describes the transfer of genetic information from DNA to RNA (transcription) and from RNA to proteins (translation). The sequence of nucleotides in the nucleic acid (DNA and RNA) determines the sequence of amino acids in the protein produced.

(Graphics by George Rice, Montana State University [2])

For example, the process of transcription, which denotes the synthesis of a particular messenger RNA molecule by copying a part (operon) of the original DNA sequence with the help of DNA polymerase, is regulated by a certain type of proteins, the transcription

factors. These regulatory proteins bind to specific sites (operators) on the DNA contour and stabilize or block the binding of DNA polymerase near these sites. This way, transcription of the structure genes the DNA operon behind such a site consists of is either facilitated or suppressed. According to the effect they have on transcription, stabilizing proteins are called activators, blocking proteins repressors. In order to find their specific binding sites, the transcription factors have to translocate along the contour of the double-stranded DNA in a target search process.

The process of translation, during which the single-stranded messenger RNA molecule produced by transcribing a DNA operon unfolds and traverses the narrow channel of a ribosome, is another example for translocation in the central dogma. During translation, the ribosome reads the information encoded in the RNA molecule codon by codon and simultaneously concatenates the corresponding amino acid pendants to form the final protein according to the blue-print on the RNA.

In both types of translocation occurring in the central dogma – translocation of proteins on DNA for the regulation of the transcription process and translocation of messenger RNA through ribosomes during the process of translation – the translocation time is (within certain limits) determined by the geometry and the topology of the corresponding nucleic acid. This is true, although for protein translocation the nucleic acid (DNA) is the translocation environment and for RNA translocation the nucleic acid (RNA) is the translocated object.

The geometric determinant for the transport properties in either case are intricate patterns of intra-molecular connections within the chain-shaped nucleic acid molecules: For entropic reasons DNA molecules have numerous loops facilitating protein transport by allowing intersegment transfer between DNA sites near in Euclidean space but possibly distant in contour space. On the other hand, the base-pair bonds of an RNA molecule which is in a folded state for enthalpic reasons constitute a geometrically complex three-dimensional tertiary structure. Since this structure is voluminous, it has to be destroyed in order to allow single-stranded translocation of the RNA molecule through the narrow ribosomal channel. Therefore, the pattern of DNA loops facilitates protein translocation, whereas the pattern of RNA base-pair bonds impedes RNA translocation.

1.2 Translocation of RNA hairpins and pseudoknots through ribosomes

Characterizing the process of RNA translocation through ribosomes kinetically and dynamically is of great importance for lots of biological questions because genetic information encoded in RNA is read codon by codon when translated in the ribosome. Codons are units of information, each consisting of three successive bases along the RNA chain. Every type of codon corresponds uniquely to a distinct amino acid. Thus, the sequence of codons along the RNA chain determines the sequence of amino acids in the protein chain produced by the ribosome during the RNA translocation process.

Under certain conditions pausing during the process of translocation (*ribosomal pausing*) causes *-1 ribosomal frameshift*, which is a 1 base upstream shift of the reading frame the ribosome uses for reading out the codons and changes which base belongs to which codon: every 3-base-codon gives its upstream rim base to the codon that is immediately upstream, receiving in turn one base from its downstream neighbor codon. Due to this massive change in codon structure, the resulting sequence of amino-acids in the protein chain generated by the ribosome is in general totally different from the sequence which would have been generated without ribosomal frameshift (cf. Fig. 1.2, which shows the opposite case of a +1 frameshift). So, ribosomal frameshift causes the production of proteins highly different from the non-frameshift case. That means, more than one protein structure can be encoded within the same RNA base sequence, an advantage heavily used in nature, especially by many RNA viruses (HIV, e.g.).

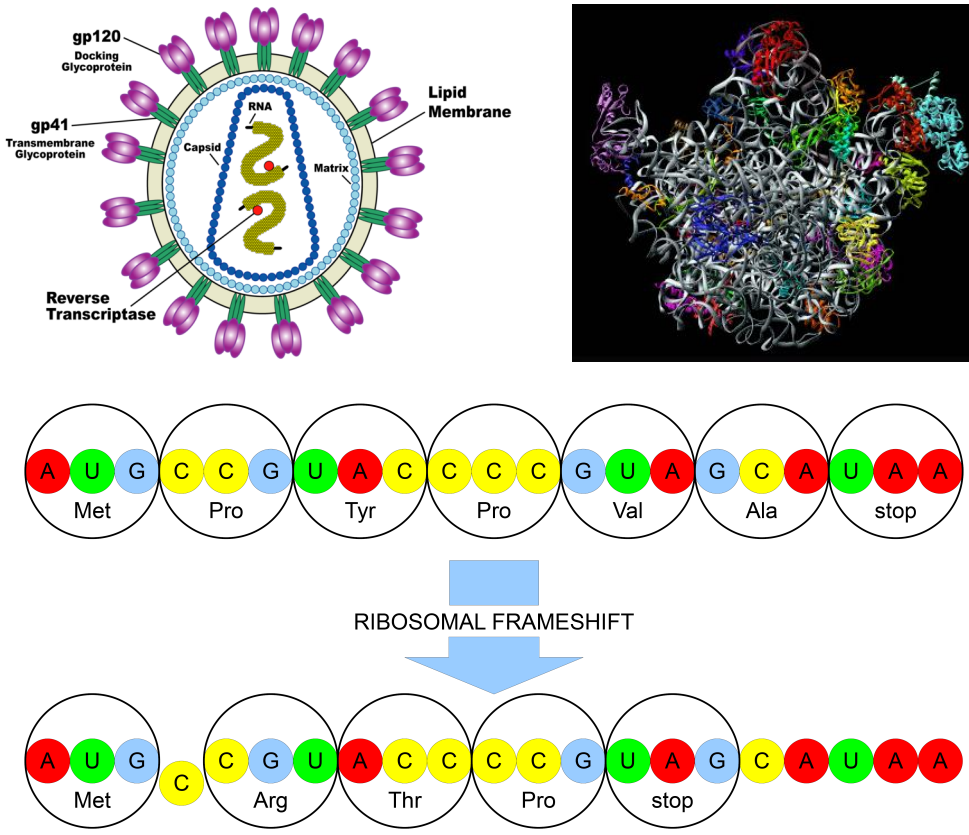


Fig. 1.2: The HI virus, as depicted top left, is one of numerous RNA viruses whose RNA, when passing through a ribosome (figure top right) during translation, undergoes ribosomal frameshift events. These events imply regrouping of the nucleotides to codons, so that proteins with a different sequence of amino acids are produced (bottom figure).

(HIV graphics by Dr. Ebert-May, Michigan State University [3], ribosome graphics by George Rice, Montana State University [4])

As ribosomal pausing, although not sufficient for -1 ribosomal frameshift to occur [5], is one necessary main condition [6][7], and as a change of pausing times affects frameshift

efficiency (as reviewed in [8]), the mechanisms causing the RNA translocation process to stop are of great interest from a biological as well as from a physics point of view. Only special tertiary structure motifs such as some types of pseudoknots have been observed to cause high level ribosomal pausing [7]. By contrast, simple hairpin structures of comparable length lack this property completely. Therefore, the subject of the numerical studies done with this RNA model will be to explain how a pseudo-knotted structure is able to slow down translocation appreciably as soon as it arrives at the ribosome, whereas a hairpin structure is not. The question is whether, and if yes, how the opening process of the pseudoknot is decelerated compared to the opening process of the hairpin.

A variety of models has been proposed to explain the deceleration mechanism presumably responsible for the high frameshift efficiency measured experimentally for certain types of pseudoknots: The frameshift efficiency has been causally connected with differential transition state energy barriers [9], with the mechanical stability of the pseudoknot as measured by unzipping experiments [10] or with the degree of rotational freedom of the first stem of the pseudoknot, which is typically restricted by the second stem. This 'torsional restraint model' [11] has been supported experimentally by comparing the frameshift efficiencies of so-called pseudo-pseudoknots, which are hairpins with additional RNA strands attached to form a pseudoknot-like binding topology [12], as well as by cryo-electron microscopic imaging of ribosomes and pseudoknots [13].

All these models find possible causes for the occurrence of ribosomal pausing during pseudoknot translocation, implicitly assuming that such pausing can increase the frameshift efficiency, as explained by the 9-angstrom model [14]: If an upstream pseudoknot motif blocks the entrance of the ribosome, the mechanical tension a 9 angstrom step of downstream translocation generates in the piece of mRNA already in the ribosome channel can only be released by shifting the reading frame by one base upstream (-1 frameshift).

So the question is whether the different translocation times for pseudoknots and hairpins emerge from topological, geometrical or energetic differences between the two types of structure. In order to address this question by the means of numerical simulations it is necessary to implement a model for RNA folding in three dimensions. The process of RNA folding is relatively well understood on the secondary structure level, that means as far as structure formation in the abstract space of base-pairing patterns is concerned. On this level not only numerous prediction algorithms, which often account even for pseudoknots, have been developed [15][16][17][18][19], but also the dynamics of RNA base-pairing has been studied explicitly [20][21][22][23][24] (including the translocation of RNA through nanopores [25][26]) using the method of calculating partition functions [27]. However, on the level of the three-dimensional structure in real space, there are hardly any modeling approaches, apart from fully-fledged molecular dynamics simulations [28][29], which are challenging even for small RNA molecules, or hybrid prediction algorithms like Kinefold [30], which operate still on the secondary structure level but include topological and geometrical constraints emerging from tertiary structure. Towards the ultimate goal of filling this gap, in this work a coarse-grained bead-spring type polymer model for RNA is constructed, which behaves like a freely jointed

rouse chain as long as the bases are unpaired. However, when the short-range sequence-dependent interactions between the bases set in, more complex interactions between adjacent base-pairs act to spontaneously create double-helical segments with a non-vanishing bending rigidity and torsion stiffness, mechanical properties of folded nucleic acids well-studied in experiments [31][32]. The rich behavior of this model, including folding dynamics as well as static and dynamic properties of the folded tertiary structure and – last but not least – the unfolding and translocation of differently-structured model molecules (RNA hairpins and pseudoknots) through narrow pores is studied by the use of Brownian dynamics simulation techniques, a well-established procedure for solving many-body problems with complex interactions numerically, particularly appropriate for simulating polymers [33][34]. Polynucleotides have also been translocated through nanopores experimentally, studying the translocation of ssDNA [35], the unzipping kinetics of dsDNA [36] and especially of DNA hairpins [37][38], as well as the unzipping of DNA duplexes [39] and the translocation of ssRNA and folded dsRNA [40].

1.3 Translocation of proteins on DNA: A target search process

As suggested above, search processes are wide-spread throughout molecular biology: they are relevant for gene regulation (cf. Fig. 1.3), where transcription factors have to find their specific binding sites on a DNA molecule (the original Jacob-Monod hypothesis for gene regulation [41] was modified by Bourgeois and colleagues, who showed that the regulatory function of lacI is mediated by a protein instead of an RNA transcript [42]), as well as for bacterial self-defense against invading viruses, for which restriction enzymes have to find cleavage sites marked by specific DNA sequences [43][44]. Despite their molecular biological nature these search processes combine features and concepts of various mathematical and physical areas: polymer physics, statistical physics, the mathematics of stochastic processes and the theory of transport and transport networks [45][46][47][48][49][50][51][52][53][54][55]. Especially, characterizing protein target search for specific binding sites on DNA quantitatively has become a paradigmatic question in biological physics [56]. Furthermore, *in vitro* [57][58][59][60] and meanwhile even *in vivo* single-molecule experiments [61] allow the direct observation of this search process.

Therefore, the search of a particle (protein) for a specific site on a heterogeneous polymer (DNA) is an interesting physics problem posed by the molecular biology of gene regulation. Historically, the study of this problem was initiated by early *in vitro* experiments [62], which showed that the *lac* repressor finds its specific binding site on short pieces of DNA faster than expected from the limit of three-dimensional diffusion. Since, as Adam and Delbrück pointed out [63], reaction rates can be increased in principle by a reduction of dimensionality, Richter and Eigen proposed a two step search process to explain the experimental results [64]: after a phase of three-dimensional diffusion in the solvent the transcription factor binds non-specifically to the DNA and eventually finds its specific binding site by one-dimensional sliding along the DNA. Finally, Berg, Winter and von Hippel established an inspiring theoretical framework highlighting a great vari-

ety of the facets of protein DNA search kinetics [65][66][67][68][69]. By this theory they were able to explain their striking experimental results: The occurrence of an optimum in the dependence of the search time on non-specific binding strength.

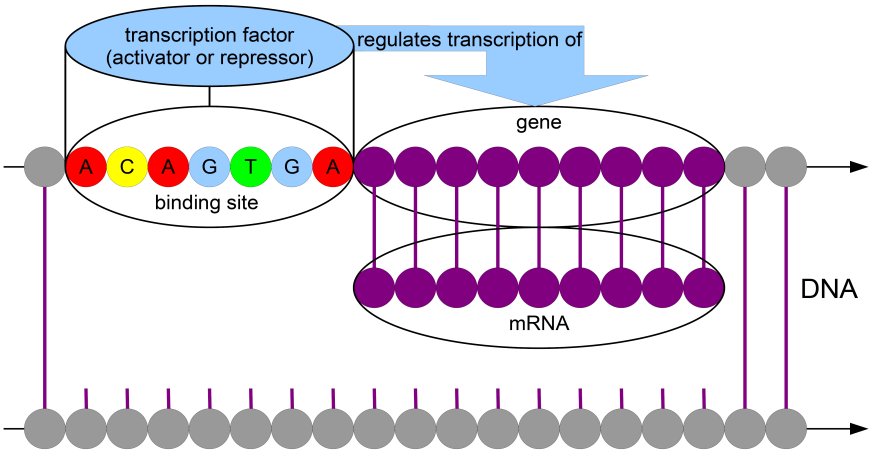


Fig. 1.3: Transcription factors are proteins that regulate the transcription of genes on the DNA into pieces of messenger RNA by binding to specific DNA target sites near these genes. The optimality of the target search process the transcription factors perform in order to find such sites is of crucial importance for gene regulation.

From the perspective of the theory of stochastic processes, the existence of such an optimal search time is a generic feature in search processes with hidden targets [70]. Especially, if the search process consists of periods of one-dimensional sliding along the DNA contour and periods of three-dimensional diffusion in the solvent, the search time is optimal if both types of diffusion have equal duration on average, i. e. if the probabilities to find the protein in the solvent or non-specifically bound to the DNA are equal [68]. The reason is that stochastic local search, represented here by sliding along the DNA, is exhaustive but redundant. This redundancy is destroyed here by periods of three-dimensional diffusion, which allow the protein to reattach at remote and therefore unexplored regions of the DNA contour. However, these excursions into the solvent cost time as well, which is the cause for the condition for the search time optimum [49].

Another possibility to destroy redundancy and thus enhance a search process are intersegment transfers [53][55][66]. An intersegment transfer is the translocation of a protein with two DNA-binding domains from one segment of the DNA contour to another segment possibly distant in contour space without entering the solvent, given both DNA segments are neighbors in three-dimensional space due to DNA looping. Since three-dimensional diffusion is rare under natural conditions in bacterial cells [71], where the generic (electrostatic) attraction between the protein and DNA is strong, and as their time cost is very low compared to excursions into the solvent, intersegment transfers are a very important alternative.

Whereas starting and ending point of 3D-diffusion events are uncorrelated sites on the DNA that are relatively independent of the chain conformation, intersegment trans-

fers happen only between DNA sites correlated because of back-looping of the chain. Very similar to intersegment transfers – as far as their correlatedness and their ability to destroy redundancy in search processes without high time cost are concerned – are hopping events to DNA sites nearby in three-dimensional space, which by contrast include short protein passages through the solvent [66].

To imitate the regime relevant in nature, this study will concentrate on protein target search on a dynamic DNA conformation proceeding by local one-dimensional sliding along the DNA and intersegment transfers. For the first time, the influence of the DNA dynamics on the search dynamics of the protein will be analyzed. As the protein has to cross an intersegment link, which is a geometrical feature of the DNA conformation, every time it performs an intersegment transfer, such an influence exists indeed.

If the time between two intersegment transfer events is sufficiently long, such that the DNA conformations at subsequent events are uncorrelated, the dynamics of this search process can be described [51] within the elegant framework of the fractional Fokker-Planck equation [72], which predicts superdiffusive behavior. However, such a treatment is only valid if the DNA dynamics is very fast compared to protein dynamics, thus eliminating any correlations ("annealed limit").

If on the other hand the DNA conformation is frozen ("quenched limit"), the correlations between subsequent intersegment transfer events persist and the search dynamics changes drastically, displaying quasi-diffusive instead of superdiffusive behavior. This "paradoxical" result has been found by numerical studies [73].

Since in biological systems typically neither of these limits (quenched or annealed) with their contradicting types of transport behavior is realized (relevant biological timescales: $1\mu s - 1s$) [68], studying the physics of the intermediate regime will prove of great relevance. This treatment of the full target search problem, where not only the protein dynamics but also the DNA dynamics is implemented explicitly, is done for the first time within this work: In the intermediate regime, where the rate of DNA dynamics is finite, a non-trivial rate-dependent crossover between the two limits of transport behavior is observed, which is due to a dynamical breakdown of the correlations present in the intersegment linking patterns of frozen DNA conformations. These correlations and the mechanism by which they render the expected superdiffusive transport behavior to the paradoxical quasi-diffusive transport behavior observed for frozen DNA conformations are characterized in detail: There, quasi-diffusion is caused by the formation of regions of the DNA contour which are special in terms of their local intersegment linking patterns. These regions are a manifestation of the postulated strong geometrical correlations. Their remarkable properties cause a competition between long jumps and long trapping times of the protein within these regions, which finally leads to quasi-diffusion.

2 RNA folding, deformation and pore translocation

2.1 Model

2.1.1 Geometry

An RNA molecule is a chain whose backbone monomers consist of phosphate (5 atoms) and sugar (ribose, 20 atoms) to which bases with 8 (uracil) to 16 (guanine) atoms are connected, predominantly occurring in a double-helical folding geometry which is known as canonical A-type RNA and whose parameters were measured in diffraction experiments for the first time by Arnott et al. [74][75]. (A survey of the most important angles and distances for A-RNA can be found in [76].)

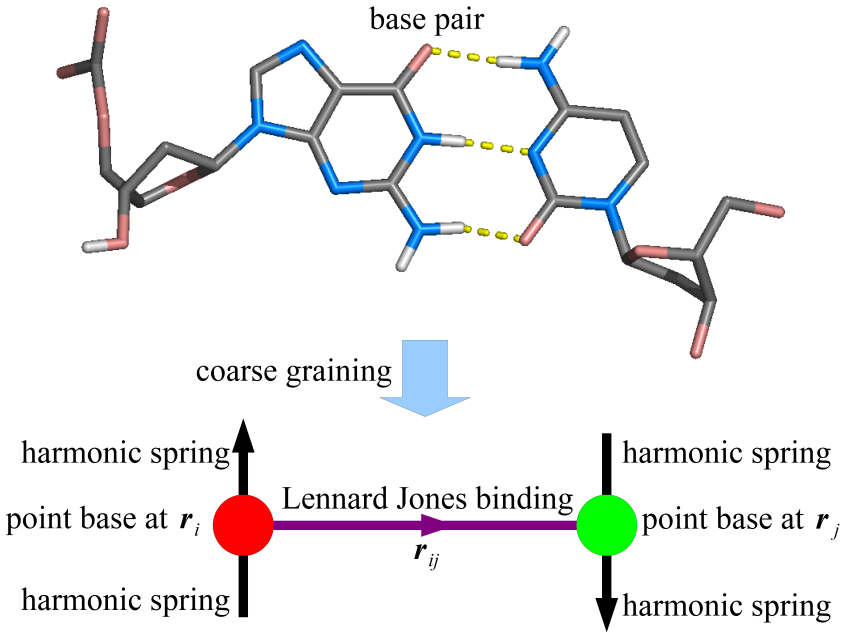


Fig. 2.1: Coarse-graining.

The coarse-grained model for RNA to be established here substitutes this comparably large number of atoms per backbone/base-monomer by one single point-shaped monomer (cf. Fig. 2.1). This is a much higher degree of coarse-graining than current non-atomistic three-dimensional RNA models exhibit (Hyeon and Thirumalai represent one base by

three beads [77], Ouldridge et al. by one backbone rod and one binding rod [78]). The advantage of this high degree of coarse-graining, the simpler and clearer three-dimensional structure, has the price of intricate considerations for the local coordinate systems needed for the angular potentials.

To begin with, a N-base RNA molecule is modeled essentially as a chain of N point-beads at positions $\boldsymbol{r}_k$ connected along the chain contour by harmonic springs. Vectors connecting two such bases to a pair (base connectors) - especially important for the base-pair forming Lennard-Jones potential in the process of folding - are denoted

$$\boldsymbol{r}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i \,. \tag{2.1}$$

Vectors connecting the centers of mass of two adjacent pairs (**base-pair connectors**) are denoted:

$$\boldsymbol{z}_{ij} = \frac{1}{2}\left(\boldsymbol{r}_{i+1} + \boldsymbol{r}_{j-1} - \boldsymbol{r}_i - \boldsymbol{r}_j\right) \tag{2.2}$$

Their direction can be identified as direction of the local stem axis if an RNA stem is formed during folding. They are important for all angular potentials, which endow the stem with its characteristic double-helical geometry. In the following the relevant angles are listed:

The local **bending angle** β_{ij} of the stem axis is defined by

$$\cos(\beta_{ij}) = \hat{\boldsymbol{z}}_{ij} \cdot \hat{\boldsymbol{z}}_{i+1,j-1} \tag{2.3}$$

The local torsion angle ϕ_{ij} between adjacent base-pairs is defined by

$$\cos(\phi_{ij}) = \tilde{\boldsymbol{r}}_{ij} \cdot \tilde{\boldsymbol{r}}_{i+1,j-1} \tag{2.4}$$

and

$$\sin(\phi_{ij}) = \left(\tilde{\boldsymbol{r}}_{ij} \times \tilde{\boldsymbol{r}}_{i+1,j-1}\right) \cdot \hat{\boldsymbol{z}}_{ij} \,, \tag{2.5}$$

where the $\tilde{\boldsymbol{r}}_{ij}$ denote the directions of the base-pair vector components perpendicular to the local stem axis:

$$\tilde{\boldsymbol{r}}_{ij} = \left[\boldsymbol{r}_{ij} - \left(\hat{\boldsymbol{z}}_{ij} \cdot \boldsymbol{r}_{ij}\right)\hat{\boldsymbol{z}}_{ij}\right]_0 \tag{2.6}$$

The local **tilt angle** θ_{ij} of base-pair inclination towards the local stem axis is defined by

$$\cos(\theta_{ij}) = \hat{\boldsymbol{r}}_{ij} \cdot \left(\boldsymbol{z}_{ij} + \boldsymbol{z}_{i+1,j-1}\right)_0 \,. \tag{2.7}$$

Fig. 2.2 shows the - although coarse-grained still complex - three-dimensional double-helical structure of a fully folded RNA hairpin (right) consisting of N bases compared to its relatively simple base-pairing scheme (left) used by RNA models that operate on base-pairing space only. The two helices of the stem are marked red and green, the loop is marked yellow.

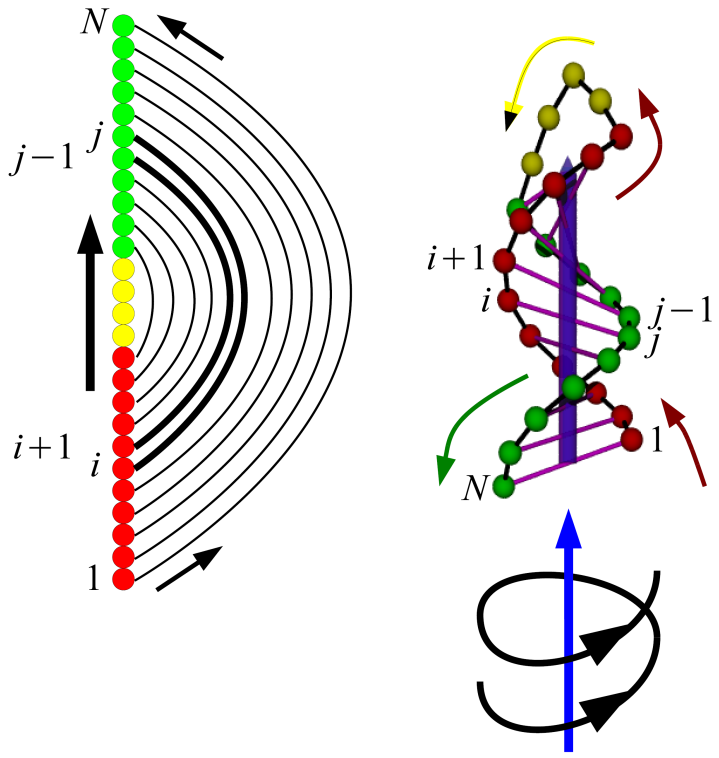


Fig. 2.2: Space of base-pairing and 3D space.

In order to visualize all vectors and angles defined above it is helpful to zoom into the part of the stem in the neighborhood of the base-pair (i, j), as is shown in Fig. 2.3.

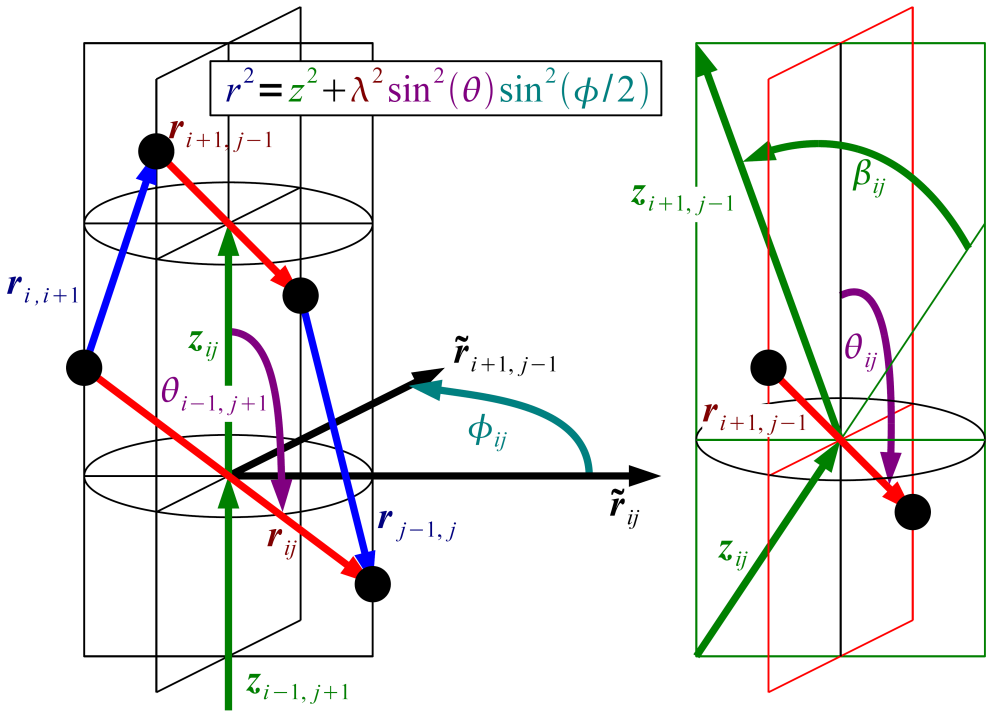


Fig. 2.3: Denotation of all vectors and angles in the model.

On the left part of the figure, where a so-called **local cell of the stem** (two fully formed adjacent base-pairs) is shown, the local bending angle vanishes, so that the tilt angle can be measured relative to the constant direction of the local stem axis. The (red) base-pair vectors are always directed from a low index base to its high index partner base, the (blue) backbone vectors are directed from one base to its successor along the backbone and the (green) base-pair connectors describing the local stem axis are always oriented in the direction of the low index backbone.

In general, however, the bending angle does not vanish, as shown to the right, and therefore the tilt angle has to be measured relative to an effective stem axis direction determined by the two stem axes touching the tilted base-pair.

If (and only if) a local cell consists of two equally-sized base-pairs ($r_{ij}=r_{i+1,j-1}\equiv\lambda$) that have equal tilt angles ($\theta_{i-1,j+1}=\theta_{ij}\equiv\theta$) and are connected by two equally-sized backbone segments ($r_{i,i+1}=r_{j-1,j}\equiv r$), some analytical geometry yields the following relation with the torsion angle $\phi_{ij}\equiv\phi$ and the cell axis length $z_{ij}\equiv z$:

$$r^2=z^2+\lambda^2\sin^2(\theta)\sin^2(\phi/2) \tag{2.8}$$

This is a kind of Pythagoras relating the relevant dimensions and angles of a local cell of a folded stem to each other. As one will see later, this local relation is extremely useful in describing global properties of stems in symmetric situations.

2.1.2 Dynamics

Using the Forward-Euler-Method, the time-dependent positions $\boldsymbol{r}_k(t)$ of the N point-shaped bases the modelized RNA chain consists of are determined numerically by integration of the Langevin equation of motion:

$$\partial_t \boldsymbol{r}_k=\frac{1}{\zeta}\left[-\nabla_k\left(U_{\mathrm{RNA}}+U_{\mathrm{pore}}\right)+\boldsymbol{\Gamma}_k(t)\right] \quad \forall\, k:\ 1\le k\le N \tag{2.9}$$

The motion of each base is governed by forces from potential gradients and by random forces within the framework of a Brownian dynamics simulation [34][79]. For the thermal random forces $\boldsymbol{\Gamma}_k(t)$ isotropic white noise is used, i.e. there are no correlations between random forces at different points of time:

$$\begin{aligned}\left\langle \Gamma_{i\alpha}(t)\Gamma_{j\beta}(t')\right\rangle&=2kT\zeta\,\delta_{\alpha\beta}\delta_{ij}\delta(t-t')\\ \left\langle \Gamma_{i\alpha}(t)\right\rangle&=0\end{aligned} \qquad \forall\, i,j,\alpha,\beta:\ 1\le i\le N,\ i<j\le N,\ \alpha,\beta=1,2,3 \tag{2.10}$$

The forces from potential gradients can be subdivided into forces within the RNA molecule itself (intra-molecular interactions) and into external forces exerted on the molecule by a pore.

The **interaction between the RNA molecule and the pore** is modelized as a superposition of two potentials (cf. Fig. 2.4):

- The pore wall potential is a Gaussian wall (diameter l_{pore}, height E_{barrier}) in the xy-plane with a single axially symmetric Gaussian-shaped hole (diameter d_{pore}) at the origin. This potential is repulsive enough to allow the RNA molecule to pass the xy-plane only through a small cylindrical channel at the origin.

- The pore drift potential exerts a force on the RNA molecule that is strongest for bases in the pore channel, thus translocating the molecule base for base through the pore. The energy gain per translocated base is E_{drift}, the force on a base at the origin $E_{\text{drift}}/l_{\text{pore}}$.

$$U_{\text{pore}}=\sum_{k=1}^{N} E_{\text{barrier}}\left[1-\exp\left(-\frac{(\boldsymbol{r}_k)_x^2+(\boldsymbol{r}_k)_y^2}{2\,d_{\text{pore}}^2}\right)\right]\exp\left(-\frac{(\boldsymbol{r}_k)_z^2}{2l_{\text{pore}}^2}\right)+\frac{1}{2}E_{\text{drift}}\tanh\left(\frac{2(\boldsymbol{r}_k)_z}{l_{\text{pore}}}\right) \tag{2.11}$$

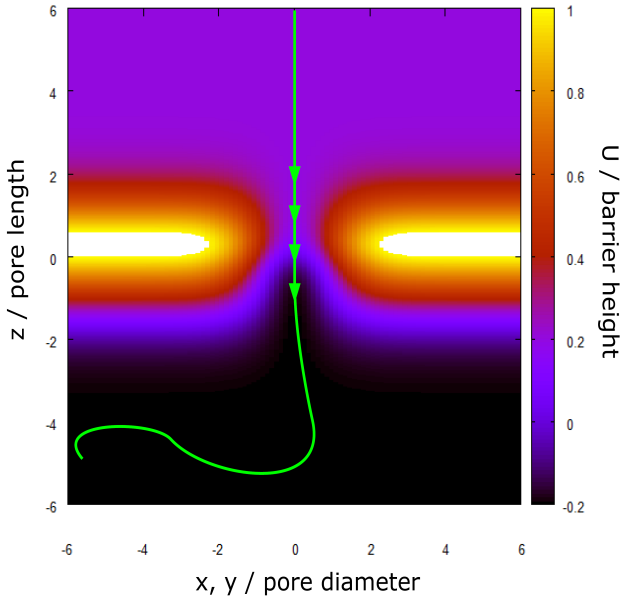


Fig. 2.4: RNA chain in pore potential.

The **intra-molecular interactions** of the model consist of the harmonic springs between next neighbor bases along the RNA backbone, of the Lennard-Jones interactions between chemically distant bases and of the angle-stabilizing interactions between adjacent base-pairs:

$$U_{\text{RNA}}=U_{\text{bond}}+U_{\text{LJ}}+U_{\text{bp}} \tag{2.12}$$

Harmonic springs between successive bases along the RNA backbone:

The RNA backbone is modelized as a bead spring chain, i.e. the N point-shaped bases are inter-connected along the chain contour by $N-1$ harmonic springs (spring constant K_{bond}, spring length r_0). These (almost rod-like) springs are the main carrier of geometric stability of the model molecule.

$$U_{\text{bond}}=\frac{1}{2}K_{\text{bond}}\sum_{k=1}^{N}\left(r_{k,\,k+1}-r_0\right)^2 \tag{2.13}$$

Lennard-Jones interaction between compatible and incompatible bases:

$$U_{\text{LJ}}=\sum_{i=1}^{N}\sum_{j=i+1+\Delta}^{N} c_{ij}\Phi_{\text{LJ}}^{\text{comp}}(r_{ij})+(1-c_{ij})\Phi_{\text{LJ}}^{\text{incomp}}(r_{ij}) \tag{2.14}$$

To enable the formation of base-pairs and hence folding of the chain, one has to introduce pair interactions between bases that are compatible (compatibility tensor $c_{ij}=1$).

The interaction potential used is of the Lennard-Jones type (potential depth E_{LJ}, equilibrium distance R):

$$\Phi_{\mathrm{LJ}}^{\mathrm{comp}}(r)=E_{\mathrm{LJ}}\left[\left(\frac{R}{r}\right)^{12}-2\left(\frac{R}{r}\right)^{6}\right] \tag{2.15}$$

If base-pair-sizes are large compared to backbone segment lengths ($R>r_0$), in order to achieve self-avoidance of the chain without hindering the folding process a much shallower version of the Lennard-Jones potential ($E_{\mathrm{LJ}}^{\mathrm{incomp}} \ll E_{\mathrm{LJ}}$) has to be used for the incompatible bases ($c_{ij}=0$). However, to avoid divergences when integrating the equation of motion numerically (with time steps whose duration minimum is limited by computation capacity), one has to circumvent the divergence in potential curvature for small distances. This is done by substituting the Lennard-Jones form by a harmonic potential there:

$$\Phi_{\mathrm{LJ}}^{\mathrm{incomp}}(r)=\left\{\begin{array}{ll} E_{\mathrm{LJ}}^{\mathrm{incomp}}\left[\left(\frac{R}{r}\right)^{12}-2\left(\frac{R}{r}\right)^{6}\right] & \text{for } r\geq\frac{R}{2} \\ ar^2+br+c & \text{for } 0\leq r<\frac{R}{2} \end{array}\right\} \quad a,b,c \text{ from } \begin{pmatrix} \Phi_{\mathrm{LJ}}^{\mathrm{incomp}} \\ \partial_r \Phi_{\mathrm{LJ}}^{\mathrm{incomp}} \\ \partial_r^2 \Phi_{\mathrm{LJ}}^{\mathrm{incomp}} \end{pmatrix} \text{ continuous} \tag{2.16}$$

Interactions between pairs of bases:

An RNA stem forms a regular (A type) double helix which is resistant to a certain degree to bending of its axis, to torsion and to shearing. In nature these mechanical properties are mainly caused by interactions of overlapping orbitals of neighboring base-pairs. That is why in this coarse-grained model local interactions between base-pairs are employed to simulate these properties. These interactions control the relative orientation of a base-pair relative to the local coordinate system made up by its neighboring base-pairs. Up to three subsequent pairs of bound bases are needed to establish such a coordinate system. Thus the weight of the interaction contribution (e.g. $U_{\mathrm{torsion}}^{ij}$) depends on the states of boundness of each of these base-pairs, which are expressed by predicates H_i^j continuously depending on inter-base distance (cf. Fig. 2.5). Therefore the total potential for interactions between base-pairs reads:

$$U_{\mathrm{bp}}=\sum_{i=1}^{N}\sum_{j=1}^{N}\theta_{j\geq i+2+\Delta}\,H_i^j\,H_{i+1}^{j-1}\left[U_{\mathrm{axis}}^{ij}+U_{\mathrm{torsion}}^{ij}\right]+\theta_{j\geq i+3+\Delta}\,H_i^j\,H_{i+1}^{j-1}\,H_{i+2}^{j-2}\left[U_{\mathrm{bending}}^{ij}+U_{\mathrm{tilt}}^{ij}\right] \tag{2.17}$$

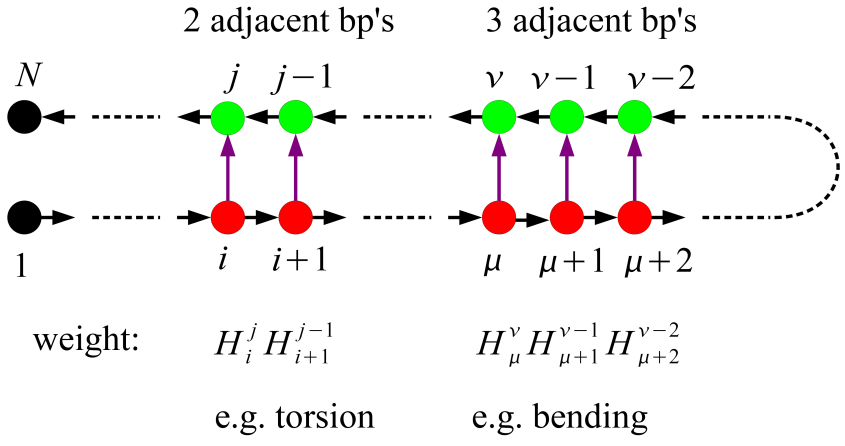


Fig. 2.5: Binding weights.

Contributions:

Only summands with weight factors near unity, i.e. only those where all participating base-pairs are bound strongly, contribute appreciably to this total potential.

The **bending potential** generates the non-vanishing bending stiffness of the stem, thus making it semi-flexible and its equilibrium form rectilinear. This bending stiffness is controlled by the bending energy parameter E_{bend} in each local contribution U^{ij}_{bending}, which is in turn weighted by the combined binding predicate of three subsequent base-pairs and tries to set the local stem bending angle β_{ij} to zero.

$$U^{ij}_{\text{bending}} = -E_{\text{bend}} \cos(\beta_{ij}) \tag{2.18}$$

The **torsion potential** is responsible for the torsional resistance of the stem and its chiral and regularly helical form in equilibrium. The resulting torsion coefficient is controlled by the torsion energy parameter E_{tor}. For the weight of each local contribution U^{ij}_{torsion} the binding predicates of two subsequent base-pairs are needed. The signed equilibrium angle between the corresponding projected base-pair vectors is ϕ_0.

$$U^{ij}_{\text{torsion}} = -E_{\text{tor}} \cos(\phi_{ij} - \phi_0) = -E_{\text{tor}} \left[\cos(\phi_{ij}) \cos(\phi_0) + \sin(\phi_{ij}) \sin(\phi_0) \right] \tag{2.19}$$

The **tilt potential** effects the shear resistance of the stem, which is controlled by the tilt energy parameter E_{tilt}. Furthermore, the radial dimensions of the stem are stabilized by keeping the angle of base-pair inclination relative to the local stem axis at θ_0 in equilibrium. As for the bending potential, again three subsequent base-pairs are needed for each local contribution, because the axis θ_{ij} is related to goes through the centers of mass of the two base-pairs neighboring the inclined base-pair symmetrically.

$$U^{ij}_{\text{tilt}} = -E_{\text{tilt}} \cos(\theta_{ij} - \theta_0) = -E_{\text{tor}} \left[\cos(\theta_{ij}) \cos(\theta_0) + \sqrt{1 - \cos^2(\theta_{ij})} \sin(\theta_0) \right] \tag{2.20}$$

Sigmoidal Hill function as a continuous predicate indicating the degree of existence of a base-pair:

The total potential of interaction between base-pairs U_{bp} consists of three angular potentials for the angles of bending, torsion and inclination. The contributions to such an angular potential are local, i.e. they describe the interactions within a local group of a few subsequent base-pairs (e.g. 2 group members per torsion contribution, 3 group members per bending contribution etc., cf. Fig. 2.5). These local groups are uniquely identified by their first base-pair (i, j). A priori, every pair (i, j) can be the first base-pair of such an angular potential contribution. However, a local group is only fully valid if all of its members are fully formed base-pairs. This gives rise to continuous weighting factors between zero and unity describing the degree of *simultaneous* boundness of all members of the corresponding local group. This is expressed by a product of distance-dependent Hill-type functions (one function per group member) which reflect how strongly a group member (i.e. base-pair) is bound. These functions H^j_i are sigmoidal with values between zero (base-pair unbound) and unity (base-pair maximally bound) and can be seen as continuous logical predicates for the statement "The base-pair (i, j) is fully formed.":

$$H^j_i = c_{ij}\, h\left(\frac{1 - \Phi^{\text{comp}}_{\text{LJ}}(r_{ij}) / \Phi^{\text{comp}}_{\text{LJ}}(R)}{\delta} \right) \tag{2.21}$$

$$h(x)=\frac{1}{1+x^2} \tag{2.22}$$

A visualization of these predicate functions is given in Fig. 2.6.

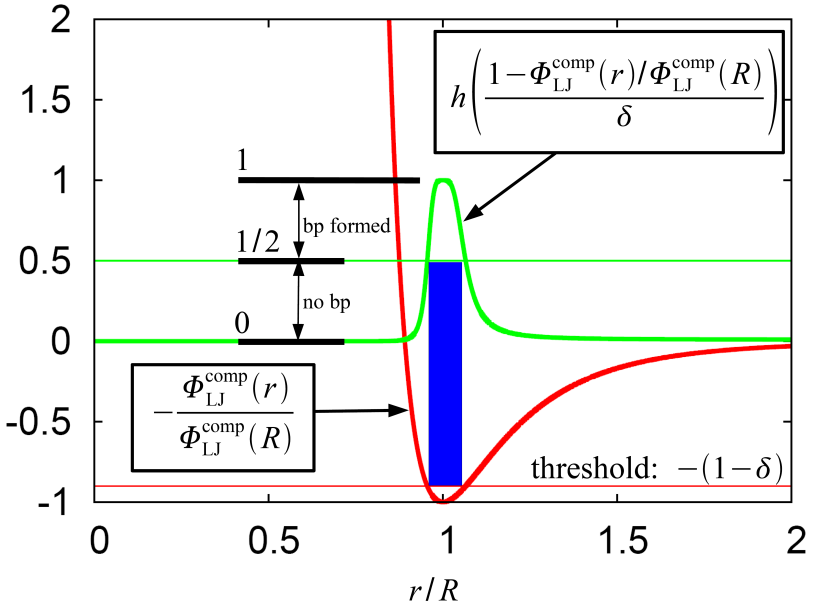


Fig. 2.6: Binding predicate function.

For incompatible pairs of bases the compatibility tensor c_{ij} in H_i^j vanishes. That is, local angular potential contributions containing at least one incompatible pair of bases have vanishing weighting factors. For compatible pairs of bases the threshold parameter δ (between zero and unity) determines how small the deviation between actual inter-base-distance and equilibrium distance R has to be in order to consider a pair of bases bound. Low threshold parameters correspond to small maximal deviations.

The complete algorithm for simulating the model numerically by using the Euler-Maruyama integration method for Brownian dynamics is described in appendix A.

2.2 Folding of RNA

In the **Watson-Crick model** all bases within so-called Watson-Crick pairs (AU, CG, GU) are considered compatible, which yields a compatibility tensor with possibly more than one non-zero entry in a row or column. That means, a base can have more than one possible binding partner (Fig. 2.7, left). In nature simultaneous binding of a base to more than one partner base is impossible due to steric effects caused by the internal structure of the nucleotides. However, the coarse-grained RNA model introduced above neglects such effects. Therefore, chains with homopolymeric sequence segments tend to form clusters of bases instead of pairs in the process of binding. In a hairpin with sequence $(A)_8(C)_5(U)_8$ for example, the A-bases do not only interact with their exactly op-

posite U-partners but with all neighboring U-bases as well, as one can see from the non-vanishing off-diagonal Lennard-Jones contributions in the left part of Fig. 2.8.

In order to find the compatibility tensor in the more restrictive **Go model** [80], one has to determine the (Watson-Crick) base-pairs formed in the ground state of a RNA molecule. Only the bases within this native base-pairs are considered compatible in the Go model. The resulting compatibility tensor contains at most one non-zero entry in a row or column, i.e. a base has at most one possible binding partner (cf. Fig. 2.7, right). Testing the interaction pattern of the example hairpin again (Fig. 2.8, right) results in finding only diagonal non-zero Lennard-Jones contributions. Therefore, base-clustering is impossible and the binding process in the Go model will lead to the formation of base-pairs.

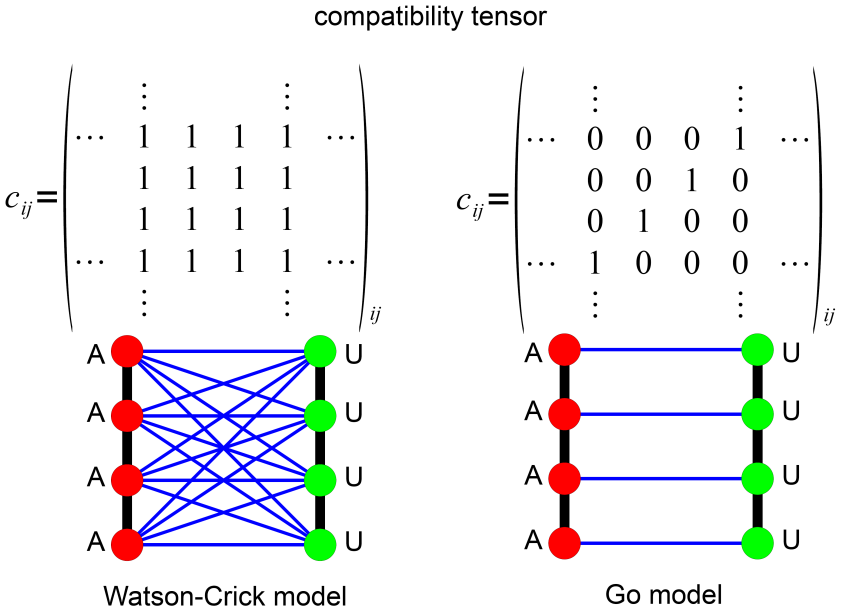


Fig. 2.7: Base compatibility in the Watson-Crick model and in the Go model.

One can modify the Watson-Crick model in a way that base-clustering is prevented as strictly as in the Go model by the means of continuously suppressing bound states with more than two participating bases. This model, referred to as "**suppressor model**" here, shares its compatibility tensor with the Watson-Crick model, thus being less restrictive than the Go model, where only native base-pairs can occur. However, in the Lennard-Jones potential U_{LJ} each summand concerning compatible pairs of bases has to be multiplied by a weighting function S_{ij} (suppressor function):

$$S_{ij}=\prod_{\substack{k=1\\k\neq i,\,j}}^{N}\left(1-H_i^k\right)\left(1-H_k^j\right) \tag{2.23}$$

This function relying on the binding predicates defined above and thus depending continuously on distances between bases has values between zero and unity and reflects how strongly the two bases i and j are already bound to competing bases. If at least one of them is already bound strongly to at least one other base k, the suppressor function

takes small values, whereas if none of them is strongly bound to other bases, the weight is almost unity. Again the Lennard-Jones interaction matrix for the example hairpin is purely diagonal, as already seen in the Go model (Fig. 2.8, bottom).

This suppressor model (Watson-Crick model with suppressor function) combines the advantages of both the Watson-Crick model (not only native base-pairs can form) and the Go model (multiple-base binding states are avoided). It allows perfect folding for short base-pairs ($R \leq r_0$). For long base-pairs ($R > r_0$) the rising mean number of competing binding partners per base causes very small weights, thus reducing the folding probability drastically, so that in this case the Go model has to be used. For very short base-pairs ($R \leq r_0/2$) the Watson Crick model is sufficient (multiple-base binding states are improbable there).

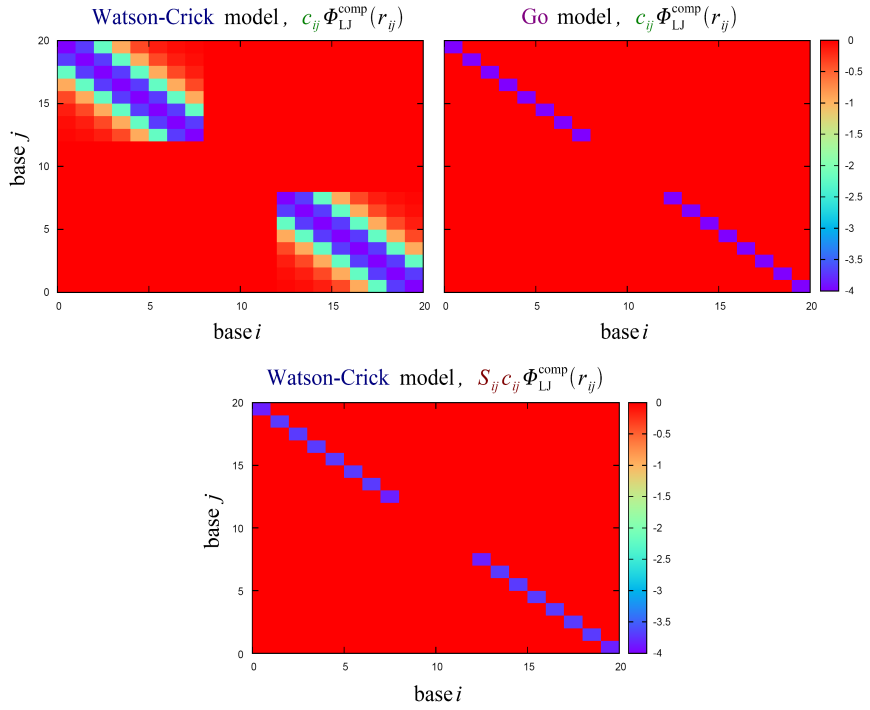


Fig. 2.8: Lennard-Jones energy contributions in several binding models.

In order to test the **folding properties of the suppressor model** for an RNA with short base-pairs ($R = r_0$), the time-dependent behavior of a heteropolymeric RNA chain (sequence AUCGGGCCCGAU) is observed. Initially, the chain is not folded, all of its backbone springs already have their equilibrium extensions and its joint angles are distributed randomly (Fig. 2.9, left part).

After some time, the first base-pair forms (Fig. 2.9, middle part) due to Lennard-Jones attraction and serves as a nucleation center for further base-pairs. Because of the suppressor function, these base-pairs form without showing multiple-base binding states, which would have been the case if an unmodified Watson-Crick model had been used.

As soon as local groups of adjacent base-pairs emerge, the corresponding local stem-fragments begin to obey the thereby continuously activated angular potentials. Finally

(Fig. 2.9, right part), all possible Watson-Crick compatible bases are zipped to pairs, the RNA chain is folded completely to a hairpin with a 4-bp stem and a 4-base loop and the angular potentials cause a perfectly regular chiral helicality of the stem (torsion potential) with constant base-pair inclination angle (tilt potential) and vanishing axis bending angle (bending potential).

This example for simulated folding is consistent with the concept of hierarchical RNA folding as suggested by Tinoco and Bustamante [81]: Since the energies for secondary structure formation are larger then those for tertiary structure formation, there exists a hierarchy of thermodynamical stability with secondary structure elements (helices and hairpin loops etc.) on its top. Moreover, for RNA molecules (as opposed to proteins) the folding kinetics, i. e. the shape of the folding path, is linked closely to the thermodynamics: tertiary structure elements (loop-loop base-pairing, base triples etc.) will form *after* the formation of the secondary structure.

In the short example sequence used here only the process of secondary structure formation plays a role. Nevertheless, the suggested close correspondence between thermodynamics and folding kinetics can be observed: The initial base-pair is stabilized by the formation of consecutive base-pairs and the helical geometry is stabilized in turn by the existence of adjacent base-pairs for the angular interactions.

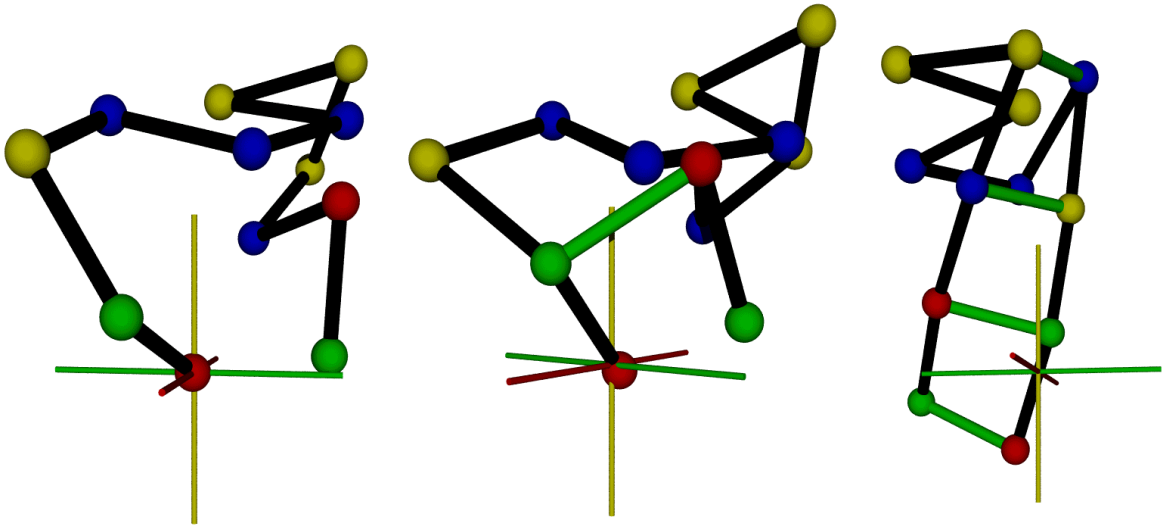


Fig. 2.9: Process of hairpin folding.

Next, the **folding properties of the Go model** at different temperatures are tested. This is done by equilibrating an initially stretched sequence which can form maximally 8 stem base-pairs (all adjacent), whose equilibrium size is $3r_0$ (long base-pairs), and - as soon as equilibration is complete – by averaging the number of fully formed base-pairs over a long interval of time. The resulting probabilities $P(n)$ for observing a distinct number n of base-pairs are shown in Fig. 2.10.

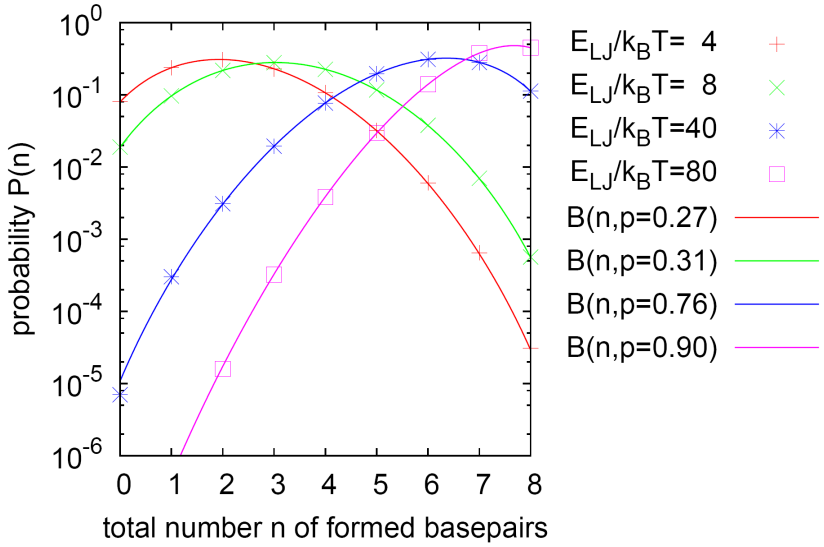


Fig. 2.10: Distribution of the base-pair number at several temperatures.

From there one can see that the number n of formed base-pairs in a stem with maximally m possible base-pairs obeys the binomial distribution:

$$B(n,p)=\binom{m}{n}p^n(1-p)^{m-n} \tag{2.24}$$

That means that in the Go model the probabilities $p_{ij}=Prob(H_i^j>0.5)$ to form single base-pairs are statistically independent, as expected. For the m pairs of Go-compatible bases the probability p decreases with increasing temperature.

The temperature-dependence of hairpin stem opening and closing processes has been studied in experiments using fluorophore-quencher set-ups [82][83]. Bonnet et al. [83] measured the temperature dependence of the opening (-) and closing (+) rates of a (DNA) hairpin stem. From these rates the probability that the hairpin is totally open is deduced easily:

$$P(n=0,T)=\frac{k_-(T)}{k_-(T)+k_+(T)} \tag{2.25}$$

This can be used to calibrate the model parameters relevant for folding (Lennard Jones potential depth, energy parameters of the angular potentials). However, as will be discussed in chapter 2.4 where the RNA model will be calibrated quantitatively, angular energy parameters producing realistic values for double-stranded persistence length and torsional modulus necessarily make this probability very small.

2.3 Deformation of folded RNA

2.3.1 Stretching a hairpin

As will be seen, a good way to test the interplay of some important model potentials is stretching a completely folded hairpin structure at temperatures far below the melting temperature (then all base-pair binding distances as well as the number of fully bound base-pairs can be regarded constant).

Experimentally, the stretching of DNA is well-studied. Smith, Finzi and Bustamante were the first to measure the elasticity of single (double-stranded) DNA molecules and demonstrated the freely jointed chain model to be inconsistent with their data [84]. Below stretching forces of 15 pN entropy dominates such that the DNA behaves like a stretched elastic tube under thermal fluctuations (worm-like chain model) [85]. However, at forces of about 70 pN, where entropic effects are already negligible, there exists an overstretching regime, where the chain undergoes a structural transition during which its length increases appreciably [86]: Computer simulations suggest [87][88] that the final state of this transition is a so-called "S-DNA", which is either planar and ladder-like or still double-helical but with inclined base-pairs [86].

The coarse-grained RNA model constructed in this thesis is able to reproduce this structural transition in principle. In order to observe and identify clearly the relevant geometrical effects and since in the overstretching regime entropic effects can be neglected anyway, all thermal fluctuations have been switched off. Pulling symmetrically at the centers of mass of the base-pairs at the model molecule's ends, by that avoiding any torsional constraints or unzipping forces, results of course in a ladder-like final conformation, which is consistent with the mentioned numerical simulations.

By contrast, Mameren et al. [89] showed experimentally (by the use of multicolor, single-molecule fluorescence imaging) that for double-stranded B-DNA this transition implies force-induced melting into single strands instead of double-stranded unwinding, even for symmetric pulling without torsional constraints. The corresponding theoretical framework was set up earlier by Rouzina and Bloomfield [90][91].

On the other hand, Rief, Clausen-Schaumann and Gaub had shown in 1999 [92] by single-molecule force spectroscopy based on atomic force microscope technology that there is an additional transition at a much higher force than the already known 70-pN transition: At 150 pN a non-equilibrium force-induced melting transition with strong hysteresis upon relaxation had been observed. This relaxation had proven indistinguishable from single-stranded stretching and relaxation, thus indicating a molten DNA state above the transition.

Eventually, the experiments of Fu et al. [93] gave evidence that both force-induced melting ("unpeeling") and the formation of a stretched double-stranded "S-DNA" conformation are competing processes occurring during the overstretching transition, the percentage of each depending on parameters that influence base-pair stability, like sequence, salt concentration and temperature.

However, it is still unknown what happens to A-RNA hairpins when overstretched. If they exhibit a double-stranded stretching mode as well, the RNA model constructed here can be used to obtain additional insight into the geometrical effects of the unwinding process involved. Within this RNA model torsion effects may impede the unwinding process occurring during axial overstretching, thereby effectively decreasing double-stranded stretching elasticity in this force regime. This interplay between backbone elasticity and stem torsion stiffness in causing double-stranded stretching elasticity will be studied quantitatively in the following.

The first task is trying to predict theoretically the behavior of the model molecule when performing a numerical stretching experiment, i. e. the relation between the (kept fix) axial extension of a hairpin stem L and the corresponding response force F_z which attempts to equilibrate this extension. For that, it is useful to remember that due to reasons of symmetry the stem can be subdivided into independent segments, each consisting of two subsequent base-pairs. The geometry of such a segment is determined by the direction of the two base-pairs (ϕ, θ), their binding distance $\lambda (= \text{const} = R)$, the distance z of their centers of mass, i.e. the segment extension, and the current length r of the two connecting backbone springs. All angles and lengths just mentioned are connected by the Pythagoras relation of the local cell (cf. equation 2.8), which will be extremely useful in what follows.

For unbent stems the segment extension is equal to the axial extension per segment $(L = z \cdot N_{\text{seg}})$. Similarly, one gets for the total potential energy of the stem $U_{\text{stem}} = U \cdot N_{\text{seg}}$, where U denotes the contribution of a single segment, from which in turn the response force F_z is obtained.

In the two the simplest cases, which yield the first of the two analytically tractable limit curves (green line) in the force extension diagram of Fig. 2.12, the torsion energy is set to zero and the response force is caused purely by the deformation of the back-bone springs:

a)vanishing torsion and tilt energy

$$U = U_{\text{bond}}(r(\phi, \theta, z)) = K_{\text{bond}}\left[r(\phi, \theta, z) - r_0\right]^2 \tag{2.26}$$

b)vanishing torsion energy, but large tilt energy (i.e. inclination angle fixed)

$$U = U_{\text{bond}}(r(\phi, \theta_0, z)) + U_{\text{tilt}}(\theta_0) = K_{\text{bond}}\left[r(\phi, \theta_0, z) - r_0\right]^2 - E_{\text{tilt}} \tag{2.27}$$

Although simply harmonic, these potentials lead to rich geometrical and dynamical behavior, which can be summarized within three distinct regimes (cf. Fig. 2.11). The expected Hookean relation between axial extension and response force is only observed for the regime of backbone spring stretching (third column).

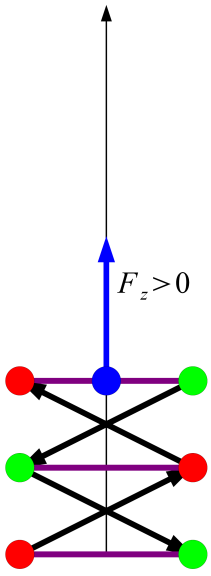
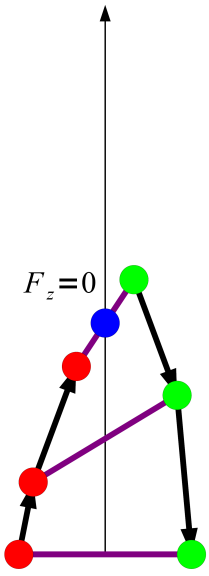
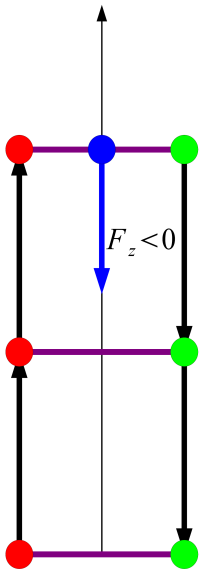
regime of spring compression (exists only if $R<r_0$)	regime of equilibrium spring length	regime of spring stretching
range of segment extension z: a) $0<z<\sqrt{r_0^2-R^2}$ b) $0<z<\sqrt{r_0^2-R^2\sin^2\theta_0}$	range of segment extension z: a) and b) $0<z<r_0$ if $R>r_0$ else: a) $\sqrt{r_0^2-R^2}<z<r_0$ b) $\sqrt{r_0^2-R^2\sin^2\theta_0}<z<r_0$	range of segment extension z: a) and b) $z>r_0$
geometry in energy minimum: a) $r_*(z)=\sqrt{z^2+R^2}<r_0$ $\phi_*=\pm\pi$ $\theta_*=\pi/2$ b) $r_*(z)=\sqrt{z^2+R^2\sin^2\theta_0}<r_0$ $\phi_*=\pm\pi$	geometry in energy minimum: a) $r_*(z)=r_0$ $\sin^2(\phi_*/2)=\frac{r_0^2-z^2}{R^2\sin^2(\theta_*)}$ b) $r_*(z)=r_0$ $\sin^2(\phi_*/2)=\frac{r_0^2-z^2}{R^2\sin^2(\theta_0)}$	geometry in energy minimum: a) $r_*(z)=z>r_0$ $\phi_*=0$ $\theta_*=$ arbitrary b) $r_*(z)=z>r_0$ $\phi_*=0$
non-linear response force: $F_z^0(z)=2K_{\text{bond}}[r_0-r_*(z)]\frac{z}{r_*(z)}>0$	vanishing response force: $F_z^0(z)=0$	Hookean response force: $F_z^0(z)=-2K_{\text{bond}}(z-r_0)<0$
		

Fig. 2.11: Regimes of axial stem deformation for vanishing torsion energy.

From the form of the potential one determines the angles which minimize energy for a certain segment extension. Then one differentiates with respect to segment extension.

The force obtained by that is the response force trying to equilibrate segment extension. From that, a dimensionless response force can be calculated (the index 0 denotes zero torsion energy):

$$f_0(L) = -\frac{F_z^0(L/N_{\text{seg}})}{K_{\text{bond}}\, r_0} \tag{2.28}$$

However, allowing for non-zero torsion energy (while keeping the inclination potential strong) results in corrections to the pure case (b) that has already been treated. By that, the sharp regime boundaries described above cease to exist. For finite torsion energies the corresponding potential contribution reads:

$$U = U_{\text{bond}}(r(\phi,\theta_0,z)) + U_{\text{torsion}}(\phi) + U_{\text{tilt}}(\theta_0) = K_{\text{bond}}\left[r(\phi,\theta_0,z) - r_0\right]^2 - E_{\text{tor}}\cos(\phi-\phi_0) - E_{\text{tilt}} \tag{2.29}$$

Now the equilibrium torsion angle needed before differentiation cannot be determined analytically for all segment extensions z, as it had been possible for the pure case (b). Only its asymptotic behavior for very large z can be derived:

$$\cot\phi_* \approx \cot\phi_0 + \left[\frac{1}{2\sin\phi_0}\left(\frac{R\sin\theta_0}{r_0}\right)^2\right]\frac{K_{\text{bond}}\, r_0^2}{E_{\text{tor}}} \quad \text{for} \quad z \gg r_0 \tag{2.30}$$

Obviously, for very large torsion energies this changes to

$$\phi_* \approx \phi_0 \quad \text{for} \quad z \gg r_0\,, \tag{2.31}$$

as expected. However, in this asymptotic z-region the response force becomes increasingly independent of the torsion angle, so that for all torsion energy values it approaches the same Hookean form as derived for the spring stretching regime in Fig. 2.11. Therefore this Hookean form can be considered as universal asymptote for large segment extensions.

In the limiting case of vanishing torsion energy the response force is already known exactly for all possible segment extensions. For finite torsion energies it can only be predicted asymptotically for large segment extensions. In the opposite limiting case of backbone spring energies that are small compared to the torsion energy the response force is predictable again for the whole range of segment extensions. The potential contribution one has to consider in this limit is:

$$U = U_{\text{bond}}(r(\phi_0,\theta_0,z)) = K_{\text{bond}}\left[r(\phi_0,\theta_0,z) - r_0\right]^2 - E_{\text{tor}} - E_{\text{tilt}} \tag{2.32}$$

By differentiating one obtains the dimensionless response force function $f_\infty(L)$ for very large torsion energies, which deviates substantially from $f_0(L)$, the function for vanishing torsion energy (cf. equation 2.28, Fig. 2.11):

$$F_z^\infty(z) = -2K_{\text{bond}}\left[r(\phi_0,\theta_0,z) - r_0\right]\frac{z}{r(\phi_0,\theta_0,z)} \equiv -K_{\text{bond}}\, r_0\, f_\infty(N_{\text{seg}}\, z) \tag{2.33}$$

The theoretical predictions for these two functions can now be confirmed by simulation experiments. The hairpin stem under consideration consists of 4 subsequent base-pairs, i.e. of 3 stem segments or local cells ($N_{\text{seg}}=3$). In order to exclude entropic effects and keeping the base-pairs stable, the temperature is set to zero. For several values of kept fix axial extension L the stem with all its angles and lengths is allowed to equilibrate be-

fore measuring the resulting response force F_z. This force is plotted against the axial extension for several values of torsion energy (cf. Fig. 2.12). The data points for vanishing torsion energy follow exactly the theoretically predicted function $f_0(L)$ with its different regimes. Because the base-pair size is larger than the length of the backbone springs $(R>r_0)$, it is impossible to observe the regime of spring compression there. For very high torsion energy, the data points follow the prediction for the opposite limiting case, namely the function $f_\infty(L)$. For average torsion energy values the data point series interpolate systematically between the two limiting functions. Two features are independent of torsion energy: the Hookean asymptote for very large axial extensions and that the response force vanishes at the equilibrium axial extension $L=N_{seg}z_0$, where

$$z_0=\sqrt{r_0^2-R^2\sin^2(\theta_0)\sin^2(\phi_0/2)}\,. \tag{2.34}$$

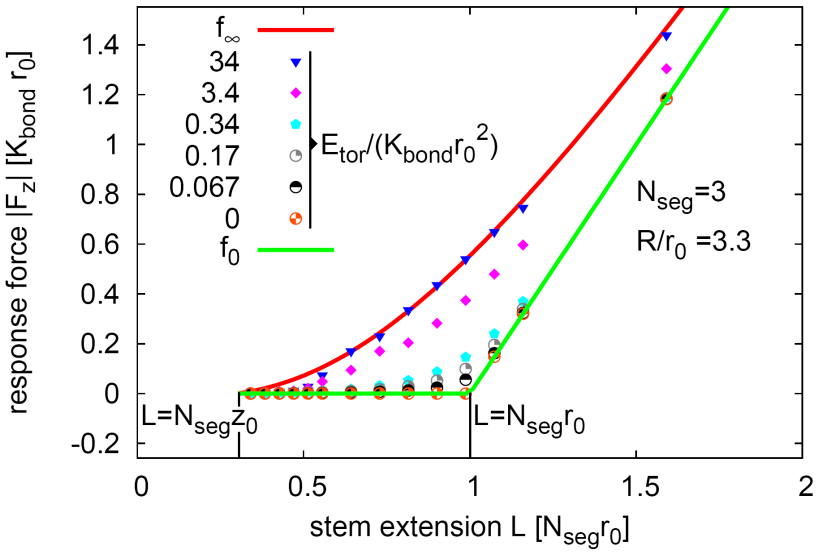


Fig. 2.12: Force extension relation for a folded hairpin when stretched axially.

2.3.2 Bending a hairpin

The total bending energy of the hairpin stem is distributed discretely over the joints between its elementary segments. The bending stiffness of each joint can be adjusted locally by varying the bending energy parameter. However, in order to calibrate the model by the use of experimental data (persistence length of folded stems in nucleic acids, which should be comparable to the results of Abels et al. [31], who used magnetic tweezers and atomic force spectroscopy: $P=55\pm2\,\mathrm{nm}$ for double-stranded DNA and $P=63.8\pm0.7\,\mathrm{nm}$ for double-stranded RNA), the effective bending stiffness of the whole stem needs to be measured. Is it simply proportional to the bending energy parameter, or are there anisotropic effects emerging from the complex geometry of the folded stem with all its potentials stabilizing lengths as well as angles? How does temperature influence this effective stem bending stiffness if the base-pairs at the bending joints are

bound more and more weakly due to thermal fluctuations and entropic forces become dominant?

The bending experiment is performed on a helical hairpin with a 4 base-pair stem folded according to the Go-model. The base-pair directions are perpendicular to the initially straight stem axis. (For a quantitative calibration of the persistence length by measuring the decay of tangent correlations in an 11-bp stem see chapter 2.4 .)

For simplicity, the closed stem can be considered as a chain of N_{seg} line segments with lengths z_{ij} (=stem chain), each of them connecting the centers of mass (=bending joints with inter-segmental bending angles θ_{ij}) of two subsequent stem base-pairs (Fig. 2.13). The direction of the loop-next segment and the position of its bending joint are kept fix during the bending procedure, yielding the z-axis and the origin of the coordinate system. An external harmonic force (spring stiffness K_{spring}) acting in x-direction on the first bead of the stem chain (=hairpin foot) tries to shift it from the z-axis into the x_0-plane:

$$\boldsymbol{F}_{spring} = -K_{spring}\left(x - x_0\right)\hat{\boldsymbol{x}} \tag{2.35}$$

For small x_0, stem stretching effects are negligible, so that the mean transverse elongation $\langle x \rangle$ of the already equilibrated stem originates purely from stem bending. Similarly, the mean external force is given by

$$\langle F_{spring} \rangle = -K_{spring}\left(\langle x \rangle - x_0\right) \tag{2.36}$$

The measured relation between the mean transverse elongation of the stem and the mean external force for several values of x_0 and/or K_{spring} allows to determine the transverse elasticity of the stem and from that its effective bending stiffness.

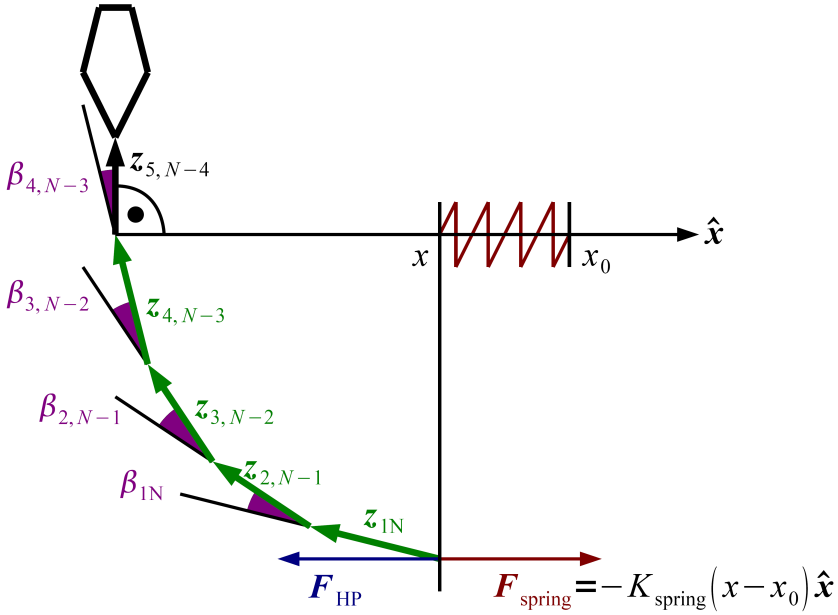


Fig. 2.13: Geometry of the bending experiment.

However, if x_0 and K_{spring} are chosen such that the resulting mean transverse elongations are large enough to stretch the stem segments considerably, the force elongation relation is not linear anymore (Fig. 2.14), at least for low temperatures. Therefore, in order

to determine the effective bending stiffness correctly one has to find the regime of small transverse elongations:

In this regime one observes a relation of linear response between the external force exerted on the hairpin foot and the resulting elongation. This linearity is maintained for all temperatures and all bending energy parameters, thus allowing the definition of an effective transverse spring constant of the hairpin stem from the inverse slope of the force elongation diagrams:

$$K_{\mathrm{HP}}(T, E_{\mathrm{bend}}) = -\frac{\Delta F_{\mathrm{spring}}}{\Delta x}\left(T, E_{\mathrm{bend}}; x \ll N_{\mathrm{seg}} z_0\right) \tag{2.37}$$

Fig. 2.15 shows such a diagram: For a bending energy parameter of $E_{\mathrm{bend}} = 1.25\, E_{\mathrm{LJ}}$ and temperatures ranging from $kT = 0.025\, E_{\mathrm{LJ}}$ to $kT = 2.5\, E_{\mathrm{LJ}}$ one obtains straight lines crossing the origin, indicating Hookean behavior for all temperatures. However, it is already clear from this diagram that $K_{\mathrm{HP}}(T, E_{\mathrm{bend}})$ depends on temperature in a non-trivial way.

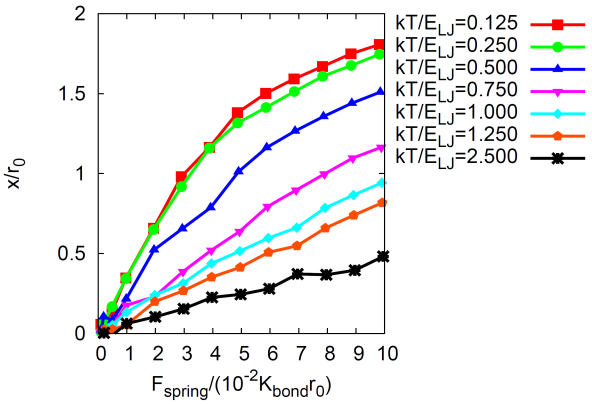


Fig. 2.14: Force elongation diagram for large elongations.

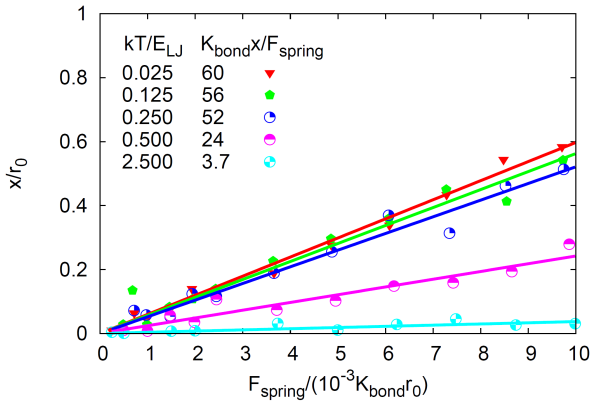


Fig. 2.15: Force elongation diagram for small elongations.

For low temperatures one obtains large slopes (small effective transverse spring constants) that are independent from temperature. For high enough temperatures, though, slopes decrease with increasing temperature. This fact is demonstrated more clearly by Fig. 2.16 (identical slopes for low temperatures) and Fig. 2.17 (slopes decreasing with increasing temperature for high temperatures), where a larger bending energy parameter, $E_{\mathrm{bend}} = 2.5\, E_{\mathrm{LJ}}$, is used. Obviously the dependence of $K_{\mathrm{HP}}(T, E_{\mathrm{bend}})$ from temperature is different in the low and the high temperature regimes.

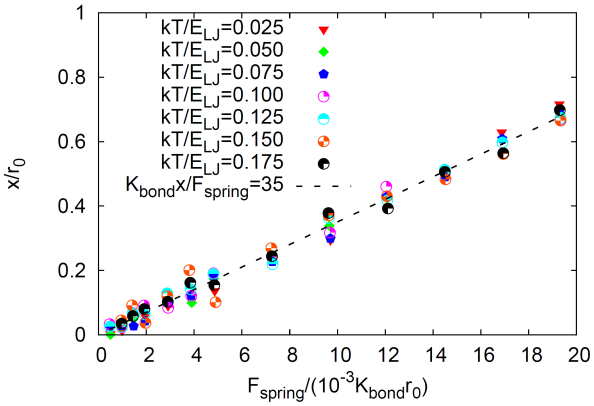


Fig. 2.16: Collapsing curves due to identical slopes for low temperatures.

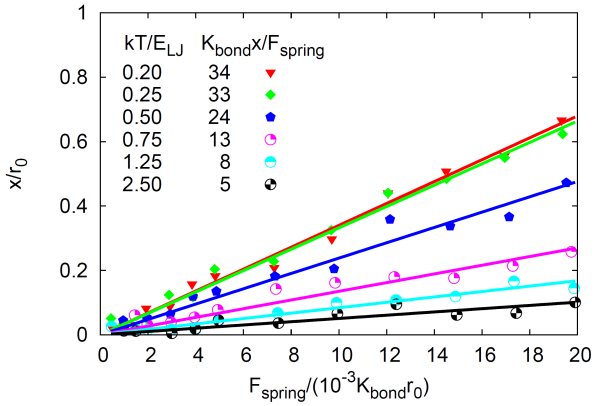


Fig. 2.17: Slopes decreasing with increasing temperature for high temperatures.

Furthermore, its dependence from the bending energy parameter differs substantially in the two regimes, as can be seen from Fig. 2.18 (low temperature regime) and Fig. 2.19 (high temperature regime). Again both figures show force elongation diagrams, but now temperature is kept fix within each diagram ($kT = 0.025\,E_{LJ}$ resp. $kT = 0.75\,E_{LJ}$), varying the bending energy parameter instead: $E_{bend} = 1.25\,E_{LJ}$, $E_{bend} = 2.5\,E_{LJ}$, $E_{bend} = 5\,E_{LJ}$. In the low temperature diagram one observes a decrease of the slope with increasing bending energy parameter, whereas in the high temperature diagram the slope is independent of the bending energy parameter.

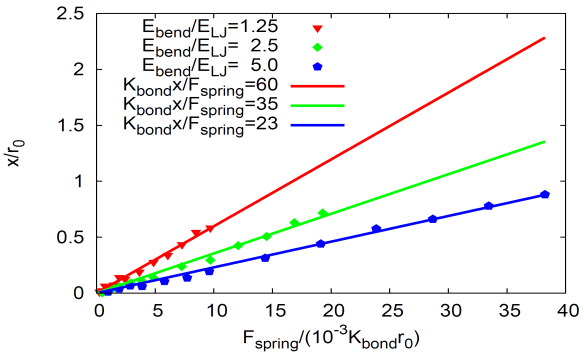


Fig. 2.18: Slopes are sensitive of the bending energy parameter for low temperatures.

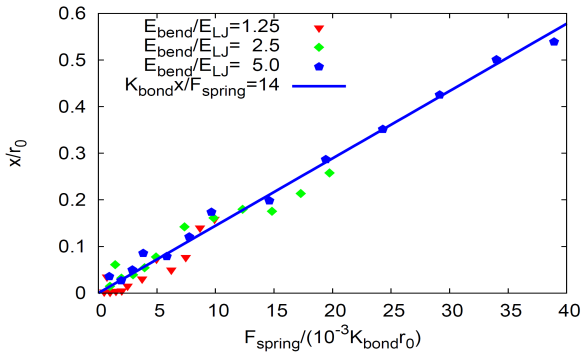


Fig. 2.19: Slopes are insensitive of the bending energy parameter for high temperatures.

How can the observed behavior of the inverse slope $K_{HP}(T, E_{bend})$ as a function of temperature and the bending energy parameter be explained? Comparing for several bending energy parameters, $E_{bend} = 1.25\,E_{LJ}$, $E_{bend} = 2.5\,E_{LJ}$, $E_{bend} = 5\,E_{LJ}$, the temperature dependence of the slope (cf. Fig. 2.20 and Fig. 2.21) with the temperature dependence of the probability that the hairpin is closed completely (Fig. 2.22) reveals that the difference between the low and the high temperature regime has structural reasons:

The melting temperature, above which the hairpin is totally unfolded due to thermal fluctuations, separates the low and the high temperature regimes. Below the melting temperature energy dominates over entropy, causing the hairpin to be closed and all base-pairs to be stable, which leads to maximal binding weights and well-defined bend-

ing joints. However, above the melting temperature entropy dominating over energy destabilizes all stem base-pairs, thus opening the hairpin and leaving it without well-defined bending joints (the binding weights are very small).

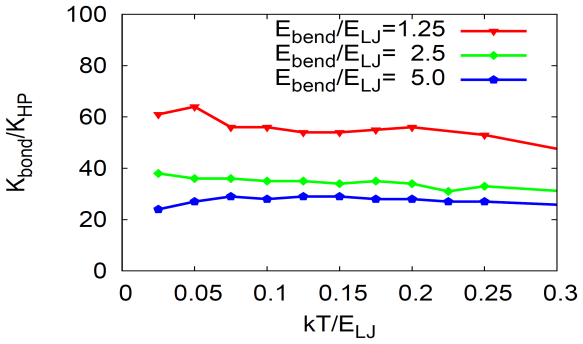


Fig. 2.20: Slope as function of temperature for several bending energy parameters (low temperature regime).

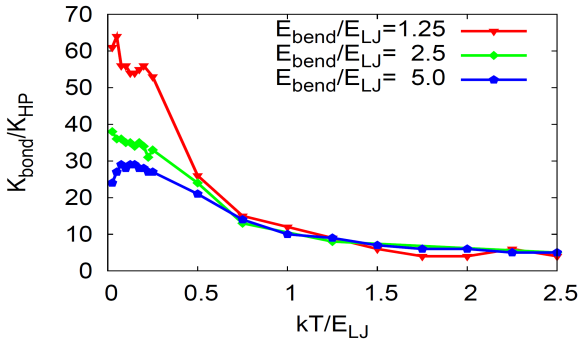


Fig. 2.21: Slope as function of temperature for several bending energy parameters (all temperatures).

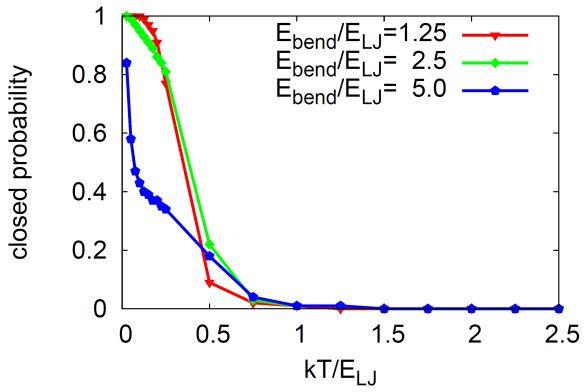


Fig. 2.22: Probability that the stem is closed as function of temperature for several bending energy parameters.

As a consequence, $K_{HP}(T, E_{bend})$ is a measure for the effective bending stiffness κ of the stem only below the melting temperature. Above this temperature its value is a mere artifact resulting from entropic forces (Although the RNA chain is totally unfolded, the fixation of its loop-next bases during the measurements causes a bias towards low transverse elongations of the hairpin foot. This bias increases with temperature, yielding in turn an increasing effective transverse spring constant.).

A priori one would expect $K_{HP}(T, E_{bend})$ to be proportional to the bending energy parameter. However, looking at the inverse slopes in Fig. 2.20 one observes deviations from this proportionality. The reason is that the non-vanishing torsion potential prevents the stem conformation from changing from helical to planar (This has already been the cause for the non-Hookean stretching behavior studied in 2.3.1.) during bending. This leads to a bending anisotropy, because for some base-pair orientations in the helix the need to deform backbone springs in order to bend the stem chain (Fig. 2.23 left) results

in an additional contribution to bending stiffness independent from the bending energy parameter. Only if the torsion potential vanishes, the stem is allowed to become planar during bending, making deformation of backbone springs unnecessary (Fig. 2.23 right) and the expected proportionality is restored (Fig. 2.24).

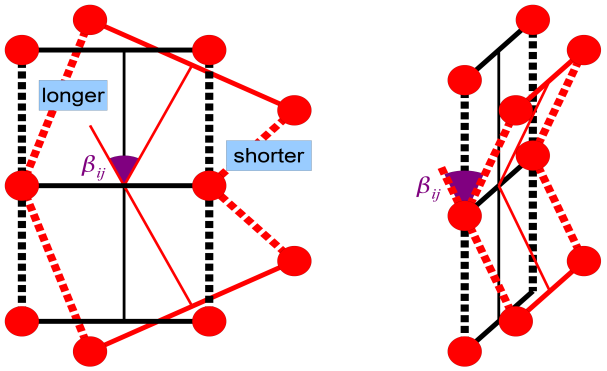


Fig. 2.23: Bending within the local stem plane enforces deformations of the backbone (left), whereas bending perpendicular to it does not (right).

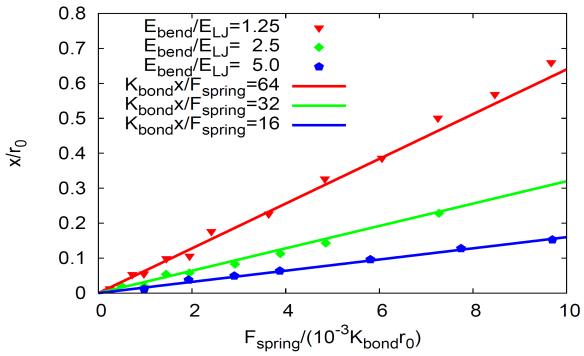


Fig. 2.24: A vanishing torsion potential restores the proportionality between the bending energy parameter and the effective bending stiffness of the whole stem.

In this chapter the qualitative behavior of a hairpin stem under bending forces has been studied, especially the complex interplay between bending, torsion and backbone potentials in producing the effective bending stiffness. The quantitative calibration of the bending energy parameter E_{bend} with respect to experiment will be carried out in chapter 2.4 .

2.3.3 Drilling a hairpin

The torsional resistance of hairpin stems in nucleic acids is well-studied in experiments: Zev Bryant and co-workers recently tested the linearity of DNA's twist elasticity and directly measured the torsional modulus ($C_{\text{tor}}=410\pm30\,\text{pN}\,\text{nm}^2$) [32]. In order to be able to adjust the model to these experimental results it is necessary to perform simulations similar to these experiments. In a first qualitative model experiment, a completely folded, regularly helical 4-base-pair stem with a straight axis, to which all base-pairs are perpendicular, is used (Later, in chapter 2.4 , the torsional modulus of an 11-bp hairpin stem will be measured directly from the thermal fluctuations of the total rotation angle of the stem helix.): It is drilled initially by rotating the loop-near and the loop-far base-pair around the stem axis - out of their equilibrium positions - into opposite directions by equal amounts, thus gaining a total drill angle $\Delta\Phi$. Then the stem is fixed in space at its loop-near end by keeping fix the corresponding base-pair. Fig. 2.25 shows such a hairpin before drilling (above) and after drilling it by a positive total angle (below).

Different from the torsion energy parameter ($E_{tor}=0.25\,E_{LJ}$), whose effect is to be studied here, both the bending and the tilt energy parameters are set to zero.

During the simulation the total drill angle $\Delta\Phi$ between the loop-far base-pair and the (spatially fixed) loop-near base-pair is kept constant. As soon as the simulation is started, the stem begins to equilibrate internally by rotating base-pairs and changing its axis length. This equilibration process endures until all single drill angles $\Delta\phi$ belonging to the N_{seg} elementary cells of the hairpin stem are equal on average (cf. Fig. 2.26):

$$\Delta\Phi = N_{seg}\,\Delta\phi = N_{seg}(\phi-\phi_0) \tag{2.38}$$

As Fig. 2.26 shows, the resulting torque is obtained by measuring the torsion forces $\boldsymbol{F}_{tor}$ that try to rotate back the base-pairs (vectors $\boldsymbol{r}$) at the loop-near and the loop-far end of the stem:

$$\boldsymbol{\tau} = \boldsymbol{r}\times\boldsymbol{F}_{tor} \tag{2.39}$$

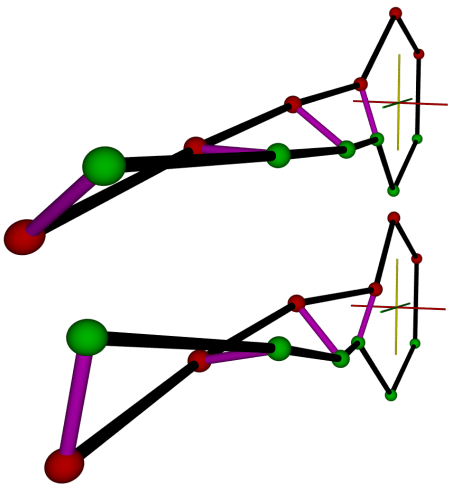


Fig. 2.25: 3D representation of the hairpin used in the torsion experiment (figure above: zero total drill angle, figure below: positive total drill angle).

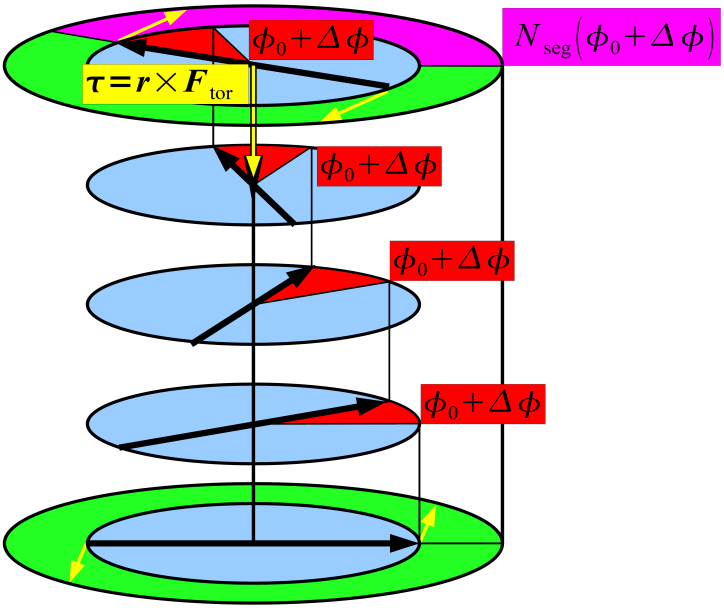


Fig. 2.26: The drill angles from the elementary cells of the stem contribute additively to the total drill angle. The torque within elementary cells compensates for stems equilibrated at constant total drill angle. The response torque can be measured at the ends of the stem.

The effect the equilibration process of the stem has on the time-dependence of the axial component $\tau_{\|}$ of this torque can be seen from Fig. 2.27, where this dependence is shown at zero temperature for several total drill angles $\Delta\Phi$ (All non-axial components vanish, data not shown.). The long-time horizontal asymptotes define the values of the (axial) response torque after equilibration. These values are plotted versus their corresponding total drill angles in Fig. 2.28, each curve belonging to a distinct temperature. For sufficiently low temperatures ($T\ll T_{melt}$), all base-pairs are bound strongly and the torsion

potential contributions $U_{\text{torsion}}^{\text{seg}} = -E_{\text{tor}} \cos(\phi - \phi_0)$ from the N_{seg} elementary cells yield the expected sine-shaped behavior of the axial torque after equilibration:

$$\tau \equiv \tau_{\parallel}^{\text{equ}} \equiv \left(\boldsymbol{r} \times \boldsymbol{F}_{\text{tor}}\right)_{\parallel}^{\text{equ}} \approx -\frac{\partial U_{\text{torsion}}^{\text{seg}}(\phi)}{\partial \phi} = -E_{\text{tor}} \sin\left(\Delta\Phi / N_{\text{seg}}\right) \tag{2.40}$$

For high temperatures, however, one obtains

$$\tau \approx -\alpha(T)\, E_{\text{tor}} \sin\left(\Delta\Phi / N_{\text{seg}}\right) \tag{2.41}$$

with an sine amplitude $\alpha(T) < 1$ that decreases considerably with increasing temperature because of the loss of torsion joints if the base-pairs are bound more and more weakly.

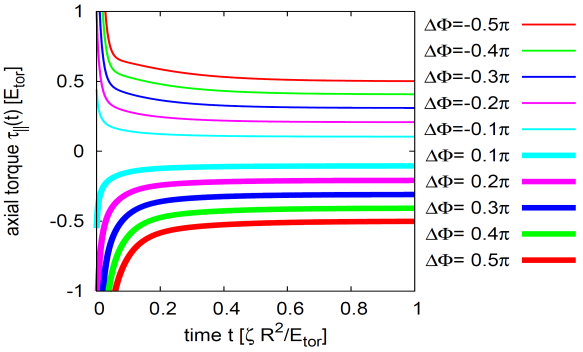


Fig. 2.27: Axial component of the torque over time for different constant total drill angles (positive and negative drill). Note the asymptotic long-time behavior after the process of stem equilibration.

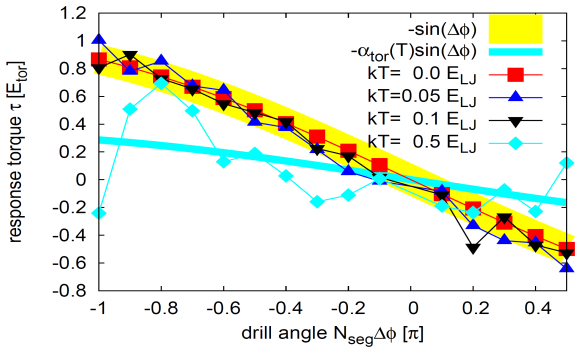


Fig. 2.28: Equilibrium axial torque as a function of the total drill angle, parametrized by temperature. All curves at low temperatures obey the same sine-law, for high temperatures the amplitude decreases, however.

To summarize, for temperatures so low that the base-pairs are stable the torsional resistance of the model hairpin stem is directly given by the torsion energy parameter. For high temperatures, the base-pairs begin to melt, yielding a much lower torsional resistance. The quantitative calibration of the torsion energy parameter E_{tor} with respect to experiment will be carried out in chapter 2.4 .

2.4 Quantitative calibration of the RNA model

Up to now, the highly entangled interplay of backbone bond, helix torsion and helix bending potentials has been analyzed by simulated mechanical stretching, twisting and bending of a short hairpin. The goal is now to exactly calibrate the energy parameters belonging to those potentials in a way that the torsional modulus and the persistence length of a 11-base-pair model hairpin are comparable to experimental data. This will be done by equilibrium measurements instead of mechanical deformation by external forces.

First of all, the **geometric properties** of the hairpin, which consists of an 11-bp stem ($N_{seg}=10$) and a 4-base loop, have to be set, i.e. its characteristic equilibrium length ratios and equilibrium angles:

$$\frac{R}{r_0}=3,\ \ \phi_0=\frac{2\pi}{11},\ \ \theta_0=0.39\pi \tag{2.42}$$

First the helix twist ϕ_0 and the base-pair tilt θ_0 are set to their experimentally measured values directly and the base-pair distance R is set three times the bond length r_0. Then the bond length is chosen such that the resulting rise z_0 per base-pair (in Å) equals experiment ($r_0=4.2$ Å). This is done by making use of the stem cell Pythagoras relation, which describes equilibrium stem geometry in this coarse-grained model:

$$r_0^2=z_0^2+R^2\sin^2(\theta_0)\sin^2(\phi_0/2) \tag{2.43}$$

In this geometry the helix axis contains all base-pair centers of mass. However, in real A-RNA stems these centers of mass form a helix around the stem axis themselves, by that making the diameter of real A-RNA larger than in the coarse-grained model, where it is given by:

$$d_{stem}=R\sin(\theta_0) \tag{2.44}$$

Apart from that, comparing the geometric parameters of the model with experiment yields geometric properties matching perfectly (Fig. 2.29):

		Simulated RNA translocation	RNA/DNA parameters [76]	A-RNA parameters	A-DNA parameters [94]
Bond length	r_0[Å]	4.2	/	4.3 [95]	/
Base-pair distance	R[Å]	12.6	/	/	/
Rise per base-pair	z_0[Å]	2.52	2.52	2.81 [76]	2.4
Helix diameter	d_{stem}[Å]	11.9	/	22 [95]	26
Helix twist	ϕ_0[°]	32.7	33.3	32.7 [76]	33.6
Base-pair tilt	$\left\|\frac{\pi}{2}-\theta_0\right\|$[°]	19	19.3	16.7 [76]	19

Fig. 2.29: Geometric parameters of RNA.

Having implemented the hairpin geometry correctly, the next step is to calibrate the **energy parameters** of the model. This is done by measuring geometric properties describing the whole stem, such as the fluctuations of the total helix rotation angle or the mean cosine of the angle between the vectors of the first and the last tangent of the helix contour.

The necessary definitions read:

- Torsional modulus $\frac{C_{\mathrm{tor}}}{\mathrm{k}\,T}=\frac{\langle L\rangle}{\langle \Phi^2\rangle-\langle\Phi\rangle^2}$

  with total helix rotation angle $\Phi=\sum_{k=1}^{N_{\mathrm{seg}}}\phi_k$ and contour length $L=\sum_{k=1}^{N_{\mathrm{seg}}} z_k$

- Persistence length per segment length or number of persistent segments: $s_p=-\frac{|i-j|}{\ln\langle \hat{\boldsymbol{z}}_i\cdot\hat{\boldsymbol{z}}_j\rangle}$ where $i=1,\ j=N_{\mathrm{seg}}$

In these definitions $\langle\,\rangle$ denotes a time average over one single trajectory.

The energy scale of the model is set to body temperature ($\mathrm{k}T=4.28$ pN nm), its length scale has been set earlier to $r_0=4.2$ Å. Base-pair binding energy and single-stranded stretching modulus are experimentally known:

- Mathews et al. [96] summarize known thermodynamic parameters for RNA, which have been obtained by performing calculations on experimental results. This yields the following free energy per base-pair:

$$G_{\mathrm{bind}}^{\mathrm{RNA\text{-}bp}}\approx 2.5\ \mathrm{kcal\ mol^{-1}}=17\ \mathrm{pN\ nm}=4\ \mathrm{k}\,T \tag{2.45}$$

- Stretching experiments with single-stranded DNA carried out by Smith, Cui and Bustamante [97] yield the following stretching modulus:

$$S_{\mathrm{stretch}}^{\mathrm{ssDNA}}\approx 800\ \mathrm{pN} \tag{2.46}$$

Therefore, the calibration starting point is:

$$\frac{E_{\mathrm{LJ}}}{\mathrm{k}\,T}=4,\quad \frac{K_{\mathrm{bond}}r_0^2}{\mathrm{k}\,T}=80 \tag{2.47}$$

Sufficient stability against base-pair tilt and axial shearing of the stem is obtained by setting $E_{\mathrm{tilt}}/\mathrm{k}\,T=4$. In order to get realistic values for the persistence "length" s_p and the torsional modulus C_{tor}, which are comparably high, one has to increase both the torsion and bending energy parameters (E_{tor}, E_{bend}) appreciably.

However, too high values of these energy parameters destabilize the stem, thereby causing negligible persistence length and torsional modulus – unless the backbone springs are hard enough. This causal connection between backbone stability, base-pair stability and angular stability is shown in Figs. 2.30 (torsional modulus vs. angular energy parameters), 2.31 (persistence length vs. angular energy parameters) and 2.32 (stem integrity vs. angular energy parameters), where the backbone spring stiffness is increased gradually ($D_{\mathrm{bond}}\equiv K_{\mathrm{bond}}r_0^2/\mathrm{k}\,T$). Obviously, the model's backbone spring stiffness has to be corrected by a factor of at least 20:

$$\frac{K_{\mathrm{bond}}r_0^2}{\mathrm{k}\,T}=1600 \tag{2.48}$$

Then one can set $E_{\mathrm{tor}}/\mathrm{k}\,T=E_{\mathrm{bend}}/\mathrm{k}\,T=160$ without loosing stem stability, which results in realistic values for persistence length and torsional modulus.

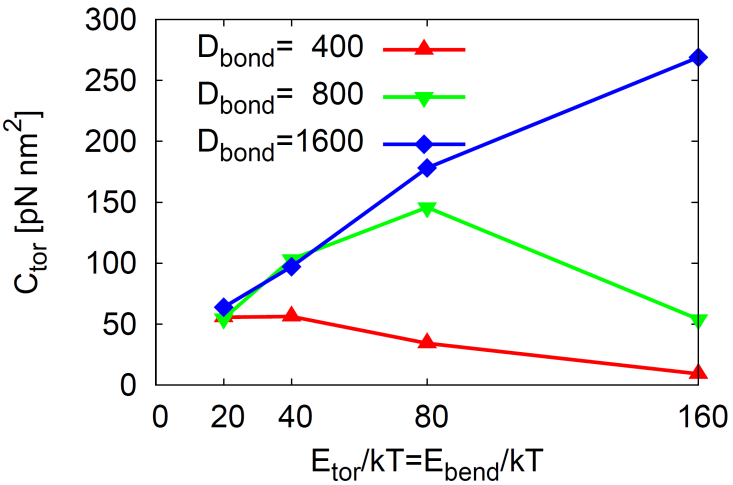


Fig. 2.30: The torsional modulus of the model RNA hairpin stem increases monotonically with the torsion energy parameter only below a critical value. The harder the backbone springs are, the higher this critical value is.

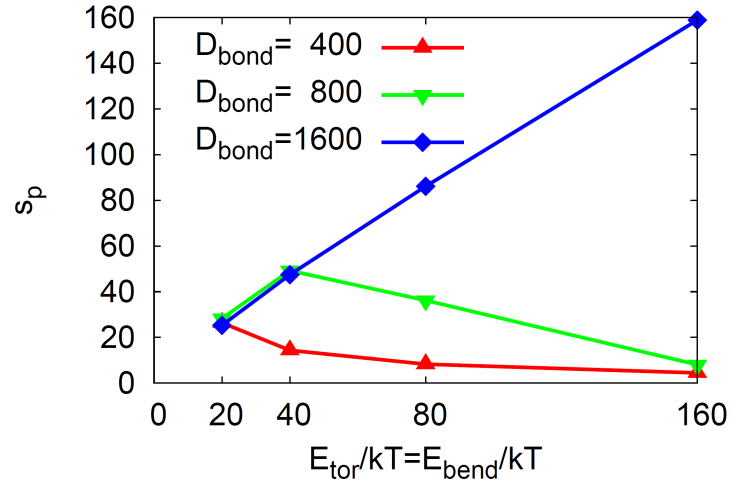


Fig. 2.31: The persistence length (per segment length) of the model RNA hairpin stem increases monotonically with the bending energy parameter only below a critical value. The harder the backbone springs are, the higher this critical value is.

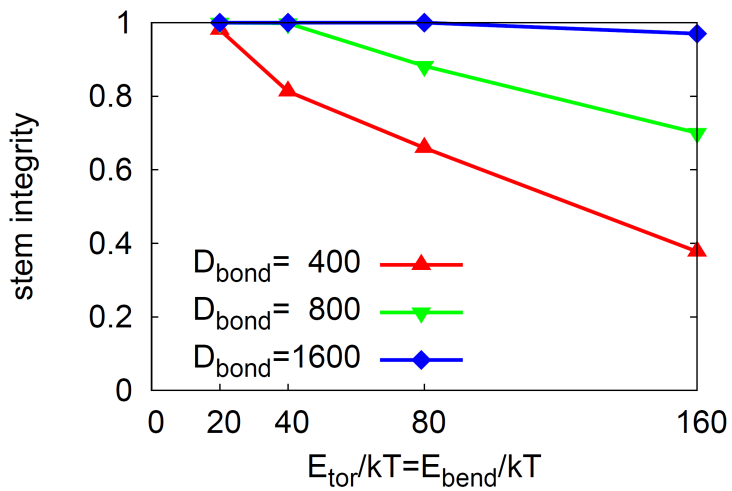


Fig. 2.32: The base-pair stability of the model RNA hairpin stem decreases with increasing torsion and bending energy parameters. This destabilization effect is the weaker, the harder the backbone springs are, and causes monotonicity breaking of torsional modulus and persistence length for high torsion and bending energy parameters.

In order to further validate this result, 40 independent runs with 4 millions of steps each are performed with the same geometric and energy parameters as above (case 1 of table 2.33) and compared with contrasting parameter sets (case 2, 3 and 4) as well as with experiment:

$$\frac{E_{\mathrm{LJ}}}{\mathrm{k}T}=4,\ \frac{K_{\mathrm{bond}}r_0^{\ 2}}{\mathrm{k}T}=1600,\ \frac{E_{\mathrm{tilt}}}{\mathrm{k}T}=4 \tag{2.49}$$

	Case 1	Case 2	Case 3	Case 4	
	$E_{\mathrm{tor}}/\mathrm{k}T=160$ $E_{\mathrm{bend}}/\mathrm{k}T=160$	$E_{\mathrm{tor}}/\mathrm{k}T=0$ $E_{\mathrm{bend}}/\mathrm{k}T=160$	$E_{\mathrm{tor}}/\mathrm{k}T=160$ $E_{\mathrm{bend}}/\mathrm{k}T=0$	$E_{\mathrm{tor}}/\mathrm{k}T=0$ $E_{\mathrm{bend}}/\mathrm{k}T=0$	Experiment
$C_{\mathrm{tor}}/\mathrm{k}Tr_0$	158±7	5.44±1.12	133±3	1.83±0.10	220 [32]
C_{tor} [pN nm^2]	285±13	9.87±2.01	235±6	3.29±0.18	400 [32]
s_p	155±5	167±5	4.01±0.56	3.39±0.54	240 [31]
$\langle\hat{\boldsymbol{z}}_i\cdot\hat{\boldsymbol{z}}_j\rangle$	0.9436±0.0018	0.9476±0.0014	0.107±0.032	0.072±0.029	0.96 [31]
					ideal
$\langle\Phi/N_{\mathrm{seg}}\rangle$ [°]	31	0.52	32	-0.34	32.7 [76]
$\langle L/N_{\mathrm{seg}}\rangle$ [Å]	2.7	3.4	2.9	3.4	2.52 [76]
$\langle\cos(\theta)\rangle$	-0.30	-0.37	-0.04	-0.29	-0.33 [76]
Stem integrity	0.97	1.00	1.00	0.98	/

Fig. 2.33: Torsional modulus and persistence length for correctly strong torsion and bending energy parameters (case 1) and contrasting test simulations in which one or both energy parameters vanish (case 1 to 3). Only in case 1 both torsional modulus and persistence length are comparable with experiment.

Case 1 is the parameter set that causes both persistence length and torsional modulus to be in the same range as the corresponding experimental values, case 2 to 4 are only tests of significance, where either the torsion potential, the bending potential or both are switched of. One observes that persistence length and torsional modulus depend significantly on bending and torsion energy parameters respectively. The correct parameter set (case 1) furthermore leads to high stem stability and approximately ideal total helix rotation angle, stem contour length and inclination cosine.

To conclude, the parameter set

$$\frac{E_{\mathrm{LJ}}}{\mathrm{k}T}=4,\ \frac{K_{\mathrm{bond}}r_0^2}{\mathrm{k}T}=1600,\ \frac{E_{\mathrm{tor}}}{\mathrm{k}T}=160=\frac{E_{\mathrm{bend}}}{\mathrm{k}T},\ \frac{E_{\mathrm{tilt}}}{\mathrm{k}T}=4,\qquad \frac{R}{r_0}=3,\ \phi_0=\frac{2\pi}{11},\ \theta_0=0.39\,\pi \tag{2.50}$$

causes realistic helix geometry (obtained from X-ray RNA crystallography data summarized by Holbrook and Kim [76]), realistic persistence length (obtained for double-stranded RNA and DNA from magnetic tweezer experiments and atomic force microscopy by Abels and co-workers [31]), realistic torsional modulus (obtained from the experiments by Zev Bryant and co-workers, who measured torque vs. twist in over- and underwound

double-stranded DNA molecules with the help of rotating beads [32]) and sufficient shear stability.

However, it does not reproduce the base-pair melting behavior correctly. The reason is an effective increase of base-pair binding energy for large angular energy parameters: If all angles are near their equilibrium values, the cosines in the corresponding angular potential contributions are near unity. Then the base-pair binding predicates (H-functions) in these angular potential contributions generate attractive forces which are the larger, the larger the corresponding angular energy parameter is. For torsion and bending energy parameters so large as in this parameter set, these attractive forces appreciably increase the effective base-pair binding energy, thereby preventing the stem from melting at realistic temperatures or from being unzipped by realistic pulling forces.

One could possibly solve this problem of base-pair stability changed by strong angular potentials by shifting the level of zero energy of the angular potential contributions. This is done by varying a_0 away from zero up to unity in the following terms:

$$H_i^j H_{i+1}^{j-1} E_{\text{tor}} [a_0 - \cos(\phi_{ij} - \phi_0)] \tag{2.51}$$

$$H_i^j H_{i+1}^{j-1} H_{i+2}^{j-2} E_{\text{bend}} [a_0 - \cos(\beta_{ij})] \tag{2.52}$$

$$H_i^j H_{i+1}^{j-1} H_{i+2}^{j-2} E_{\text{tilt}} [a_0 - \cos(\theta_{ij} - \theta_0)] \tag{2.53}$$

For a_0=0 equilibrium angles $(\phi_0, \beta_0$=0,$\theta_0)$ cause base-pair stabilizing forces, for a_0=1 these forces vanish. However, for all other angles a_0=1 leads to base-pair destabilizing forces, which is a strong bias towards unstable stems, whereas a_0=0 balances in angle-space between stabilizing and destabilizing forces (cf. Fig. 2.34 for the torsion potential). Therefore, this model stays with a_0=0 as long as the angular energy parameters are small, despite the resulting base-pair stabilization effects. For high angular energy parameters a possible solution could be setting a_0=1 and compensating the resulting base-pair destabilization effects by increasing $E_{\text{LJ}}/\text{k}T$.

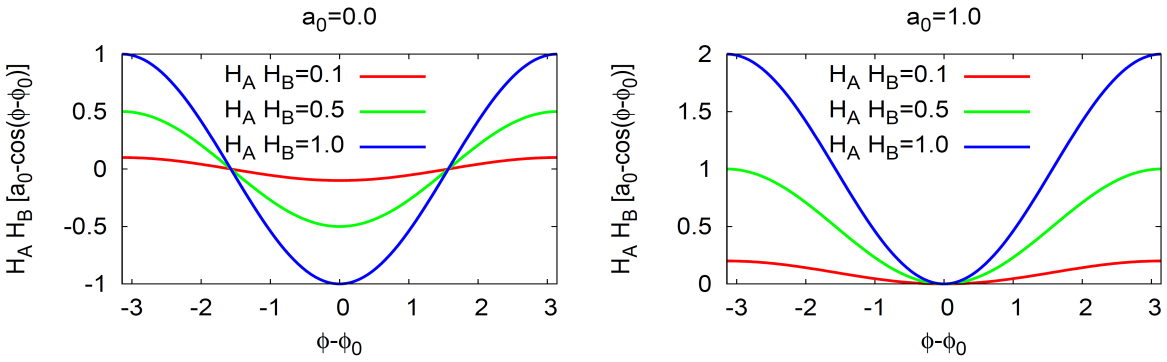


Fig. 2.34: Torsion potential contribution for several levels of base-pair binding. The parameter a_0 controls for which ranges of twist angles the contribution produces base-pair stabilizing and destabilizing forces, respectively. For a_0=0.0 destabilization and stabilization range are equally-sized and the largest stabilization effect is obtained for the equilibrium angle. For a_0=1.0 only the equilibrium angle is neutral, all other angles cause destabilization.

Because for studying translocation of RNA molecules through narrow pores, which is the goal of the following chapters, one has to rely on correct base-pair melting and unzipping behavior, a **different parameter set** will be used:

$$\frac{E_{\mathrm{LJ}}}{\mathrm{k}T}=4,\ \frac{K_{\mathrm{bond}}r_0^2}{\mathrm{k}T}=400,\ \frac{E_{\mathrm{tor}}}{\mathrm{k}T}=4=\frac{E_{\mathrm{bend}}}{\mathrm{k}T},\ \frac{E_{\mathrm{tilt}}}{\mathrm{k}T}=4,\qquad \frac{R}{r_0}=3,\ \phi_0=\frac{2\pi}{11},\ \theta_0=0.39\,\pi \tag{2.54}$$

Here, torsion and bending energy parameters are much lower than above, yielding too low torsional modulus and persistence length. However, this disadvantage is inevitable (for $a_0=0$) if base-pair melting and unzipping should be realistic. As the angular potentials are weak enough within this parameter set, their base-pair stabilizing effect plays only a negligible role.

2.5 Translocation of hairpins and pseudoknots

In the previous chapters the static and dynamic properties of isolated model RNA molecules have been investigated. Now, the model is extended by a narrow pore with local drift potential in order to study the translocation properties of these molecules. The (locally repulsive) pore plane separates space into a *cis* and a *trans* side and the pore channel is short enough to harbor only one RNA base at a time and too narrow to allow double-stranded translocation without major deformation of the folded RNA molecule. Of special interest from a biologist's as well as from a physicist's point of view is the mechanism causing highly different rates of RNA translocation through ribosomes for pseudoknot motifs on the one hand and hairpin motifs on the other hand: Whereas pseudoknots may induce ribosomal pausing and in turn ribosomal frameshift, the hairpin translocation rate is comparably high. Therefore the main attention in the following chapters will lie on a comparative study of the translocation behavior of these two RNA motifs, especially taking into account effects of geometry and binding topology.

2.5.1 Structure of hairpins and pseudoknots

An RNA hairpin is a secondary structure element consisting of one Stem (*S*) and one loop (*L*), which are arranged along the RNA chain from its 5' to its 3' end according to the sequence

$$s_{\mathrm{HP}}=(S)_i L (S)_j \,. \tag{2.55}$$

However, the most prominent RNA folding motif involved in ribosomal pausing and thus in -1 programmed ribosomal frameshift is the pseudoknot. Pseudoknots, which were identified in RNA first by Rietveld and coworkers [98][99] are ubiquitous RNA tertiary structure elements [9], occurring not only in messenger RNA [100][101][102][103][104][105], but also in ribosomal RNA [106], transfer-messenger RNA [107], catalytic and self-splicing RNA [108][109][110], ribonucleoprotein complexes [111] and viral genomic RNA (reviewed in [112]).

Among all the different pseudoknot topologies the particularly simple H-type pseudoknot, which often plays the role of a stimulatory element for ribosomal pausing and frameshift during the translation of messenger RNA, is most common [113]. It consists of two stems (S_A, S_B) and two loops (L_1, L_2). The corresponding sequence is

$$s_{PK}=(S_A)_i L_1 (S_B)_i (S_A)_j L_2 (S_B)_j \,. \tag{2.56}$$

In both the hairpin and the pseudoknot structure a loop is a sequence of unpaired bases connecting regions of the chain where bases are paired. On the other hand, a stem is a cylindrical structural motif (axis $\hat{\boldsymbol{z}}$) consisting of two complementary right-handed backbone helices (index *i* for the helix nearer to the 5' end, index *j* for its counterpart) connected to each other by base-pairs.

The structure of frameshifting pseudoknots has been studied extensively [114][115][116]. The two stems of an H-type pseudoknot are stacked onto each other, forming a quasi-continuous double-helical structure [117], which – as NMR studies show – may be bent in some cases [118]. Whereas the 3' and 5' end strands of a hairpin point into the same direction, the end strand directions in an H-type pseudoknot are opposite, a topological difference proposed to be relevant as far as the ability to provoke ribosomal pausing is concerned [119][112]: Even if the total numbers of base-pairs in a pseudoknot and a hairpin structure are equal, only the pseudoknot increases frameshift efficiency appreciably [7].

In order to make clear the three-dimensional structures of both hairpin and pseudoknot, Fig. 2.35 resp. Fig. 2.36 visualize the special symmetric case where there are 10 base-pairs in the hairpin stem and 5 base-pairs in each pseudoknot stem.

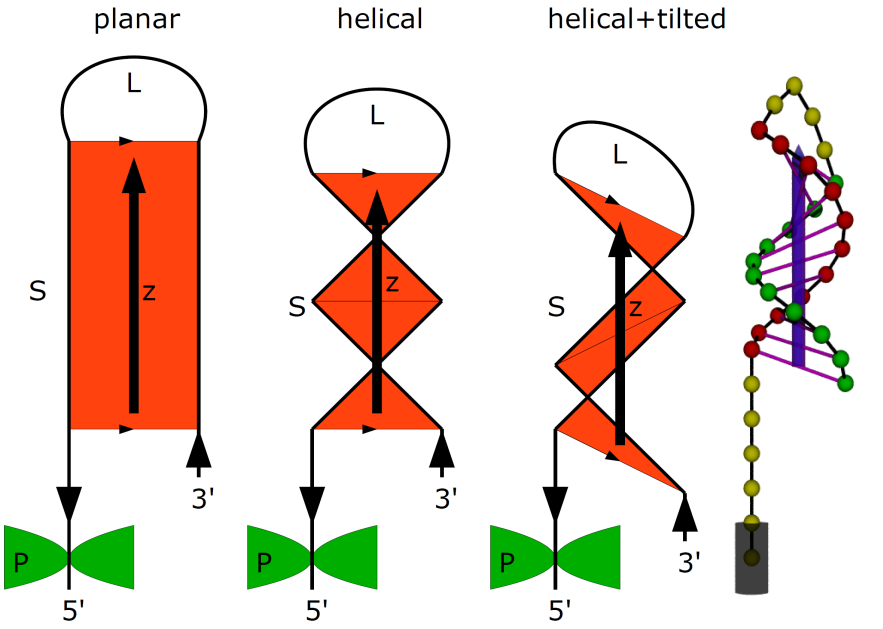


Fig. 2.35: Schematic view of the 3D geometry of a hairpin with loop L and stem S, whose 5′ end is threaded in a nanopore P. The stem is helical: its base-pairs perform one total rotation around the stem axis z. The rightmost figure shows the initial geometry for the simulations.

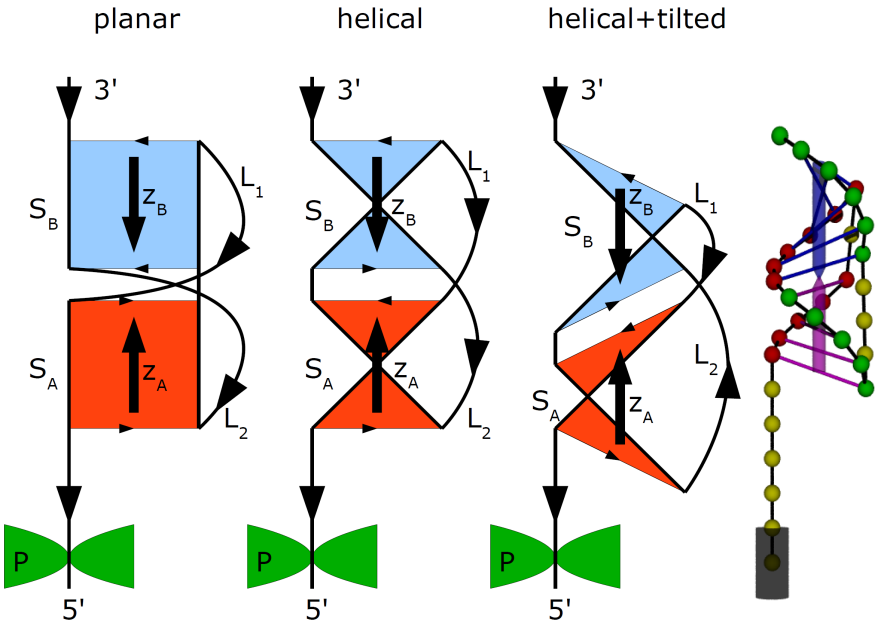


Fig. 2.36: Schematic view of the 3D geometry of a H-type pseudoknot with two loops L_1, L_2 and two stems S_A, S_B, whose 5′ end is threaded in a nanopore P. The stems are helical: their base-pairs perform a half rotation around each of the stem axes z_A, z_B. The rightmost figure shows the initial geometry for the simulations.

This causes a full rotation of the base-pair direction around the hairpin stem axis and half rotations around the axes of each pseudoknot stem. For both hairpin and pseudoknot this is shown by the comparison between the planar (first column in Fig. 2.35 and Fig. 2.36) and the helical representation (second column). The two pseudoknot stems connected to each other at their heads complement one another by forming a right-handed double helix that continues at the connection site and thus encompasses the whole pseudoknot. Adding the feature of base-pair inclination ($\theta_0 > \pi/2$, third column) to the 3D representation scheme, one recognizes that the loop length symmetry observed in the planar (first column) and helical (second column) representations of the pseudoknot geometry disappears: For inclination angles larger than 90 degrees loop L_1 has to be shorter than loop L_2.

The previous geometrical contemplations help to construct three-dimensional and fully folded initial configurations for hairpins and pseudoknots that are near to mechanical equilibrium (cf. rightmost column in Fig. 2.35 and Fig. 2.36). After some further equilibration steps with the chain's 5' end kept fix within the channel of the narrow pore (denoted *P* in the figures) the translocation process of such RNA structures can start.

2.5.2 Characterization of the system using the hairpin paradigm

Before the translocation process of a pseudoknot can be compared to that of a hairpin the behavior of the system "pore – RNA chain" has to be characterized. This will be done for the simpler case of the RNA hairpin, from which one can then abstract results easily to the more complex pseudoknot case. Before the hairpin translocation process and its various time regimes are studied in detail it is most important to answer the questions which types of translocation behavior exist and – because of the complexity of the model with its large collection of energy parameters ($E_{\mathrm{drift}}, E_{\mathrm{LJ}}, kT, E_{\mathrm{tilt}} \ldots$) – which parameter ratios are relevant for the characterization of the translocation behavior. Having characterized this behavior with respect to these energy ratios it will be easy to find the appropriate regime in parameter space to perform the hairpin-pseudoknot comparison.

2.5.2.1 Exploration of the parameter space: finding the relevant regime

In order to study the parameter space for hairpin translocation by variation of temperature T and pore drift energy E_{drift}, a regularly folded hairpin ($\phi_0 = 2\pi/11,\ \theta_0 = 1.9,\ R = 3r_0$) with an 11-bp stem and a 4-base loop is initially threaded into the pore channel ($l_{\mathrm{pore}} = 2r_0,\ d_{\mathrm{pore}} = r_0,\ E_{\mathrm{barrier}} = 500\,E_{\mathrm{LJ}}$) with its first base, which is connected to the stem by a chain of 6 bases. The energy parameters of the angular potentials are equal to the Lennard-Jones energy parameter ($E_{\mathrm{tor}} = E_{\mathrm{bend}} = E_{\mathrm{tilt}} = E_{\mathrm{LJ}}$). As usual, the backbone springs are strong ($K_{\mathrm{bond}} r_0^2 = 100\,E_{\mathrm{LJ}}$), the base-pairs form according to the Go model with low binding threshold ($\delta = 0.1$) and the repulsion between incompatible bases is weak ($E_{\mathrm{LJ}}^{\mathrm{incomp}} = 10^{-4}\,E_{\mathrm{LJ}}$).

A first subdivision of parameter space is shown in Fig. 2.37, where for high thermal energies kT (compared to the energy gain per translocated base E_{drift}) the chain is rejected from the pore into backward direction ("rejected, -") instead of being translocated through the pore channel. This effect is due to entropic forces outweighing the pore drift

force: The deeper the chain is in the pore the less is its conformational entropy. Since this entropy gradient between chain configurations differing from each other only by one translocated base is largest at the begin (and the end) of translocation, it causes an initial rejection of the chain if temperature is so high that $\Delta S(T)T > E_{\text{drift}}$. This explains qualitatively the shape of the monotonically increasing coexistence line between the regime with chain rejection and the one without (Fig. 2.37, "coex. with (-)").

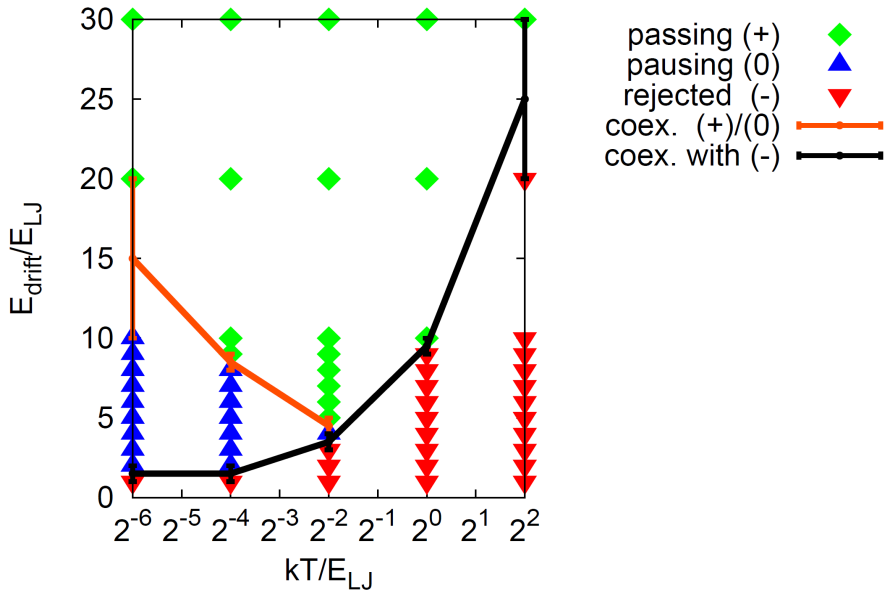


Fig. 2.37: Regimes of chain passing, chain pausing and chain rejection with their respective coexistence lines in parameter space.

However, the main interest lies in the complementary non-rejection regime, which in turn consists of two sub-regimes, as one can see from Fig. 2.37. In the regime of single-stranded translocation ("passing, +") the stem is opened totally in order to allow all bases to pass the narrow pore. By contrast, in the pause regime ("pausing, 0") the stem remains closed causing the translocation process to stop as soon as the stem, which is not thin enough to pass the pore in its closed state, reaches the pore (cf. Fig. 2.48 later).

The ability of the pore to open the whole stem in the course of single-stranded translocation depends on the rate for breaking up the base-pair containing the base immediately in front of the pore and subsequently translocating this base through the pore channel. For simplicity one can assume that this rate is – in zeroth order – a monotonic function of the ratio

$$g(E_{\text{drift}}/E_{\text{LJ}}, kT/E_{\text{LJ}}) = \frac{\alpha_{\text{bind}} E_{\text{LJ}} - E_{\text{drift}}}{kT} \tag{2.57}$$

(where $\alpha_{\text{bind}} > 1$ is a dimensionless factor accounting for contributions to the base-pair binding energy that do not originate from the Lennard-Jones potential but from the effectively attractive binding predicates H_i^j in the angular potentials). This assumption uses the fact that in order to unbind and translocate one base an energy barrier has to

be crossed which increases with the base-pair binding energy and decreases with the pore drift energy. If the ratio g is greater than some constant g_0, one is supposed to observe chain pausing, else chain passing. And indeed this order of regimes is observed in simulation, moreover, the coexistence line between the passing and pausing regimes in Fig. 2.37 ("coex. (+)/(0)") is monotonically decreasing, just as predicted from the assumption (2.57) by setting $g = g_0$:

$$\left(\frac{E_{\text{drift}}}{E_{\text{LJ}}}\right)_{\text{coex}} = \alpha_{\text{bind}} - \frac{kT}{E_{\text{LJ}}} g_0 \tag{2.58}$$

To conclude, as long as the deformability of the stem is low, one observes three regimes, separated by coexistence lines in parameter space. The coexistence line between the rejection and the non-rejection regimes is determined by the competition between entropic forces and the pore force. The coexistence line between the two non-rejection sub-regimes of single-stranded translocation and chain pausing can be explained as a competition between an energetic barrier, which is increased by the base-pair binding energy and decreased by the pore drift energy, on the one hand and thermal fluctuations on the other hand.

But what if the stem deformability cannot be assumed low? In that situation a fourth regime emerges, the regime of double-stranded translocation. To observe it, it is sufficient to use a relatively low temperature ($kT/E_{\text{LJ}} = 2^{-3}$) and energy parameters for the angular potentials that are considerably higher than in the previous simulation ($E_{\text{tor}} = E_{\text{bend}} = E_{\text{tilt}} = 10\,E_{\text{LJ}}$). All other parameters have their old values. The pore drift energy is varied between two values, the results are shown in Fig. 2.38 for $E_{\text{drift}}/E_{\text{LJ}} = 30$ and in Fig. 2.39 for $E_{\text{drift}}/E_{\text{LJ}} = 56$.

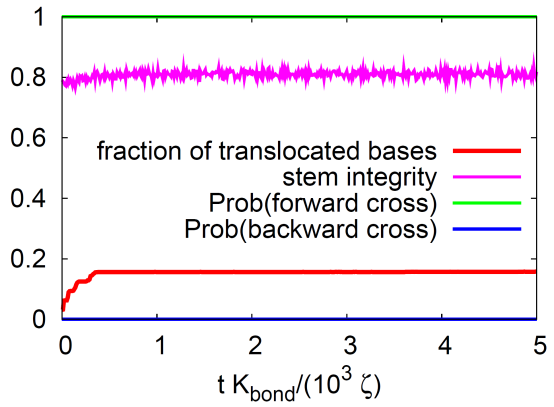


Fig. 2.38: The chain gets stuck within the pore (fraction of translocated bases does not reach unity) because the stem neither opens (constantly high stem integrity) for single-stranded translocation nor deforms sufficiently to allow double stranded-translocation.

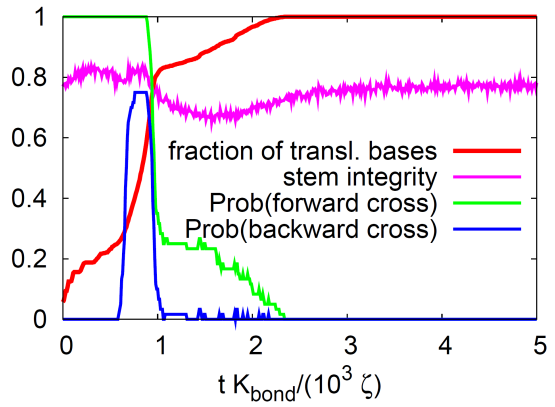


Fig. 2.39: The chain translocates totally (fraction of translocated bases reaches unity) although the stem does not open (constantly high stem integrity): The stem is translocated double-stranded (high backward cross probability during period of high translocation rate).

For $E_{\text{drift}}/E_{\text{LJ}}=30$, translocation stops as soon as the stem begins to sense the barrier potential of the pore, as can be seen from the fact that the mean fraction of translocated bases – after a short initial increase – never exceeds a relatively low value. That the pore drift force is not able to open the stem in order to allow single-stranded translocation is clear from the constantly high stem integrity, which is the mean fraction of closed base-pairs within the stem. The chain line always crosses the pore plane exactly once: The probability for crossing the plane in cis-trans direction is always one, the probability for crossing the chain in trans-cis direction is always zero (cf. probabilities for forward respectively backward crosses in Fig. 2.38). Putting this evidence together, one concludes that the chain gets stuck in the pore due to a stem that neither opens nor deforms considerably.

By contrast, for $E_{\text{drift}}/E_{\text{LJ}}=56$ all bases finally reach the trans side of the pore, as is shown by a mean fraction of translocated bases increasing with time up to unity. However, this does not happen by single-stranded translocation: First, the mean stem integrity remains high, indicating a stem relatively closed for all times. Secondly, one observes a time interval of high translocation rate during which the chain crosses the pore plane not only in cis-trans direction but in trans-cis direction as well (cf. high probability for backward crosses in Fig. 2.39). The only possible conclusion is that within this time interval one has double-stranded translocation, that is, the stem is translocated through the pore in its closed state, necessarily deformed to fit the narrow pore channel.

The explanation for this transition from the pausing chain regime to a regime of double-stranded translocation is as follows: The relatively high angular potential energy parameters shifts the coexistence line between the pausing and the passing regimes in Fig. 2.37 considerably upwards by rising the relative base-pair binding energy according to:

$$\alpha_{\text{bind}} \approx 1+E_{\text{tor}}/E_{\text{LJ}}+E_{\text{bend}}/E_{\text{LJ}}+E_{\text{tilt}}/E_{\text{LJ}} \tag{2.59}$$

Therefore both parameter points $\left(kT/E_{\text{LJ}};E_{\text{drift}}/E_{\text{LJ}}\right)$ used in the simulation lie below this coexistence line, causing the hairpin stem to remain closed in both cases and thus forbidding single-stranded translocation. In the second case (Fig. 2.39) the pore drift energy is high enough to deform the stem to an extent where double-stranded translocation is possible, in contrast to the first case (Fig. 2.38), where the stem cannot cross the pore channel. The value of the critical pore drift energy separating these two cases is determined in theory by the energy cost needed in order to reduce the stem radius such that it equals the diameter of the pore channel:

$$E_{\text{drift}}^{\text{crit}}=E_{\text{deform}}\left(d_{\text{pore}}\right) \tag{2.60}$$

Here the deformation energy cost E_{deform} is a monotonically decreasing function of the stem radius and vanishes at the equilibrium stem radius. The energy scale of this function can be assumed linear in all energy parameters that stabilize the stem radius (e.g., the tilt energy parameter and the energy parameter for the repulsion between incompatible bases).

For example, a hypothetical two-dimensional hairpin stem whose resistance against radial deformation is caused solely by the inclination potential acting on its m base-pairs would yield for very narrow pore channels:

$$E_{\text{drift}}^{\text{crit}} \approx \frac{1}{2} m E_{\text{tilt}} \frac{d_{\text{pore}}^2}{R^2} \tag{2.61}$$

To summarize, the regime diagram of Fig. 2.37 has to be updated in general. The regime of chain pausing is in fact a regime where the hairpin stem remains closed (stem closed regime). Only as long as the pore drift energy lies below the critical pore drift energy the chain really pauses in the stem closed regime (as it is the case in Fig. 2.37). If within this regime the pore drift energy can be chosen such that it is higher than the critical pore drift energy, one obtains a new sub-regime of the closed stem regime, namely the regime of double-stranded translocation.

The hierarchical structure of all regimes treated so far is depicted in Fig. 2.40. The regime most interesting for the following studies will be the regime of single-stranded translocation near the coexistence line to the regime of chain pausing.

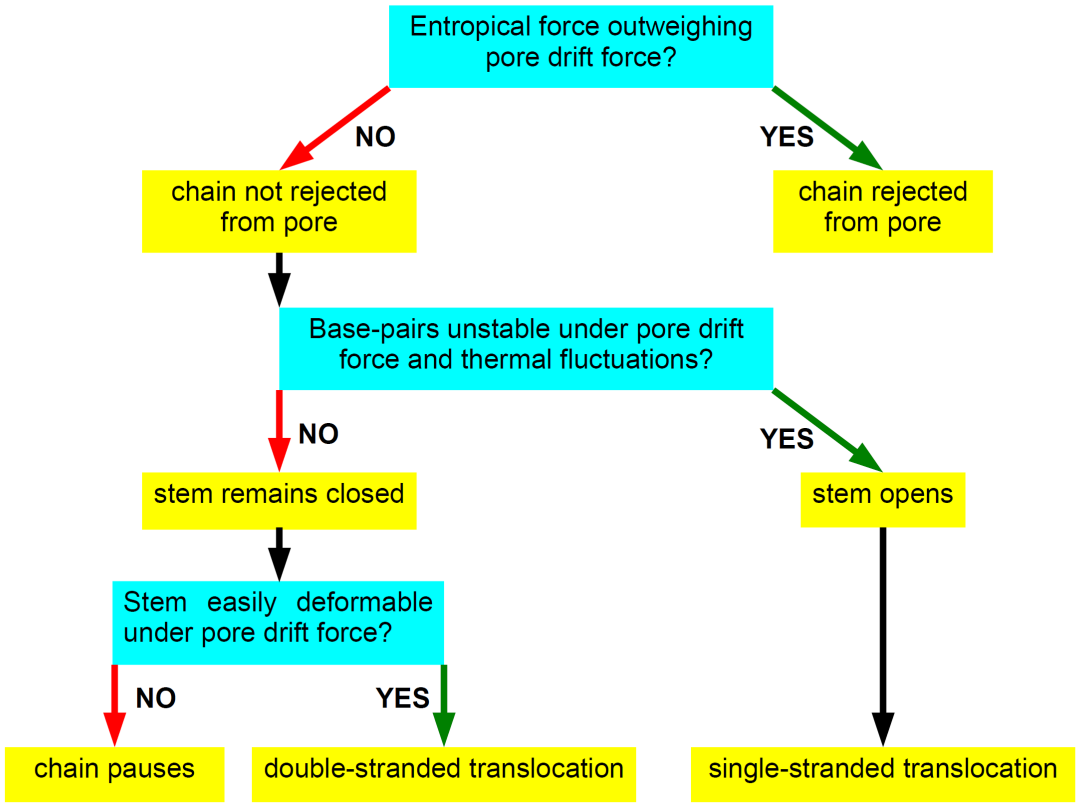


Fig. 2.40: Case differentiation resulting in the four regimes of chain behavior the model parameter space can be divided into: chain rejection, chain pausing, double and - most important - single stranded translocation of the chain.

2.5.2.2 The process of translocation in detail

After having explored the pore/chain parameter space and having mapped the four observed types of hairpin behavior to distinct energy regimes, whose origins have been able to be explained by simple theoretical arguments, the next step is to study the process of single-stranded hairpin translocation in detail and in a time-resolved way. For this purpose it is useful to return to the original values for the energy parameters of the angular potentials (cf. chapter 2.5.2.1) to avoid the existence of a regime of double-stranded-translocation and to regain the three-regime structure of the parameter space diagram shown in Fig. 2.37. In this diagram it is best to study an intersection line of constant temperature ($kT/E_{LJ}=2^{-2}$) which crosses both the coexistence line between the regime of chain rejection (-) and the regime of chain pausing (0) and the coexistence line between the regime of chain pausing and the regime of single-stranded chain passing (+), as visualized in Fig. 2.41,

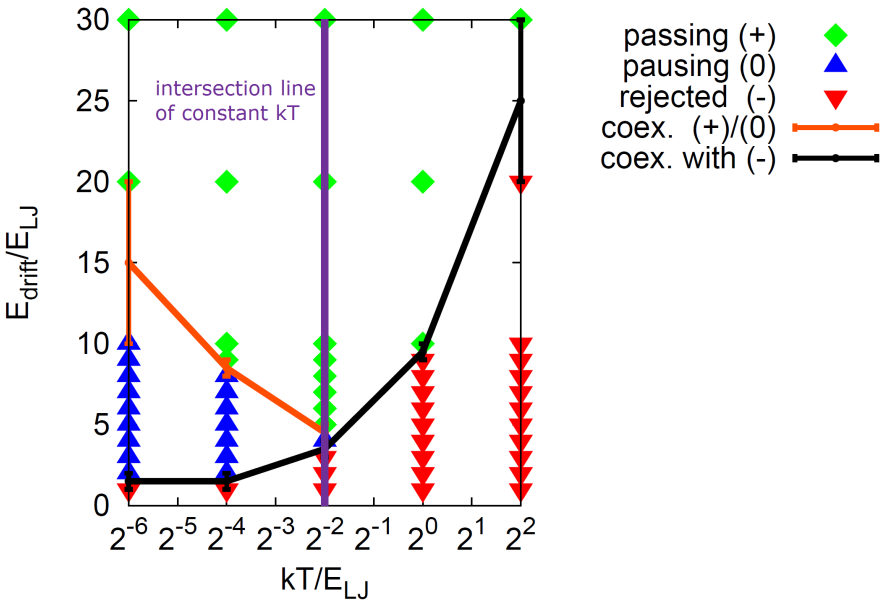


Fig. 2.41: Again the three regimes of hairpin behavior (green, blue and red) with their coexistence lines. The isothermal intersection line (magenta) used for all future studies is chosen such that it intersects both coexistence lines (orange and black).

by varying the dimensionless "pore strength",

$$\epsilon=\frac{E_{\text{drift}}}{E_{\text{LJ}}}\,. \tag{2.62}$$

These coexistence lines are crossed for

$$\epsilon=\epsilon_{-}^{0}(kT/E_{\text{LJ}}) \qquad \text{and} \qquad \epsilon=\epsilon_{0}^{+}(kT/E_{\text{LJ}}) \tag{2.63}$$

respectively, where the indices at the coexistence line functions denote the corresponding adjacent regimes.

The time regimes of the translocation process

In Fig. 2.42 and Fig. 2.43 the mean fraction of translocated bases is plotted as a function of time for a collection of pore strengths ϵ lying entirely within the regime of single-stranded translocation. Obviously, the total translocation time – the time when the plotted fraction reaches unity – decreases when the pore strength increases, as expected. But how can the different time regimes into which each curve is divided be explained?

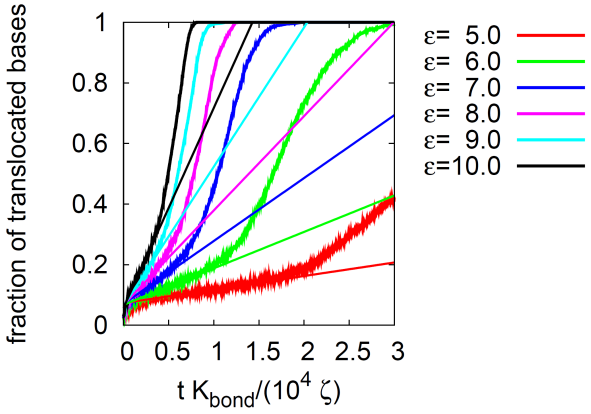


Fig. 2.42: The fraction of translocated bases increases linear in time for short times: regime of stem unzipping. The relatively low short time translocation rate (also called unzipping rate) increases with increasing pore strength.

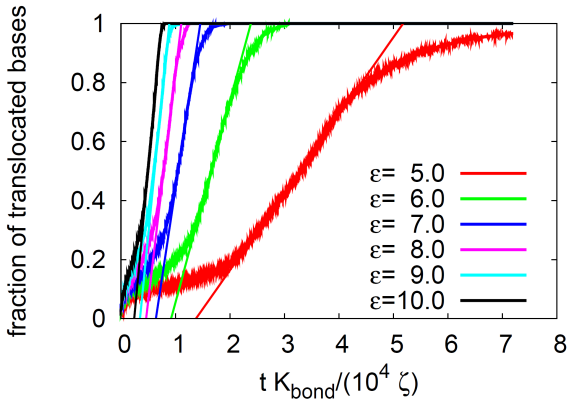


Fig. 2.43: The fraction of translocated bases increases linear in time for long times: regime of free translocation. The relatively high long time translocation rate (also called free rate) increases with increasing pore strength.

The first time regime emerges due to the translocation of the initial sequence of unbound bases located in front of the stem, which remains still closed. This **time regime of initial threading** ends as soon as the first base that belongs to a stem base-pair is reached. The fact that the bases are not bound within base-pairs makes the mean translocation rate high.

By contrast, in the second time regime, the so-called **time regime of stem unzipping**, each base is bound within a stem base-pair, which has to be forced open in order to allow the base to pass the pore channel. The rate of the combined process of pair unzipping and base translocation – the **short time translocation rate or unzipping rate** $k^{-}_{transloc}(\epsilon)$ – is constant over time and much lower than it would be if the translocated bases were unbound from the beginning.

As soon as all stem base-pairs have been unzipped, all bases to be translocated from now on are again unbound and the third (and last) time regime, the **time regime of free drift**, starts. Again the corresponding translocation rate – the **long time translocation rate or free rate** $k^{+}_{transloc}(\epsilon)$ – is constant over time, but without need for pair unzipping the translocation process is much faster.

In order to verify the physical causes for these three regimes it is useful to compare the time dependence of the mean fraction of the translocated bases with the time dependence of the base-pair integrity of the hairpin stem. This is done here for a high pore

strength, far off the coexistence line between chain pausing and chain passing (ϵ=20 , cf. Fig. 2.44). Again, one distinguishes between different time regimes (A,B,C,D,E,F):

- Regime A has to be identified with the regime of initial threading: The stem remains completely closed, the mean translocation rate is high and the maximal fraction of translocated bases is low.
- Regime B has to be identified with the regime of stem unzipping: The stem opens with initially constant rate, thus indicating a step-by-step opening of all of its base-pairs (cf. the stem integrity). Accordingly, the translocation rate is constant and relatively low (cf. the mean fraction of translocated bases). Furthermore, the fraction of bases translocated during this regime equals the fraction of the bases in the pore-next stem helix S_i (cf. the sequence representation to the right).
- Regime C and D together correspond to the regime of free drift:
  - In regime C ("free drift without stem re-zipping") the stem remains completely open whereas the translocation rate is as constant as in regime B but much higher because of the complete lack of base-pairs on the *cis*-side of the pore. Thus, the bases drift freely through the pore channel.
  - In regime D ("free drift with stem re-zipping") there are no base-pairs on the *cis*-side, just as in regime C. In this sense, there is still free drift. On the *trans*-side, however, the stem begins to close again – base-pair by base-pair. This recombination process produces an additional drift force on the bases on the *cis*-side, thus increasing effectively the translocation rate.
- With the beginning of regime E the process of base translocation is already finished. The recombination of stem base-pairs on the *trans*-side, which has already started in regime D, continues – but more slowly than there.
- With the beginning of regime F this recombination process comes to an end: The hairpin is completely folded again.

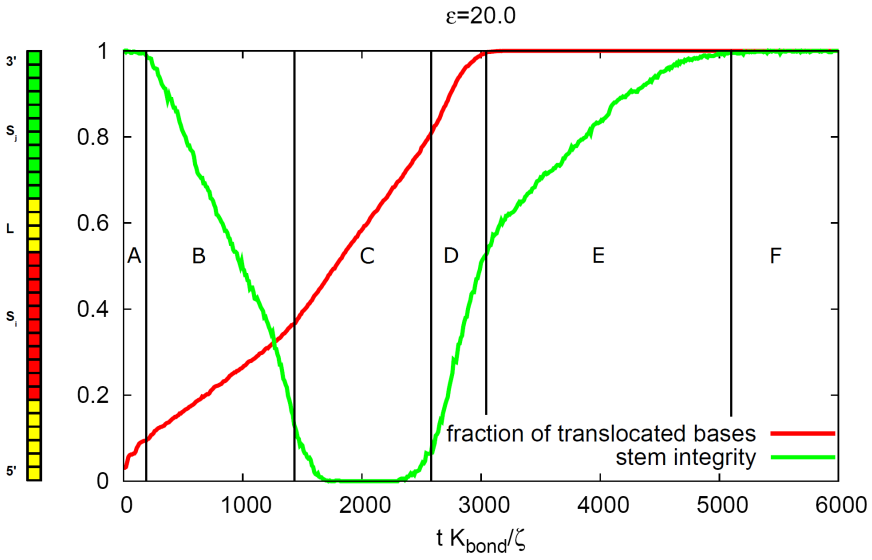


Fig. 2.44: Subdivision of the process of hairpin translocation into different time regimes by the means of a comparison between the mean fraction of translocated bases and the stem integrity.

Beginning and end of each time regime can be assigned to a distinct amount of translocated bases, as can be seen from Fig. 2.45. Accordingly, the pore position in the sequence space of the RNA chain determines the time regime: To be in the regime of initial threading, the pore has to sit within the threading sequence set in front of the hairpin stem. For the regime of *cis*-side stem unzipping, the pore position has to be within the sequence region corresponding to the first helix (S_i) of the hairpin stem. The regime of free drift starts, as soon as the pore enters the loop region of the hairpin (L). This free drift can be accompanied by *trans*-side recombination of base-pairs and stem re-zipping, as soon as the pore is within the sequence region for the second stem helix (S_j). Deviations from these predictions emerge from the fact that in the simulations the length of the pore channel is not zero and the repulsion caused by the Gaussian-shaped pore barrier destabilizes base-pairs long before they approach the pore center. This can be seen if one returns to Fig. 2.44 and compares the sequence representation to the right with the mean fraction of translocated bases.

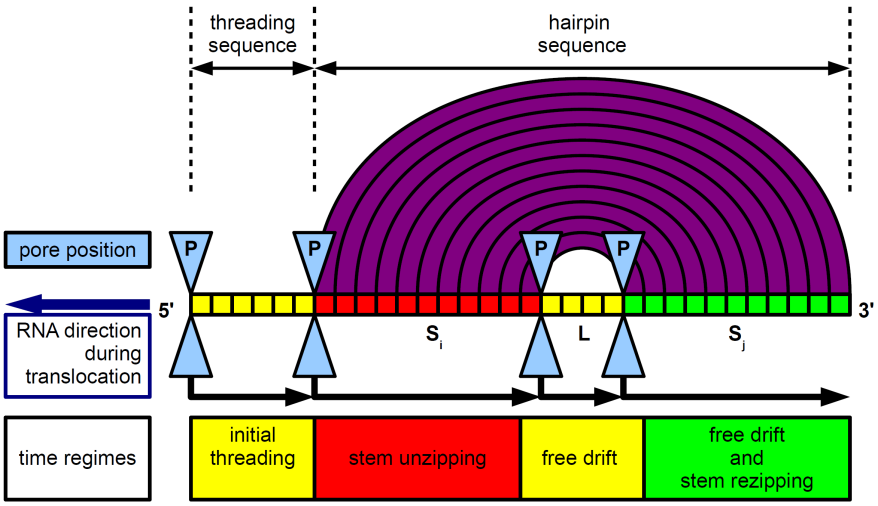


Fig. 2.45: Relation between the position of the pore P with respect to the RNA chain and the current time regime of the hairpin translocation process.

Translocation rates and their dependence on pore strength

After the discussion of the physical origin of all the time regimes within the process of translocation let us look at the regime of stem unzipping and the regime of free drift in more detail and study the dependence of the translocation rates on pore strength in both regimes. For each pore strength these two rates are determined by performing linear fits in the two time regions of constant translocation velocity, as shown in Fig. 2.42 (stem unzipping) and Fig. 2.43 (free drift). The resulting slopes are plotted as functions of the pore strength, as can be seen in Fig. 2.46 (translocation rates for low and intermediate pore strengths) and Fig. 2.47 (translocation rates for high pore strengths):

Because of the additional energetic barrier needed to cross in the time regime of stem unzipping, the unzipping rate is always lower than the free rate, except at the coexistence line between chain pausing and chain passing, where – as expected – both rates vanish:

$$k^{-}_{\text{transloc}}(\epsilon^{0}_{-})=0=k^{+}_{\text{transloc}}(\epsilon^{0}_{-}) \quad \text{with} \quad \epsilon^{0}_{-}\approx 4 \tag{2.64}$$

For pore strengths even lower (so-called low pore strengths) there is no translocation at all and both rates remain zero.

On the other hand, both rates show a linear dependence on pore strength for high pore strengths, the corresponding slope being greater for the free rate than for the unzipping rate. This linearity of both rates may serve as a criterion that the pore strength can be considered high.

When the pore strength decreases to intermediate values, this linearity continues for the unzipping rate even until the pausing/passing coexistence point is reached, whereas it is lost by the free rate. In this example the region of intermediate pore strength goes from 4 to approximately 10.

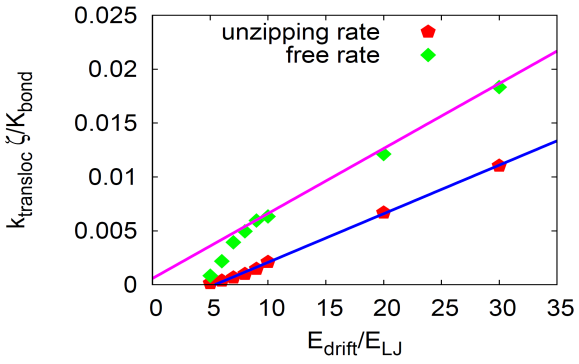


Fig. 2.46: Rates for translocation by stem unzipping and for free translocation as a function of pore strength at constant temperature. Here one focuses on the pausing/passing transition between low and intermediate pore strengths, where both rates vanish.

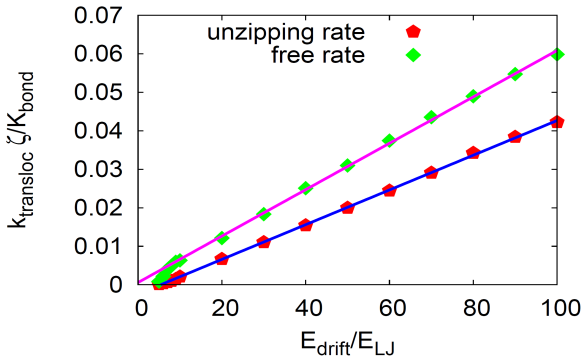


Fig. 2.47: Rates for translocation by stem unzipping and for free translocation as a function of pore strength at constant temperature. Here one focuses on the linear increase of both rates for high pore strengths.

Behavior for low, intermediate and high pore strengths

In Fig. 2.48 and Fig. 2.49 the mean fraction of translocated bases respectively the stem integrity are plotted as functions of time for low and intermediate pore strengths. One observes that two coexistence lines are crossed:

- Pore strength 3 lies in the regime of chain rejection: The fraction of translocated bases vanishes for long times, whereas the stem remains completely closed.
- Pore strength 4 lies in the regime of chain pausing: Again the stem remains closed, but the fraction of translocated bases becomes stationary at a low positive value, thus indicating that only the bases of the threading sequence are translocated.

- Pore strength 4.8 and 5.0 already lie in the regime of chain passing: More and more bases are translocated as time passes, and the stem opens completely. This has been confirmed by counting how often the chain crosses the pore plane at different points of time (data not shown). Paradoxically, the stem integrity does not reach zero during the process of stem opening. The reason is that the pore strength is still not high enough to locate the events of stem opening and stem closing within narrow time intervals. Therefore, averaging yields a stem integrity with a non-zero minimal value.

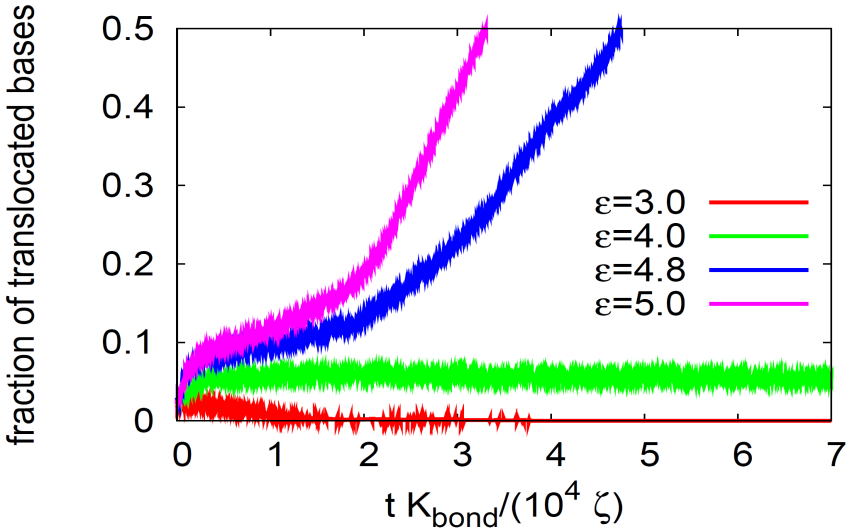


Fig. 2.48: Fraction of translocated bases as a function of time for pore strengths in the regimes of chain rejection (red), chain pausing (green) and chain passing (blue, magenta).

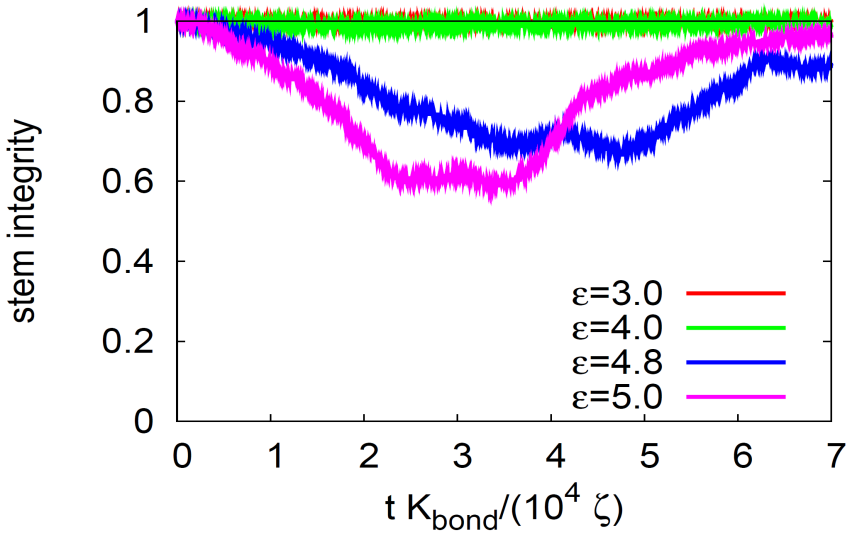


Fig. 2.49: Stem integrity as a function of time for pore strengths in the regimes of chain rejection (red), chain pausing (green) and chain passing (blue, magenta).

This non-zero minimal value of the stem integrity can be found for all intermediate pore strengths, where base translocation is still slow. The effect decreases with increasing pore strength until the minimal stem integrity becomes zero, as is shown in the stem integrity plots of Fig. 2.50. The explanation for this behavior is that the greater the pore strength is, the narrower are the time distributions for the events of stem opening and stem closing and the more similar to each other are the single integrity trajectories (each of which reaches zero at some time) one has to average to obtain the stem integrity.

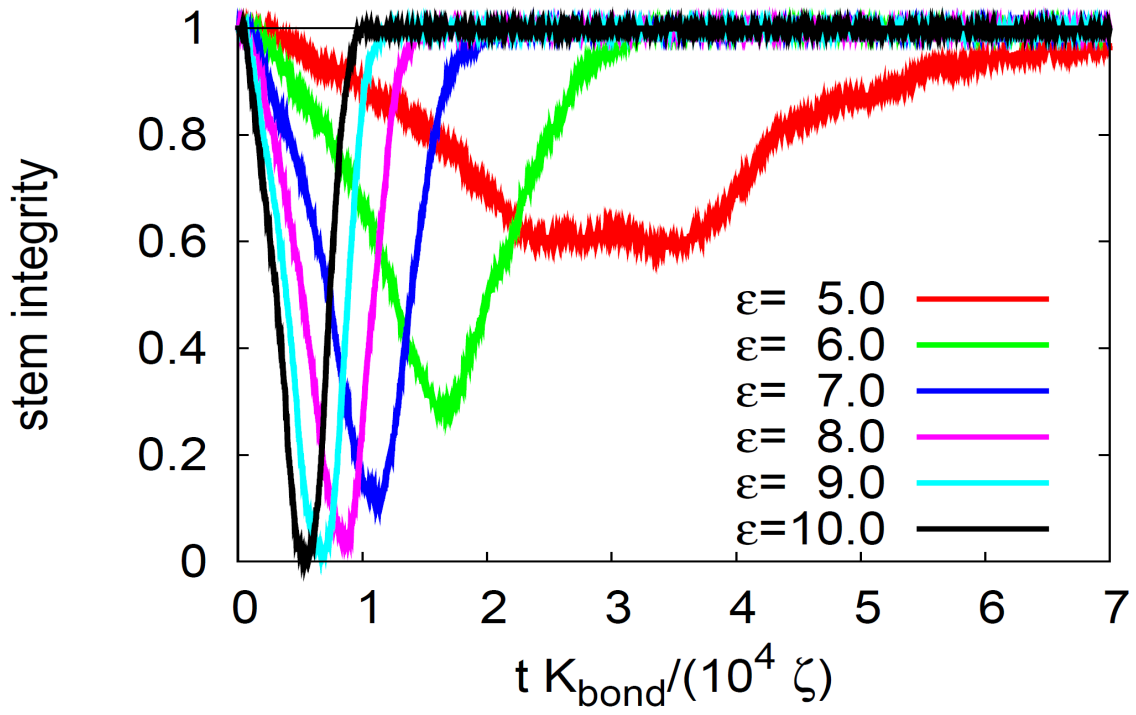


Fig. 2.50: Stem integrity as a function of time for slow translocation by intermediate pore strengths. The stem opening occurring in each single run is always complete, the non-zero minimal values result from averaging. This effect decreases with increasing pore strength, where the opening and closing events are located more sharply in time.

If on the other hand the pore strength is high, making translocation fast, as it is the case in the stem integrity plots shown in Fig. 2.51, the stem integrity always has a zero minimal value of certain time duration. The time distributions for the events of stem opening and stem closing are indeed very narrow in this pore strength regime. This can be demonstrated by collapsing those parts of the integrity curves in Fig. 2.51 onto each other that correspond to the time regime of stem unzipping, simply by rescaling the time axis by the value the unzipping rate has for the respective pore strength. This collapse is shown in Fig. 2.52. For intermediate pore strengths such a collapse is not possible.

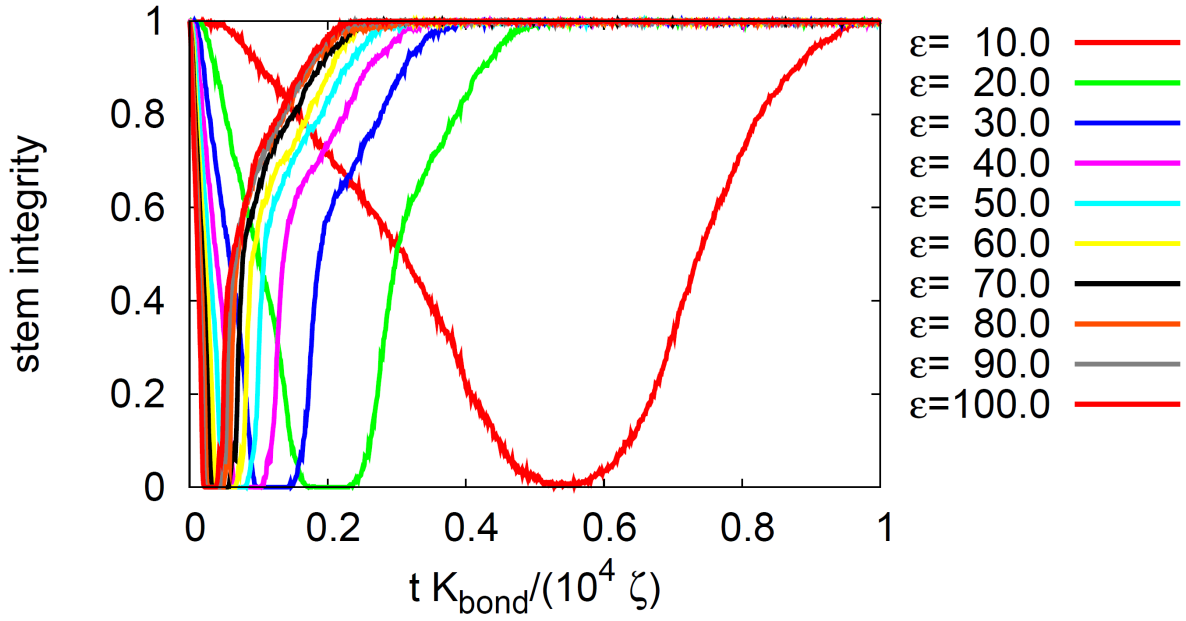


Fig. 2.51: Stem integrity as a function of (non-rescaled) time for fast translocation by high pore strengths. Due to the strong pore forces the time distributions for the events of stem opening and stem closing are narrow. Therefore the stem integrities resulting from averaging over the single runs have vanishing minimal values.

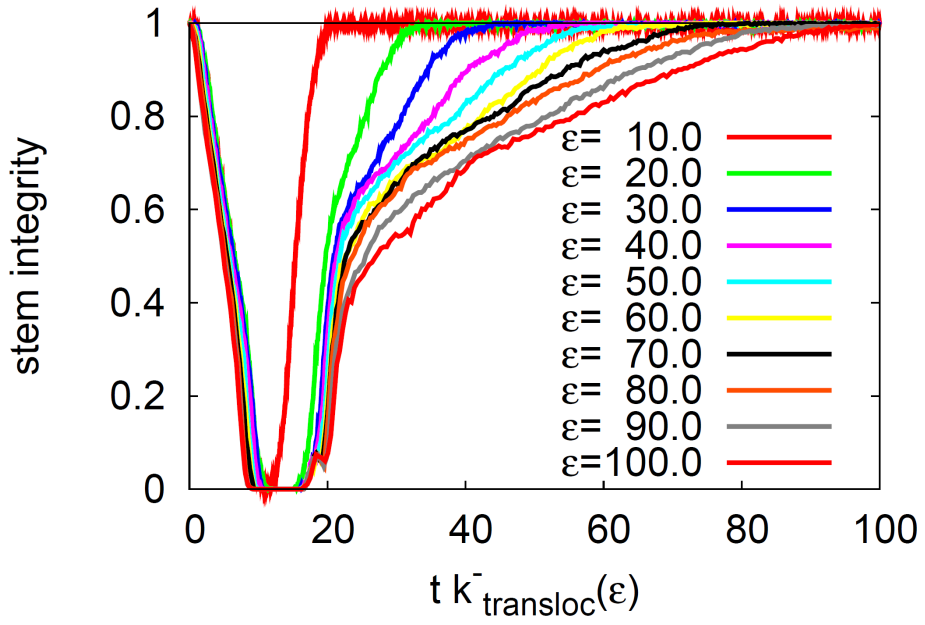


Fig. 2.52: Stem integrity as a function of time after having rescaled time by the rate of stem unzipping, which in turn is a function of pore strength. Due to the strong pore forces the time distributions for the events of stem opening and stem closing are narrow. Therefore, the parts of the curves that correspond to the time regime of unzipping can be collapsed onto each other.

2.5.3 Hairpin translocation versus pseudoknot translocation

The goal of this chapter is to identify, analyze and quantify the similarities and differences between the translocation processes of hairpins and pseudoknots, and especially to study under which circumstances pseudoknot translocation is appreciably slower than hairpin translocation, thus giving rise to such phenomena as ribosomal pausing. To this end, the translocation process of pseudoknots has to be characterized. This task is simple after having done such characterization in the previous chapter for the hairpin case. The hairpin and pseudoknot sequences used in performing the necessary comparisons are shown in the overview of Fig. 2.53. Especially the corresponding patterns of native base-pairs will prove important.

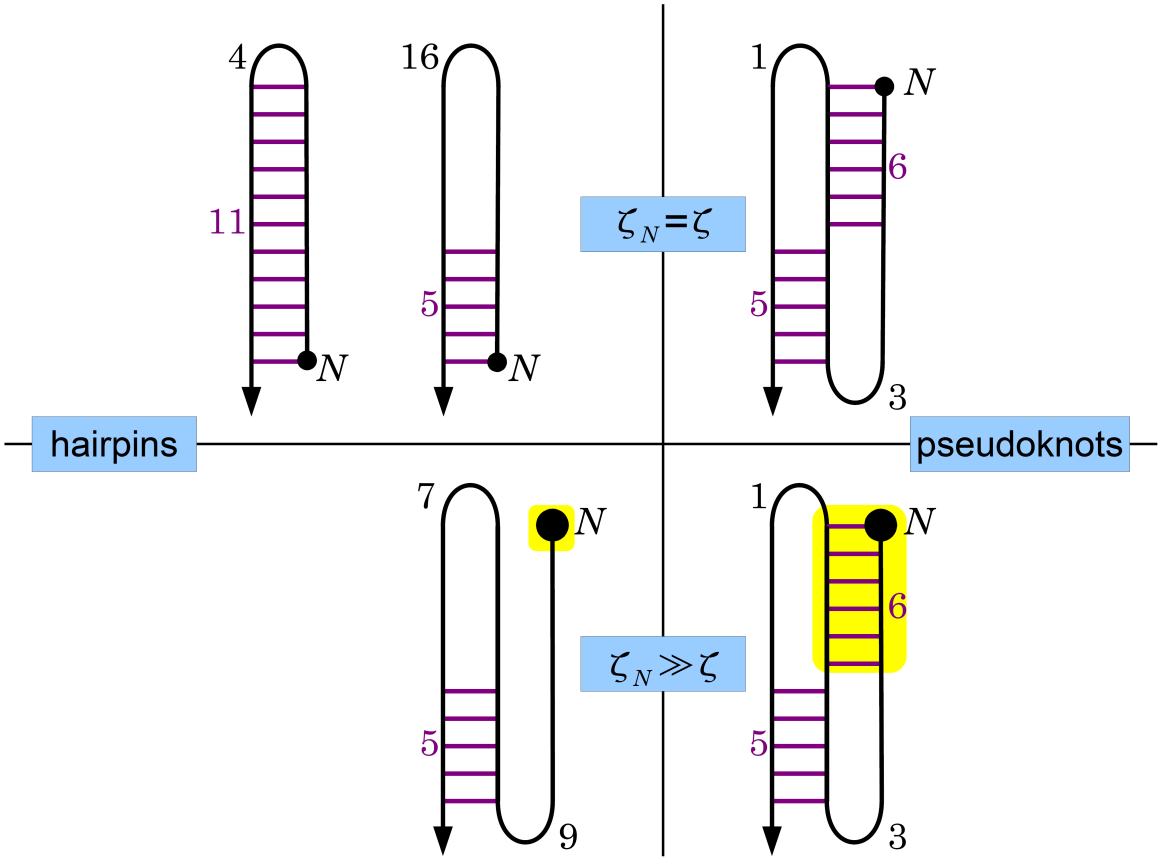


Fig. 2.53: Above: Hairpin and pseudoknot sequences without upstream chain residues (last base N with normal friction coefficient). Below: Hairpin and pseudoknot sequences with large upstream chain residues (last base N with a friction coefficient much higher than normal). In each of the sequences below the positions in three-dimensional space of the bases labeled yellow are regarded as being correlated to the position of the last base due to base-pairing. The effect of this will be discussed later.

2.5.3.1 Analogous time regimes for pseudoknot structures

If one compares especially a hairpin with a stem consisting of 5 base-pairs (the sequence can be found in the top part of Fig. 2.53) to a pseudoknot with two stems (stem A with 5

base-pairs, stem B with 6 base-pairs, again depicted in the top part of Fig. 2.53), one recognizes from Fig. 2.54, where the mean fraction of translocated bases is plotted against time, that in both cases the translocation process can be subdivided into a number of time regimes, there denoted by A, B and C (hairpin) and by A, B_1, C_1, B_2, and C_2 (pseudoknot).

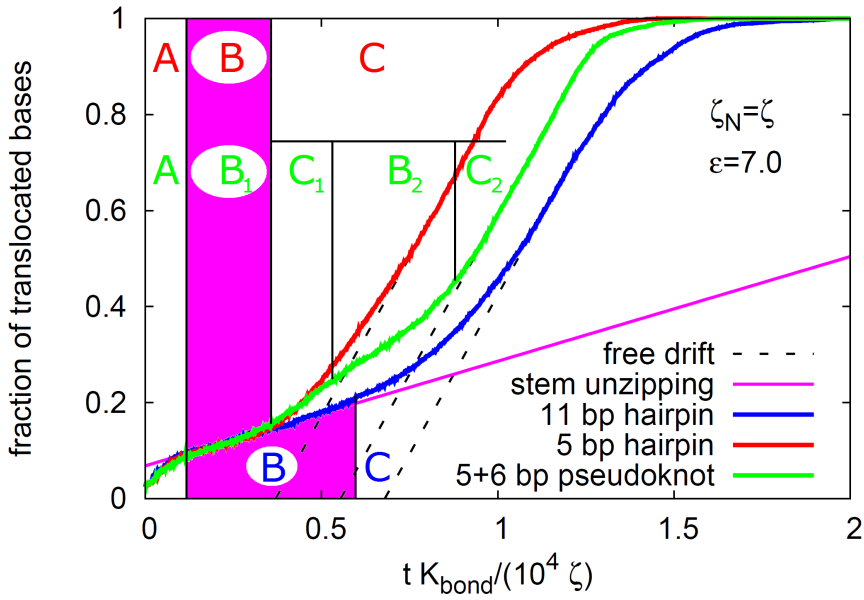


Fig. 2.54: A long hairpin (blue), a short hairpin (red) and a pseudoknot (green) are compared to each other by looking at the mean fraction of translocated bases as a function of time at constant pore strength and temperature. Stem A of the pseudoknot and the stem of the short hairpin have equal base-pair numbers, as it is the case for the whole pseudoknot and the long hairpin. The stem unzipping rates (look at stem A for the pseudoknot) are equal for all three sequences (cf. the slope of the magenta line). This is also true for the free rates (the dotted lines have equal slopes) after the stems have opened. A description of the time regimes (A,B,C, ...) can be found in the text.

Since this subdivision has already been studied in the hairpin case (cf. chapter 2.5.2.2), it is sufficient to explain the differences emerging in the pseudoknot case.

- Both the hairpin and the pseudoknot sequence feature regimes of initial threading (denoted by A).
- Whereas the hairpin has only one stem and accordingly only one regime of stem unzipping (denoted by B), the pseudoknot exhibits two regimes of stem unzipping due to its two stems. These regimes are denoted by B_1 for the first stem (stem A) and by B_2 for the second stem (stem B) and can be recognized from the comparatively low translocation rates (cf. the small slopes within the unzipping regimes in Fig. 2.54).
- Unlike the hairpin, the pseudoknot features an additional short time interval of free drift just after the first stem is completely open. This interval is denoted by C_1 and separates the regimes B_1 and B_2 of pseudoknot stem unzipping.
- Finally, both hairpin and pseudoknot show regimes of fast translocation for long times, denoted by C and C_2 respectively. There the process of translocation is dominated by the free drift of bases not bound any longer within *cis*-side base-

pairs and accompanied later even by re-zipping of the stem(s) on the *trans*-side of the pore.

Fig. 2.55 shows a visualization of the time regimes newly discovered for the pseudoknot in terms of the translocation coordinate, which is the number of bases that are already on the *trans*-side of the pore. Here this number is presented as the pore position on the chain contour. The corresponding diagram for the hairpin is already known from Fig. 2.45. The main difference to that hairpin diagram is the existence of two regimes of stem unzipping (B_1 and B_2) in the pseudoknot case.

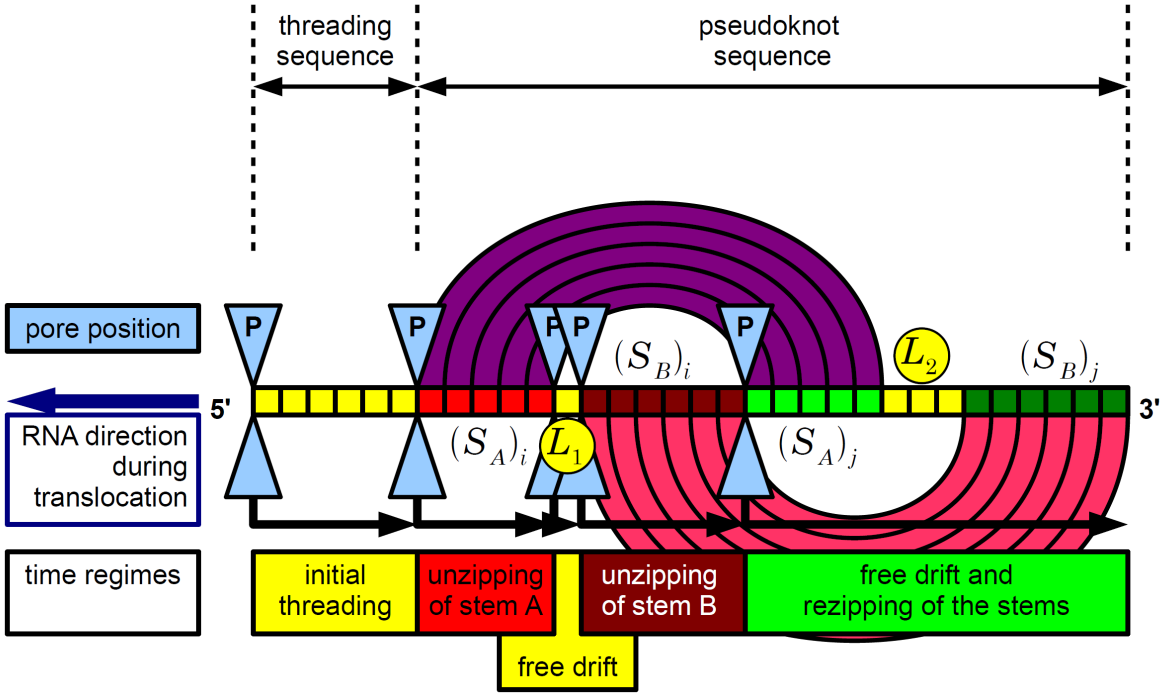


Fig. 2.55: In which time regime of pseudoknot translocation the system is depends on the pore position P on the chain, whose native base-pairs form two stems S_A and S_B. Accordingly, there are two regimes of stem unzipping whereas hairpin translocation has only one.

2.5.3.2 Geometry versus total binding energy

Translocation is much slower for pseudoknots than for hairpins. Generally, it is assumed that the reason for this fact is that pseudoknots have special geometrical properties which prevent their pore-next stems from opening fast, whereas hairpins lack these properties. This hypothesis contradicts the naive assumption that total translocation time – if dominated by the total unzipping time – is a mere function of the total binding energy and not influenced strongly by geometry:

For high enough pore strengths ϵ the total unzipping time Δt_{unzip} of all stems of a folded RNA chain is proportional to the total base-pair binding energy of those stems $E_{\text{LJ}}^{\text{tot}}$ divided by the energy gained when translocating a single base through the pore E_{drift} :

$$\Delta t_{\text{unzip}} = \frac{\Delta N_{\text{unzip}}}{k_{\text{unzip}}} \propto \frac{\Delta N_{\text{unzip}}}{\epsilon - \epsilon_c} \rightarrow \frac{\Delta N_{\text{unzip}}}{\epsilon} \quad \text{for} \quad \epsilon \gg \epsilon_c \tag{2.65}$$

Therefore:

$$\Delta t_{\text{unzip}} \propto \frac{E_{\text{LJ}} \Delta N_{\text{unzip}}}{E_{\text{drift}}} \equiv \frac{E_{\text{LJ}}^{\text{tot}}}{E_{\text{drift}}} \quad \text{for} \quad \epsilon \gg \epsilon_c \tag{2.66}$$

Especially, the unzipping time does neither depend on the concrete order of the base-pairs that are unzipped nor on their initial geometrical loci. This would predict that pseudoknots and hairpins have equal total unzipping times, as long as their total base-pair binding energies are equal. However, it has been established that the frameshift efficiency, an indirect indicator for the extent of ribosomal pausing, is not correlated with the thermodynamic stability of the folding motif [120][121].

2.5.3.3 Confirmation of the influence of total binding energy

For example, the sum of the unzipping times of two pseudoknot stems with 5 and 6 base-pairs should equal the unzipping time of a hairpin stem with 11 base-pairs. However, because the angular potentials contribute additively to the binding energy per base-pair and these contributions are weaker for the first and last base-pairs of a stem than for its interior base-pairs, it is clear that this equality does not hold exactly in this case: As the hairpin and pseudoknot unzipping rates are equal only in the interior of the stem(s), the equal numbers of base-pairs that have to be unzipped ($N_{\text{unzip}}=11$) makes the hairpin unzipping time slightly larger than the pseudoknot unzipping time. In Fig. 2.54 the time-dependence of the respective mean fractions of translocated bases is shown.

2.5.3.4 How to detect the influence of geometry

The goal is to observe a pseudoknot structure which is translocated much slower than a hairpin structure not for reasons of binding energy but for geometrical reasons. That is, differences in the total unzipping time that emerge from different total binding energies brought about in turn by different numbers of base-pairs can be regarded as irrelevant. Therefore, it is legitimate to compare a shorter 5-bp hairpin to a 11-bp pseudoknot in-

stead of using an 11-bp hairpin: One cannot decide to what extent it is the total binding energy or geometrical reasons that influence the translocation process if only global properties of the translocation process, such as the total translocation time or the total unzipping time, are examined. That is why one has to focus on the process of translocation and unzipping of the pore-next stems (the hairpin stem and stem A of the pseudoknot) instead, which now have equal numbers of base-pairs and thus are comparable in terms of their intrinsic energetic and geometric properties. The question is whether stem B of the pseudoknot is able to influence the unzipping and translocation behavior of stem A to an extent that it differs appreciably from the behavior of the single hairpin stem.

2.5.3.5 Detection impossible without upstream sequence extensions

Since the hairpin stem and stem A of the pseudoknot have the same number of base-pairs, a difference in their unzipping times can only be caused by different stem unzipping rates. However, in Fig. 2.54, which shows the mean fraction of translocated bases over time for both the 5-bp hairpin and the 11-bp pseudoknot, unzipping times as well as stem unzipping rates are equal for the pore-next stems of both sequences. Furthermore, the two respective stem integrities (Fig. 2.56) show identical time behavior, especially they become minimal at the same point of time. Altogether, this indicates that the unzipping process of stem A of the pseudoknot is unaffected by stem B, since it behaves like an isolated stem, represented here by the 5-bp hairpin.

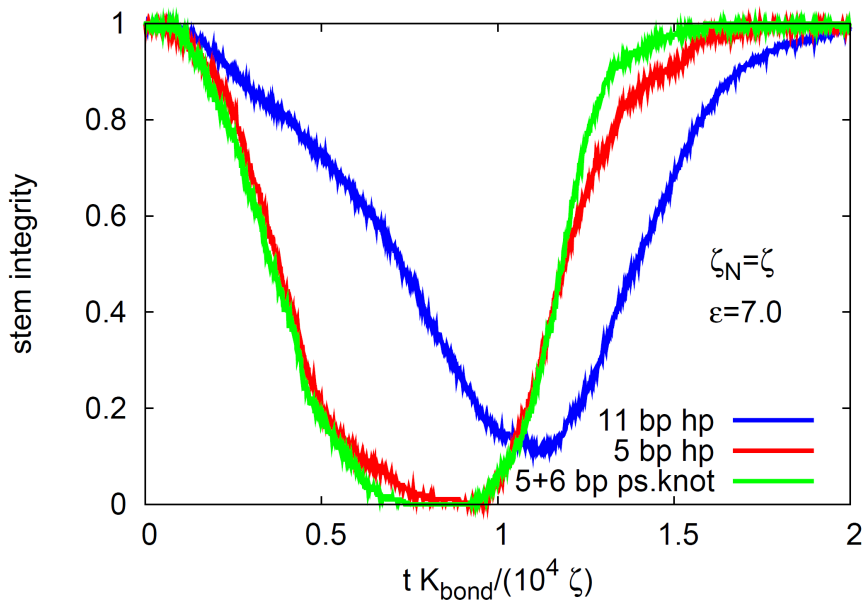


Fig. 2.56: Stem integrity as a function of time. Pore strength and temperature are as in Fig. 2.54. Again the long hairpin (blue), the short hairpin (red) and the pseudoknot are compared. The integrity of stem A of the pseudoknot and the stem integrity of the short hairpin behave similar. This indicates that stem B of the pseudoknot has no influence on the opening process of stem A. By contrast, the stem of the long hairpin opens later.

2.5.3.6 Detection successful with upstream sequence extensions

Motivation and prediction

Up to now the stem unzipping rate has been regarded as a mere function of base-pair energy, pore drift energy and temperature. One will see that under certain conditions it also depends on the geometrical properties of the RNA motif to be translocated, as does the corresponding stem integrity. Thus both stem unzipping rate and stem integrity may serve as indicators for the strong influence of geometry on the translocation process. Geometrical properties like force directions will prove important, as in recent comparative H-pseudoknot (the -1 frameshifting infectious bronchitis virus pseudoknot with 11 base-pairs in stem A and 6 base-pairs in stem B) and hairpin (a hairpin consisting of the same 17 base-pairs as both pseudoknot stems together) experiments by Green and co-workers [122]. They used optical tweezers to unfold and refold these motifs mechanically at constant force as well as with force ramps and measured the end-to-end molecular extension in order to compare the resulting unfolding and refolding characteristics. The hairpin was unzipped by a localized tensile force perpendicular to its axis, the pseudoknot, however, by an axially parallel shearing force affecting its whole structure. This situation is very similar to the pore translocation simulations studied in this thesis if one compares stem A of the pseudoknot with the hairpin stem.

In Green's experiments the (mechanically more stable) pseudoknot required a higher unfolding force, showed slower unfolding kinetics and an unfolding rate less dependent on the unfolding force than the hairpin, in spite of similar thermodynamic stabilities of the two motifs. They conclude that pseudoknot unfolding is a kinetically controlled process and suggest that in nature pseudoknots play the role of a kinetic barrier against unfolding by a ribosome.

As far as the pore-next stems are concerned, the translocation simulations in this thesis will yield similar results: If compared to the hairpin stem, the unzipping rate of stem A of the pseudoknot will be appreciably lower and furthermore depend more weakly on pore strength (in the regime of intermediate pore strength), although both stems have equal number of base-pairs and thus similar thermodynamic stabilities. These observations can be explained by the geometric influence stem B, which remains intact during the unzipping of stem A of the pseudoknot, has on the directions of the unzipping forces acting on stem A and thus on its unfolding kinetics.

Modeling the upstream sequence extensions

In order to make the geometrical influence detectable, which slows down the translocation of pseudoknots compared to the translocation of hairpins, it makes sense to extend both the pseudoknot and the hairpin sequences appreciably by appending a number of bases to their upstream ends. To keep things simple, these appended bases are represented in simulation by decreasing the motility of the last base of the chain. The resulting pseudoknot and hairpin sequences are depicted in Fig. 2.53 (bottom right and bottom left): The motility of the last base is decreased by a factor of 100.

Total translocation time: pseudoknot slower than hairpin

These sequences differ strongly in global properties of the translocation process. For example, the total translocation time of the pseudoknot is much higher than the total translocation time of the hairpin and this time difference increases with decreasing pore strength. This can be seen from Fig. 2.57 (hairpin) and Fig. 2.58 (pseudoknot), which show the probability that the chain has still not left the pore ("survival probability") over time for different pore strengths. There, the total translocation time is the time when the chain leaves the pore channel, i.e. the time when the survival probability has decreased to 50 percent. For very low pore strength (ϵ=5.0) the direct comparison of the survival probabilities yields an considerable difference τ_{Prob} in total translocation times, as shown in Fig. 2.59.

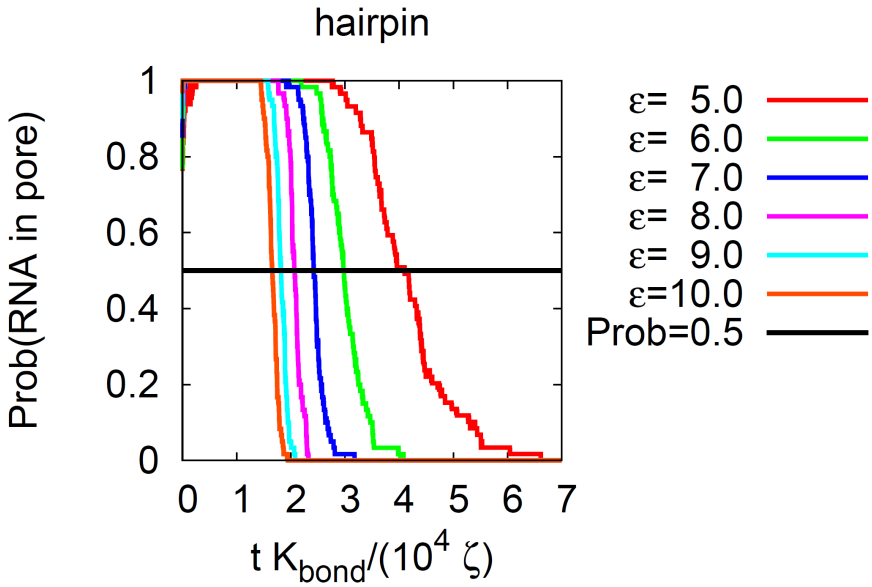


Fig. 2.57: The probability that the process of hairpin translocation is still in progress decreases from one to zero the faster the higher the pore strength is. Therefore, the total translocation time (defined here as the time when this probability sinks below 0.5) decreases with increasing pore strength.

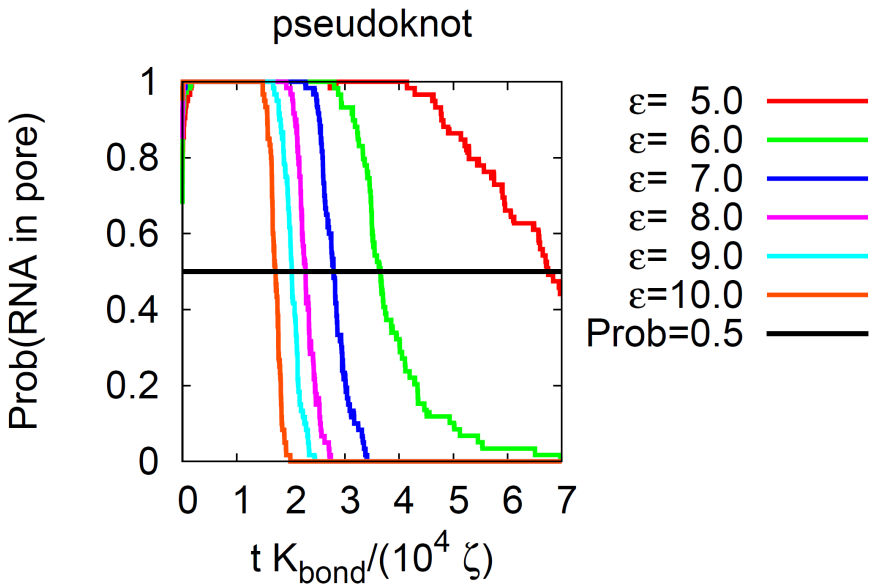


Fig. 2.58: The probability that the process of pseudoknot translocation is still in progress shows a time behavior similar to the hairpin case (Fig. 2.57): Again the total translocation time decreases with increasing pore strength.

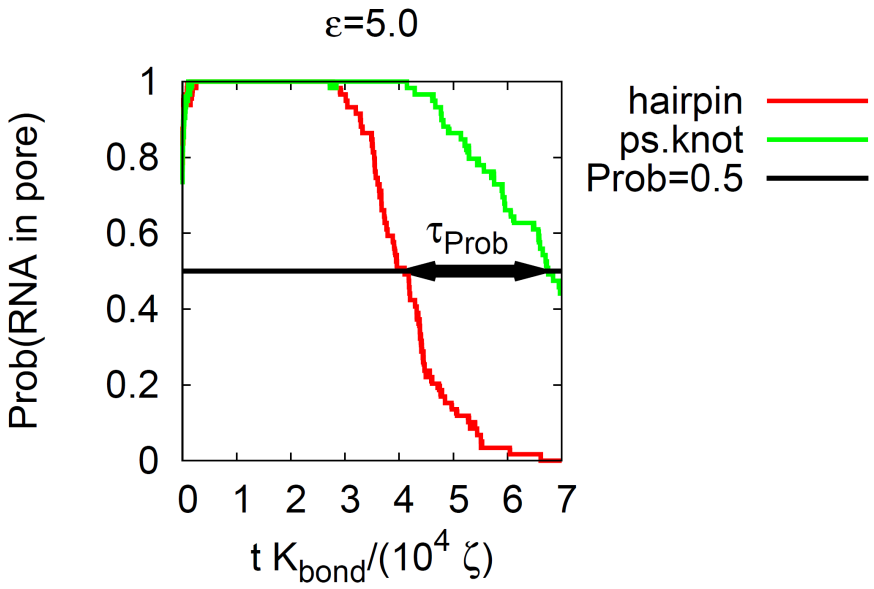


Fig. 2.59: Comparison of the probabilities that the translocation is still in progress in the hairpin and in the pseudoknot case for a low pore strength. The respective total translocation times show an appreciable difference τ_{Prob} : The pseudoknot is slower than the hairpin.

But: Insignificant for the detection of geometrical influence

Although the pseudoknot is obviously slower than the hairpin, one has no evidence that this time difference emerged from geometrical differences: Since the pseudoknot's total binding energy outweighs the total binding energy of the hairpin, the time difference could be explicable by mere binding energy arguments as well. To check doubtlessly whether stem B of the pseudoknot exerts geometric influence on stem A one has to compare directly the translocation behavior of stem A with that of the isolated hairpin stem, both again identical in terms of binding energy and local geometry, rather than to examine insignificant global translocation properties.

Primary stem unzipping rate much lower for the pseudoknot

And indeed, the corresponding stem unzipping rates differ, a feature not observable for the original sequences. If one plots the mean fraction of translocated bases against time for different pore strengths in the case of the hairpin sequence (cf. Fig. 2.60) as well as in the case of the pseudoknot sequence (cf. Fig. 2.61), the most striking observation is that in spite of equal energetic and (local) geometric conditions the stem unzipping rate (represented by the slope of the linear fit shown in the respective figures) of stem A of the pseudoknot is much lower than that of the hairpin stem. This rate difference becomes more and more manifest as pore strength decreases. This can also be seen from the slopes of the fit lines in Fig. 2.62, where a direct comparison between hairpin and pseudoknot is shown for low pore strength (ϵ=5.0) : The unzipping rate of stem A of the pseudoknot is almost four times lower than the unzipping rate of the hairpin stem.

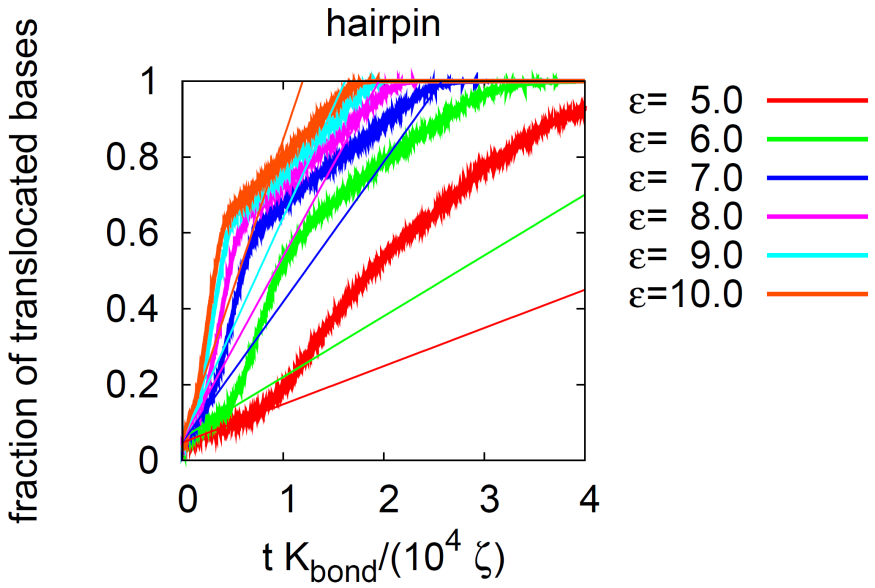


Fig. 2.60: Mean fraction of translocated hairpin bases plotted against time for different pore strengths. The stem unzipping rate, which is given by the slope of a linear fit within the time regime of stem unzipping, increases with increasing pore strength.

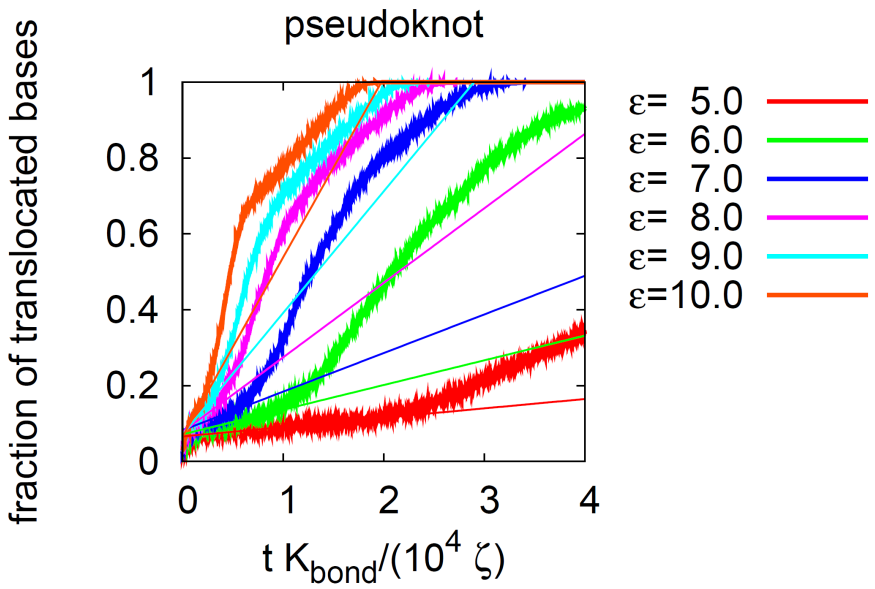


Fig. 2.61: Mean fraction of translocated pseudoknot bases plotted against time for different pore strengths. As in the hairpin case (Fig. 2.60), the stem unzipping rate obtained from the fit line slopes in the time regime of unzipping stem A increases with increasing pore strength.

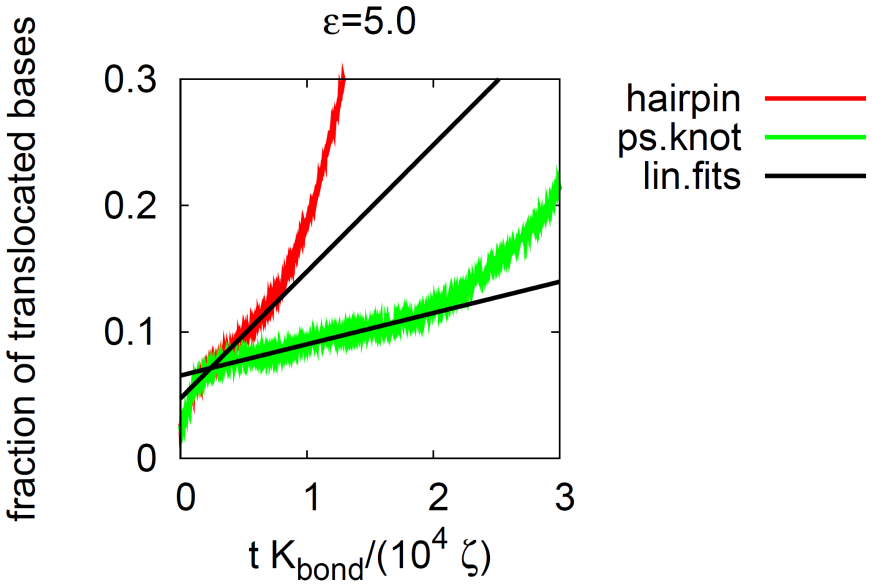


Fig. 2.62: The mean fraction of translocated bases is plotted against time for a low pore strength in order to compare the hairpin and the pseudoknot case. In the time regimes of unzipping the hairpin stem resp. stem A of the pseudoknot the translocation rate is constant. The slopes of linear fits in these regions yield the corresponding stem unzipping rates. The pseudoknot' s stem A unzipping rate is much lower than the hairpin's stem unzipping rate, although both have equal total binding energy.

Primary stem integrity decays much slower for the pseudoknot

Likewise, the time behavior of the stem integrity now shows drastic differences between the pseudoknot and the hairpin sequence whereas both sequences behaved identically in their original version. In Fig. 2.63 (hairpin) and Fig. 2.64 (pseudoknot) the integrity of the pore-next stem is plotted against time for several pore strengths. As expected, for the hairpin as well as for the pseudoknot, increasing the pore strength means an increasing rate of stem integrity reduction (cf. the slope of the falling part of each stem integrity curve) and accordingly a shorter average period of time until the stem is totally open (cf. the time positions of the stem integrity minima) and furthermore an increasing definiteness of the stem opening events in terms of time (cf. the increasing deepness of the stem integrity minima). Comparing now hairpin and pseudoknot for equal pore strength one finds: The pseudoknot has a lower rate of stem integrity reduction, its stem opening process is complete earlier and its stem opening events are less definite in terms of time. These differences are most pronounced at low pore strengths. This is especially demonstrated by Fig. 2.65 for a pore strength of ϵ=5.0 , where the time difference τ_{int} between the stem integrity minima of the hairpin and the pseudoknot (stem A) is shown to be comparatively large and even of the same order of magnitude as the difference τ_{Prob} in total translocation time.

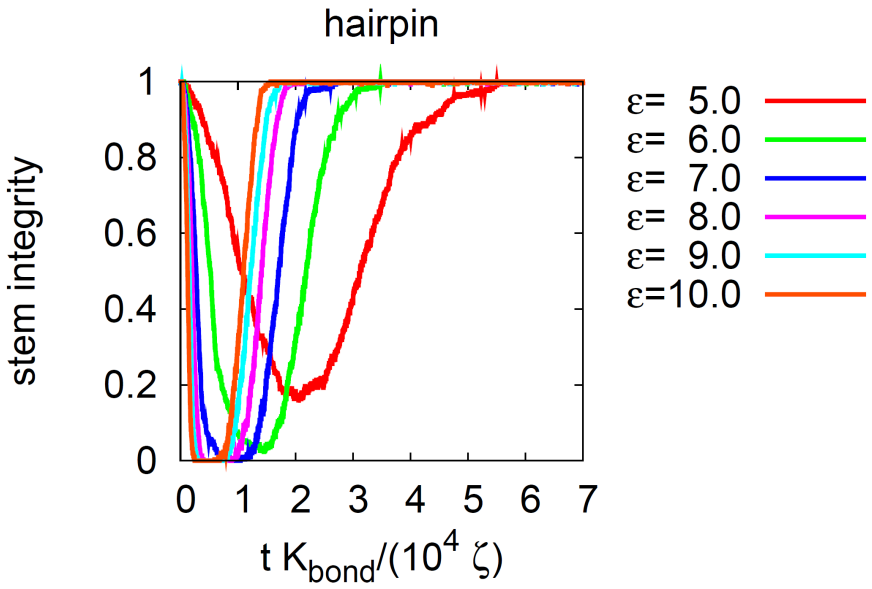


Fig. 2.63: Mean fraction of fully formed base-pairs in the hairpin stem (hairpin stem integrity) plotted against time for different pore strengths. The higher the pore strength is the faster the stem opens (slope in the declining region becoming steeper), the earlier the minimum region is reached where the stem is completely open (shift of the minimum position to short times) and the more temporally well-defined the event of complete stem opening is (deeper integrity minimum).

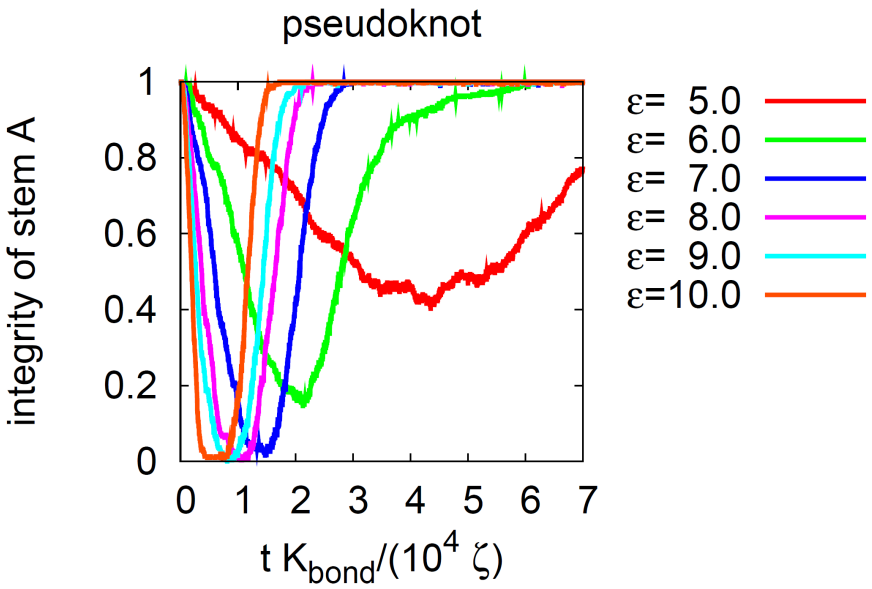


Fig. 2.64: Integrity of stem A of the pseudoknot plotted against time for different pore strengths. As in the hairpin case, the stem opens faster with increasing pore strength. Accordingly, the time interval when the stem is completely open is defined more sharply for higher pore strengths, which is indicated by a minimum value of the stem integrity approaching zero.

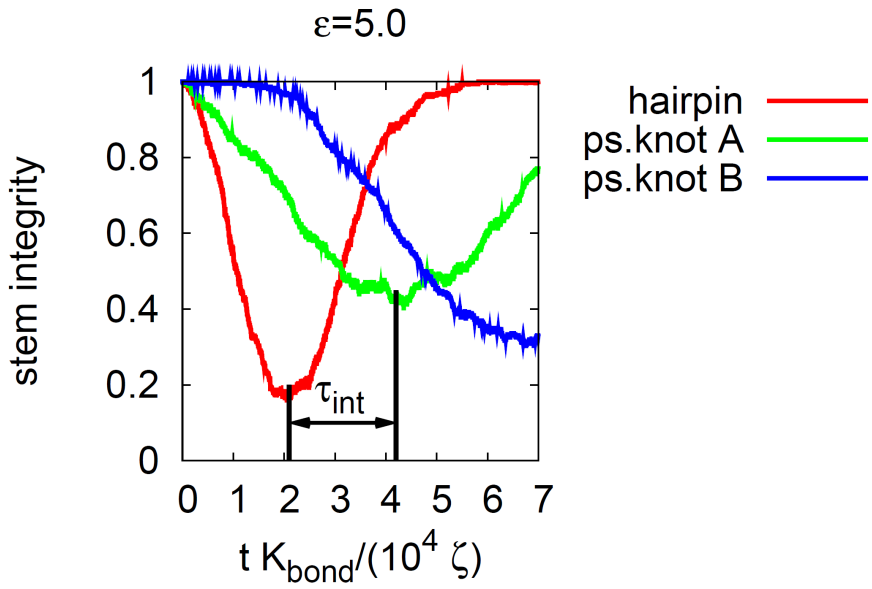


Fig. 2.65: Comparison between the integrity of the hairpin stem (red) and the integrity of stem A of the pseudoknot (green) for a low value of pore strength. The minimum positions differ strongly (time difference τ_{int}): The point of time of complete stem unzipping is much earlier in the hairpin case than in the pseudoknot case. The unzipping process of stem B of the pseudoknot (cf. blue stem integrity curve) does not start until stem A (green) is almost completely open.

Summarizing the observations

To combine all observations, as long as pore strength is low, the translocation of the pore-facing helix of stem A of the pseudoknot proceeds much slower than the translocation of the corresponding helix of the single hairpin stem. This fact is reflected in base translocation rates differing strongly during the respective time regimes of stem unzipping as well as in the corresponding stem unzipping behavior itself and may serve as evidence for the non-local geometrical influence stem B of the pseudoknot exerts on stem A, an influence totally lacking in the hairpin case.

Interpretation of the observations concerning the pseudoknot

Obviously the cause for this difference in translocation speed has to be searched for in the only difference between the pseudoknot and the hairpin sequence: Both sequences are equal in terms of base number and initial geometry, their pore-next stems even in terms of binding energy. However, unlike the hairpin sequence, the pseudoknot sequence is stabilized by additional base-pair bonds forming a second stem (stem B).

If pore strength is low, these base-pairs remain stable until stem A is totally open. This is shown in Fig. 2.65 for a pore strength of ϵ=5.0 , where the integrity of stem B does not begin to decrease until the integrity of stem A has decreased almost to its minimal value. With increasing pore strength this effect gradually disappears, until the opening processes of stem A and stem B start almost simultaneously. To see this, one has to compare Fig. 2.64 and Fig. 2.66, where the integrities of stem A resp. stem B are plotted against time for different pore strengths. This almost simultaneous stem opening for high pore strength can be explained as follows: The pore force is not only strong enough to deform and unzip the pore-next base-pairs of stem A but also – due to the

short loop 1 and the high friction coefficient of the 3'-base – to immediately exert an appreciable tensile force on the base-pairs of the pore-remote end of stem B.

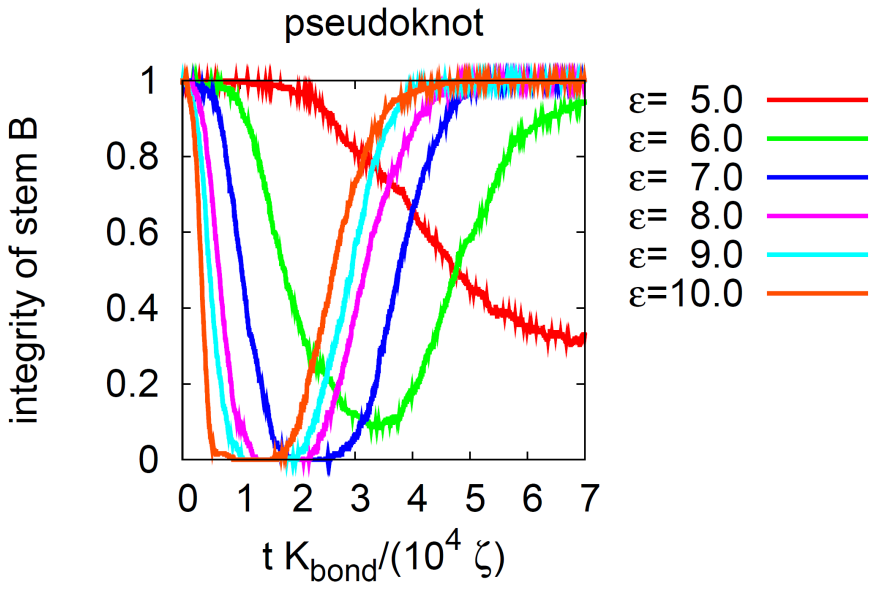


Fig. 2.66: Integrity of stem B of the pseudoknot plotted against time for different pore strengths, to be compared to Fig. 2.64, where the same is done for the integrity of stem A. For low pore strengths the unzipping process of stem B does not start until stem A is almost completely open, whereas for high pore strengths both stems start unzipping almost simultaneously.

Since for low pore strengths stem B remains closed during the whole unzipping process of stem A, it can be regarded as a single particle with very low motility: The particle's friction coefficient is the sum of the friction coefficients of all bases stem B consists of, and the large friction coefficient of the last base dominates this sum. Therefore, the drift rate of the particle during the translocation process is inversely proportional to the friction coefficient of the last base:

$$k_{\mathrm{B}}^{\mathrm{drift}} \propto \frac{1}{\zeta_{\mathrm{B}}} = \left(\sum_{\text{base } i \in B} \zeta_i \right)^{-1} \approx \frac{1}{\zeta_N} \tag{2.67}$$

The particle has to be dragged towards the pore in order to allow the translocation of the bases of the pore-facing helix of stem A after unzipping the corresponding base-pairs: Because the translocation of a single base of this helix requires a drift motion of the stem B "particle" before the corresponding base-pair in stem A can be unzipped, the translocation rate is related to the rates of both sub-processes via an addition of characteristic times:

$$\frac{1}{k_{\text{base of stem A}}^{\mathrm{transloc}}(\epsilon, \{\zeta_i\})} \approx \frac{1}{k_{\mathrm{B}}^{\mathrm{drift}}(\{\zeta_i\})} + \frac{1}{k_{\text{bp of stem A}}^{\mathrm{unzip}}(\epsilon)} \tag{2.68}$$

For large friction coefficients of the last base of the sequence ($\zeta_N \gg \zeta$) the translocation process of the pore-facing helix of stem A is impeded strongly by the dragging process of stem B, which results in a very low translocation rate:

$$k_{\text{base of stem A}}^{\mathrm{transloc}} \approx k_{\mathrm{B}}^{\mathrm{drift}} \propto \frac{1}{\zeta_N} \quad \text{because} \quad k_{\mathrm{B}}^{\mathrm{drift}} \ll k_{\text{bp of stem A}}^{\mathrm{unzip}} \tag{2.69}$$

This reducing effect on the translocation rate increases if the friction coefficient of the last base increases, as demonstrated in Fig. 2.67, where the mean fraction of translocated bases is plotted against time for different values of this friction coefficient at equal pore strength ϵ=5.0 , thereby focusing on the time regime of stem A unzipping.

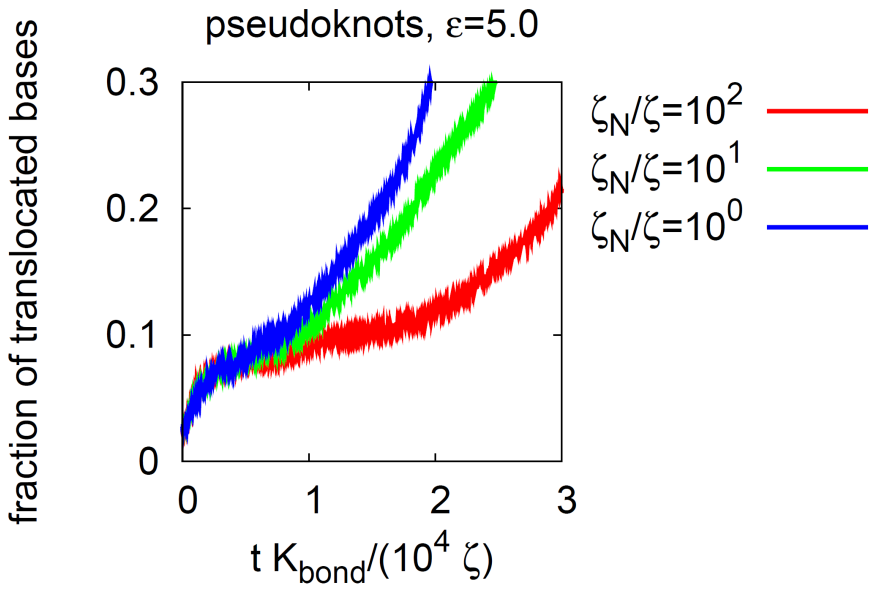


Fig. 2.67: Mean fraction of translocated pseudoknot bases plotted against time at low constant pore strength for different friction coefficients of the last base. The process of unzipping stem A (time region with low slopes) is slowed down more and more when the friction coefficient increases.

Interpretation of the observations concerning the hairpin

This rate-reducing effect is not observable in the case of the hairpin sequence (data not shown), so that, unlike the unzipping of the pore-next stem of a pseudoknot, the unzipping of a hairpin stem cannot be slowed down appreciably by increasing the friction coefficient of the 3'-end of the sequence. The reason is that, despite a large friction coefficient of the last base, the hairpin will always lack the base-pairs needed for gluing bases together to a particle which contains the last base. Therefore, the bases of the pore-facing helix of the hairpin stem do not sense the high friction coefficient of the last base as long as the stem is still unzipping: Their translocation process is not impeded by the last base of the sequence, thus being controlled only by the local stem unzipping rate:

$$k^{\text{transloc}}_{\text{base of stem A}} \approx k^{\text{unzip}}_{\text{bp of stem A}} \tag{2.70}$$

2.5.3.7 Translocation with realistic angular energy parameters

Up to now the translocation process of model RNA with realistic base-pair melting behavior but too low torsional modulus and persistence length has been studied. The goal of this chapter is to simulate the translocation of hairpins and pseudoknots with realistic torsional modulus and persistence length.

To unzip an RNA stem the pore drift energy has to be larger than the effective binding energy per base-pair. For realistic angular energy parameters, this effective binding energy is increased largely by the base-pair stabilizing forces generated by the binding predicate functions of the angular potential contributions (cf. chapter 2.4). Therefore, large pore drift energies are needed for the translocation of such model RNA molecules. Because of these high pore drift energies, the pore barrier energy has to be increased in

order to avoid barrier crossing events and the pore diameter has to be decreased in order to avoid double-stranded translocation by stem deformation. The low angular energy RNA parameter set used for the comparison between hairpin and pseudoknot translocation up to now for reasons of realistic base-pair melting behavior,

$$\frac{E_{\mathrm{LJ}}}{\mathrm{k}T}=4,\ \frac{K_{\mathrm{bond}}r_0^2}{\mathrm{k}T}=400,\ \frac{E_{\mathrm{tor}}}{\mathrm{k}T}=4=\frac{E_{\mathrm{bend}}}{\mathrm{k}T},\ \frac{E_{\mathrm{tilt}}}{\mathrm{k}T}=4,\qquad \frac{R}{r_0}=3,\ \phi_0=\frac{2\pi}{11},\ \theta_0=0.39\pi\ , \tag{2.71}$$

produces slow single-stranded translocation for a pore parametrized by

$$\frac{E_{\mathrm{barrier}}}{\mathrm{k}T}=2000,\ \frac{E_{\mathrm{drift}}}{\mathrm{k}T}=20,\qquad \frac{d_{\mathrm{pore}}}{r_0}=1,\ \frac{l_{\mathrm{pore}}}{r_0}=2\ . \tag{2.72}$$

The high angular energy RNA parameter set, which has been found by exact calibration of torsional modulus and persistence length (cf. chapter 2.4),

$$\frac{E_{\mathrm{LJ}}}{\mathrm{k}T}=4,\ \frac{K_{\mathrm{bond}}r_0^2}{\mathrm{k}T}=1600,\ \frac{E_{\mathrm{tor}}}{\mathrm{k}T}=160=\frac{E_{\mathrm{bend}}}{\mathrm{k}T},\ \frac{E_{\mathrm{tilt}}}{\mathrm{k}T}=4,\qquad \frac{R}{r_0}=3,\ \phi_0=\frac{2\pi}{11},\ \theta_0=0.39\pi\ , \tag{2.73}$$

however, needs a different pore parameter set to allow slow single-stranded translocation and thus a comparison between hairpin and pseudoknot translocation:

$$\frac{E_{\mathrm{barrier}}}{\mathrm{k}T}=10000,\ \frac{E_{\mathrm{drift}}}{\mathrm{k}T}=350,\qquad \frac{d_{\mathrm{pore}}}{r_0}=0.5,\ \frac{l_{\mathrm{pore}}}{r_0}=2 \tag{2.74}$$

This comparison is done for the known hairpin (5-bp stem) and pseudoknot (5-bp stem A, 6-bp stem B) sequences (all base beads have identical friction coefficients now). One averages over 38 non-rejected hairpin configurations (out of 60) and over 44 non-rejected pseudoknot configurations (out of 60). Chain pausing or double-stranded translocation has not been observed in this regime, only single-stranded translocation and chain rejection. The comparison between hairpin and pseudoknot translocation via the time evolution of the RNA-within-pore-channel probability (Fig. 2.68), the time evolution of the fraction of translocated bases (Fig. 2.69) and the time evolution of the stem integrity (Fig. 2.70) shows respectively:

- a non-vanishing difference τ_{Prob} in the total translocation times of hairpin and pseudoknot: the total hairpin translocation process is faster than the total pseudoknot translocation process;
- that the unzipping rate of the hairpin stem is approximately 1.6 times larger than the unzipping rate of stem A of the pseudoknot;
- that the minimum of the integrity of stem A of the pseudoknot occurs later than the minimum of the integrity of the hairpin stem, leading to a time difference τ_{int} between the respective states of open primary stems; even later, stem B of the pseudoknot is open;

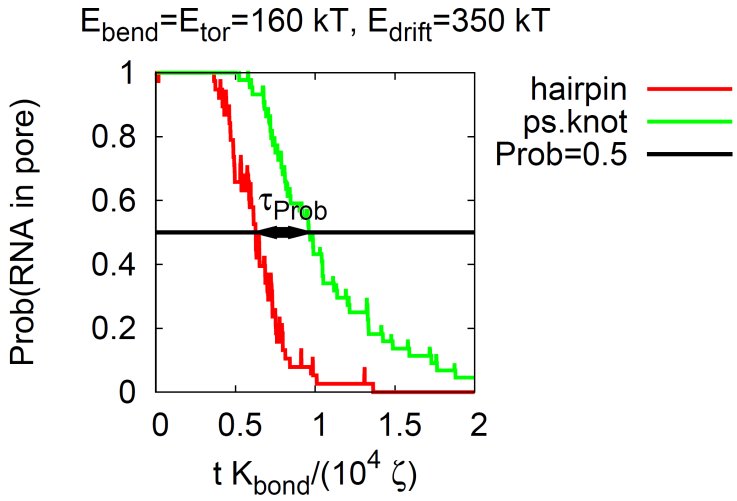


Fig. 2.68: Comparison of the probabilities that the translocation is still in progress in the hairpin and in the pseudoknot case for realistic torsional modulus and persistence length. The respective total translocation times show a difference τ_{Prob}: The pseudoknot is slower than the hairpin, as far as the total translocation process is concerned.

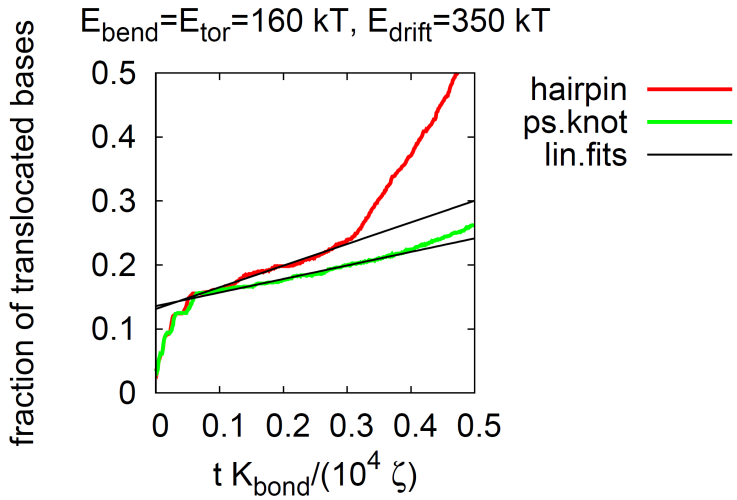


Fig. 2.69: Mean fraction of translocated bases vs. time in the pseudoknot and in the hairpin case for realistic torsional modulus and persistence length. The focus lies on the unzipping of the respective primary stems. Despite equal numbers of base-pairs, the unzipping rate of the hairpin stem is approximately 1.6 times larger than the unzipping rate of stem A of the pseudoknot (cf. slopes of the corresponding black fit lines).

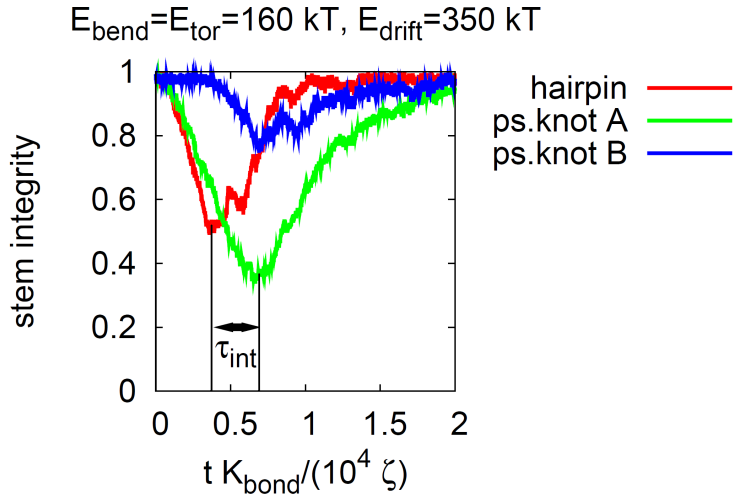


Fig. 2.70: Comparison between the integrity of the hairpin stem (red) and the integrity of stem A of the pseudoknot (green) for realistic torsional modulus and persistence length. The minimum positions differ (time difference τ_{int}): The hairpin stem is unzipped completely at an earlier point of time than stem A of the pseudoknot. Stem B of the pseudoknot does not start unzipping until stem A is almost completely open.

In the comparison between hairpin and pseudoknot performed in chapter 2.5.3.6 the high friction coefficient of the last base of the sequence caused a high effective total friction coefficient of stem B of the pseudoknot, thus impeding its axial translation, which in turn retarded the unzipping of stem A. Although the last base of the sequence has the same low friction coefficient as all other bases in the current translocation simulation now, surprisingly the unzipping of stem A is retarded again. What could be the reason for this?

In spite of high stem torsional modulus, no difference could be observed between pseudoknot and hairpin as far as axial rotation of their primary stems during unzipping is concerned (none of them rotates). However, such a difference was proposed by Plant and Dinman in their torsional restraint model [12]. That it could not be observed here is clear from the cylindrical shape of the pore channel and the point-shape of the traversing bases, which lack any geometric information about direction.

To circumvent this problem, one could choose an ellipse instead of a circle as the slice plane of the pore channel, causing the pore-adjacent cis-side base-pair to orient according to the long axis of the ellipse before being unzipped. Due to the high torsional modulus this change of base-pair direction would propagate along the stem, causing it to rotate. Such rotation of the primary stem would progress freely in the hairpin case, whereas it would be impeded by the secondary stem in the pseudoknot case. Because the current pore model is cylindrical, however, the reason for the retarded unzipping of stem A of the pseudoknot – if compared to the unzipping of the hairpin stem – cannot be impeded stem rotation.

However, the high torsional modulus has other effects: In the still folded parts of the hairpin stem and stem A of the pseudoknot as well as in the whole stem B of the pseudoknot it leads to relatively stiff torsion angles between successive base-pairs and – as a consequence – to a relatively fix short axial rise per base-pair. Due to this reduction of angular and axial degrees of freedom the presence of stem B of the pseudoknot can influence the unzipping process of stem A appreciably, which could explain that stem A of the pseudoknot unzips slower than the isolated hairpin stem.

An alternative explanation could be the stronger repulsive force field of the pore barrier, which tries to repel the closed – and due to the high torsional modulus axially shortened – stem B from the pore plane, thus impeding the translation of the pore-next strand of stem A towards and through the pore. In the hairpin case no stem B exists, making such repulsion effects much smaller than in the pseudoknot case.

On the whole, two mechanisms have been identified within this model that could cause ribosomal pausing of pseudoknots in nature: Firstly, the unzipping of stem A can be impeded by an intact stem B by preventing free translation of stem A due to the large effective friction of long base sequences upstream stem B (simulated by using a high friction coefficient for the last base of the sequence). Secondly, the unzipping of stem A can be impeded by an intact stem B due to strongly restricted angular and axial degrees of freedom in all base-paired regions (simulated by using a realistically high torsional modulus).

2.6 Conclusion

In order to study the folding, unfolding and unzipping dynamics of RNA molecules in three-dimensional space a novel coarse-grained RNA model was developed and used in Brownian dynamics simulations. Although the complex geometry and interactions of real RNA biomolecules are modeled in a very coarse-grained way (bead-spring chain with binding interactions between compatible bases and continuously activated base-pair-orientating interactions between fully-formed base-pairs), the model is able to produce geometrically exact tertiary structures from initially unfolded chain configurations. For example, for a hairpin sequence the formation of a straight regular stem double-helix could be observed. By measuring base-pair binding probabilities and comparing them to stem opening probabilities known from experiment, it is possible to calibrate the model system, which switches permanently between the folded and the unfolded state.

The mechanical properties of the folded state, which originate mainly from the orientating interactions between base-pairs, were tested below the unfolding temperature by deformation simulations and can be used to adjust the model parameters on the basis of the results obtained for stem bending stiffness, stem torsional stiffness and longitudinal stem elasticity to experimental data. A parameter set with realistic bending stiffness and torsional stiffness was found by simulating a hairpin molecule in thermal equilibrium without external forces.

After having studied the static and dynamical properties of the free model molecule in detail the model was extended by a short, narrow pore channel through which the molecule is driven by the use of a local potential gradient. By this model extension it is possible to simulate the unzipping and translocation behavior of arbitrarily complex RNA tertiary structure motifs through nanopores or ribosomal translation channels.

The translocation behavior of the model molecule was characterized – in terms of energy and temperature as well as in a time-resolved way – using a generic hairpin structure. Neglecting a regime of double-stranded translocation, which only occurs if the folded structure is deformable by the pore force to an extent where the stems do not need to unzip in order to allow their bases to pass the pore, the parameter space spanned by temperature and the ratio between pore drift force and base-pair binding energy can be divided into three relevant regimes of chain behavior: In the regime of chain rejection the folded structure is rejected from the pore by entropic forces, in the regime of chain pausing it remains in the pore channel, because the pore force is to weak to break base-pairs or to deform the stems sufficiently, and in the regime of single-stranded chain translocation the folded structure unzips in order to pass the pore.

Of course, of these three regimes the regime of single-stranded translocation proved most important for all further observations: Recording simultaneously for each point of time the mean fraction of RNA bases on either side of the pore as well as the mean frac-

tion of base-pairs fully-formed within each stem the RNA tertiary structure consists of, it was possible to decompose the translocation process into a number of distinct and subsequent time regimes whose sequence (for example initial threading, stem unzipping, free drift, stem re-zipping) corresponds directly to the used base sequence. This was shown for both the hairpin (1 stem) and the H-type pseudoknot (2 stems). Most prominent among these time regimes is the regime of stem unzipping, where the time per base passing the pore is usually dominated by the (long) unzipping time per base-pair, which in turn depends on base-pair binding energy.

In order to identify the mechanisms that slow down the translocation of certain H-type pseudoknots compared to that of hairpins in nature, thus being responsible for ribosomal pausing, which in turn is necessary for ribosomal frameshift during RNA translation, the translocation processes of both motifs were simulated and compared to each other. Since the total pore translocation time of an arbitrary RNA tertiary structure is usually dominated by the sum of its stem unzipping times, and to filter out those time differences that emerge trivially from different numbers of native base-pairs in the hairpin and the pseudoknot case, only the unzipping processes of the respective pore-next stems, which had equal length and total binding energy, were examined in the comparison.

For not too low pore forces the unzipping of these stems proceeded in an almost identical manner. Therefore, the hypothesis that the low motility of the closed second stem of the pseudoknot impeded the opening process of the pore-next stem (for example by imposing rotational constraints on it) could be falsified in this regime: The pore-next stem of the pseudoknot unzips in the same way as an isolated stem would do.

However, if pore forces are low and both the hairpin and the pseudoknot motifs have long upstream chain rests, which is usually the case in ribosomal translocation of RNA, such an impeding effect is observable noticeably for pseudoknots. Thus, in this regime the pore-next stem of the pseudoknot cannot be regarded as isolated any longer: Because the base-pair bonds in its closed second stem connect the long upstream chain rest mechanically to the 5'-helix of the pore-next stem, the low motility of this chain rest is able to slow down the translocation of that helix (and the unzipping process of the pore-next stem) appreciably. By contrast, the hairpin lacks such connecting base-pair bonds. Therefore, the low motility of the upstream chain rest cannot slow down the unzipping process of its stem.

To conclude, there is strong evidence that ribosomal pausing during the translocation of an H-type pseudoknot is caused by geometrical constraints imposed on its pore-next stem by long upstream RNA structure parts. These constraints are conveyed by the base-pairs of its second stem, which function as mechanical connections.

The obtained simulation results are consistent with the experimental findings of Plant and Dinman [12], who compared the ribosomal frameshift efficiencies of pseudo-H-type-pseudoknots. These were hairpins with a 5'-end slippery sequence (12 bases) and identical stems (13 base-pairs) but additional RNA strands (~15-20 bases) attached to their long 3' dangling ends (22 bases) in a way that a pseudo-stem B and thus a pseudoknot-like binding topology is formed. According to the 'torsional restraint model'

[11], they proposed that the appreciable frameshift efficiencies gained for those artificial topologies – if compared to a pure hairpin – were caused by a loss of the freedom of rotation of stem A due to the constructs imitating a stem B. However, no kinetic observables like unzipping rate or angular velocity of stem A were studied in these experiments. Therefore, one could as well propose an alternative (or additional) mechanism: The attached RNA strands may slow down the translation of the (remaining intact) pseudostem B towards the pore, thus impeding the translocation of the pore-next strand of stem A and the unzipping process of stem A. This mechanism was observed in the RNA translocation simulations in this thesis when giving a high friction coefficient to the last bases of the pseudoknot and hairpin sequences to be compared. Simulations with realistically high torsional modulus but normal friction coefficient of the last base as well yielded slower unzipping of stem A of the pseudoknot, possibly indicating torsional restraint effects in pseudoknot translocation (although the corresponding hairpin stem did not rotate during unzipping).

Experiments explicitly studying the unfolding and refolding kinetics of hairpins and H-type pseudoknots by the means of unzipping the corresponding RNA motifs with optical tweezers were performed by Green and co-workers [122]. Their experimental setup is comparable to the pore unzipping simulations in this thesis, at least for hairpins: The unzipping forces act perpendicular to the hairpin stem and are localized at its open end. For pseudoknots, however, the optical tweezer setup and the pore simulation setup differ sightly in the directions of unzipping forces:

The pore simulations show local forces at the pore-adjacent end of stem A, which are directed stem-perpendicular (the attractive drift force in the pore channel vs. the repulsive pore barrier force), as well as a drag force emerging from the high effective friction coefficient of the intact stem B. This axially-directed force counteracts the shearing and unzipping of stem A by impeding the translation of its pore-next strand.

By contrast, in the optical tweezer setup there is no pore and the unzipping forces act simultaneously on opposite ends of the pseudoknot sequence, by that trying to shear and unzip its two stems at the same time. Consequently, the experimentally measured pseudoknot unzipping rate corresponds to both of its stems, whereas a crucial element in the pore simulations of this thesis was the serial unzipping of stem A and stem B: This allowed to study the kinetic influence stem B exerted on the unzipping of stem A by measuring the unzipping rate of stem A separately.

Apart from these differences, the results of Green et al. are consistent with the results in this thesis: Although the studied structures in the pseudoknot case (a combination of stem A and stem B in the optical tweezer setup / stem A alone in the pore simulation setup) have been compared to thermodynamically equivalent counter-parts in the hairpin case in both setups, their unzipping kinetics – quantified by the unzipping rate in both setups – is appreciably slower in the pseudoknot case than in the hairpin case (cf. the slopes in Fig. 2.62). Furthermore, in both setups, this unzipping rate is more independent on the unzipping force in the pseudoknot case than in the hairpin case (cf. the slopes in Fig. 2.60 and Fig. 2.61).

On the whole, the main result of the comparative pore translocation simulations preformed in this thesis, that stem B of the pseudoknot impedes the unzipping of stem A by imposing geometrical constraints on it, by that making the unzipping kinetics of stem A slower than the unzipping kinetics of the hairpin stem, is consistent with the experimental findings discussed above.

3 Protein target search on DNA

3.1 Model

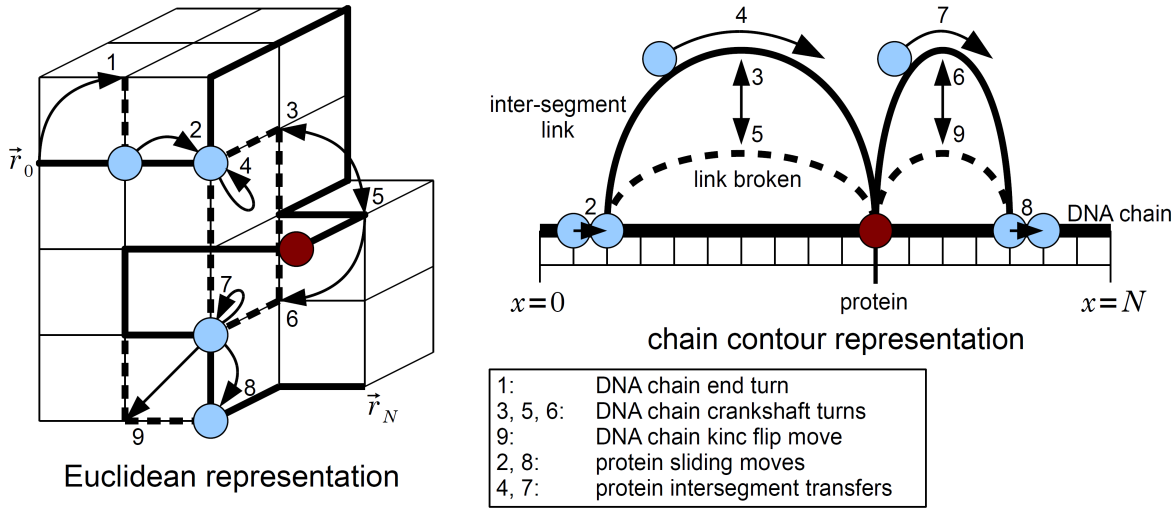


Fig. 3.1: Simultaneous dynamics of a DNA chain on a simple cubic lattice (rate K_{DNA}) and of a protein on the contour of the chain (rate K_{protein}). Each local DNA chain move (Advanced Verdier Stockmayer move set) changes not only the three-dimensional DNA conformation but also the pattern of intersegment links the protein uses besides sliding when traveling on the chain contour. Therefore the dynamics of conformational changes of the DNA affect protein transport in a significant way.

The model represents – in a coarse-grained way – the simultaneous dynamics of a **DNA chain** in three-dimensional space and of a **protein** sliding and jumping on the DNA chain's contour, as one can see from Fig. 3.1:

- Euclidean space is discretized by replacing it with an infinite cubic lattice.

- The DNA chain's initial conformation is formed by a three-dimensional trajectory of an unbiased random walker on that lattice, so that it consists of N unit length segments.
- The protein is a point-shaped particle solely residing on the N+1 points of the DNA chain, the so-called chain contour space.

The simultaneous dynamics of the DNA chain on the one hand and the protein walker on the other hand are simulated using the kinetic Monte Carlo method (Gillespie algorithm). Of these, the chain dynamics is implemented by the Advanced Verdier-Stockmayer algorithm [123][124], where every move changes only a local region of the chain. By allowing chain points to overlap (more than one chain point per lattice point) the random walk nature of the chain is preserved by that algorithm. Furthermore, the move set is ergodic and makes the chain exhibit Rouse-type dynamics. The three types of chain moves, each of which is performed with equal rate K_{DNA} are:

- **DNA chain end turn.** Both end segments of the chain can be rotated by 90 degrees, i.e. there are four possibilities to rotate an end segment.
- **DNA chain crankshaft turn.** A set of three adjacent chain segments forming a crankshaft-shaped region can be rotated by 90 degrees around its axis, either in forward or in backward direction.
- **DNA chain kink flip move.** A set of two adjacent chain segments which are perpendicular to each other ("kink") can be changed by swapping the directions of the two segments and shifting their connection point along the diagonal of the lattice square they span.

The combination of such chain moves changes position and shape of the modelized three-dimensional DNA chain conformation in the surrounding cubic lattice, i.e. in *Euclidean space*. By contrast, protein moves, which are all performed with equal rate K_{protein}, change only the position of the protein on the chain contour, which means one-dimensional transport carried out exclusively in *chain contour space*. The reason is that the protein is not allowed to detach oneself from the chain and enter surrounding lattice points but can only move from one point of the DNA chain to another. The two types of protein moves used here are:

- **Protein sliding move.** The protein performs an unbiased random walk step of unit length in chain contour space, thus arriving at an adjacent point of the DNA chain contour.
- **Protein intersegment transfer.** If the DNA chain forms a loop onto itself so that the same point in Euclidean space (in this model: the same point of the simple cubic lattice) is occupied by two distinct points of the DNA chain, which are possibly remote in chain contour space (in this model: separated by a number of chain segments), a protein sitting at one of these two points is allowed to jump to the other one. Since such a loop is a connection between different segments of the DNA chain it is called an intersegment link whereas protein jumps over such intersegment links are named intersegment transfers.

If sliding was the only type of protein move, protein transport in chain contour space would be nothing else than a diffusion process on a finite one-dimensional system. Especially, protein dynamics would be unaffected by the dynamics of the DNA chain. Adding the possibility of intersegment transfer has two important effects:

- The intersegment links connecting remote points of the chain contour form a network for protein transport with non-trivial topological and geometrical properties, even if the chain conformation is stationary (In this stationary case protein sliding is necessary in order to reach different intersegment links.).
- Although chain dynamics cannot influence the position of the protein in chain contour space directly, they do influence the protein transport process, because changes in the DNA chain conformation mean changes in the network of intersegment links.

The interplay of chain dynamics and protein dynamics is shown in Fig. 3.1, where the effects of chain moves on the shape of the DNA chain conformation in Euclidean space (left part of the figure) as well as on the structure of the network of intersegment links in chain contour space (right part of the figure) and thus on the path which the protein can take are exemplified by means of a fictitious sequence of DNA chain and protein moves (the protein sits initially at chain point 1):

1. One DNA chain end turns by 90 degrees. The intersegment linking pattern is unaffected by that.
2. The protein slides from chain point 1 to its neighbor chain point 2.
3. A DNA chain crankshaft rotates around its axis in a way that one of its points (chain point 10) coincides with the protein location in Euclidean space. By that, an intersegment link connecting chain point 2 (protein position at the moment!) and chain point 10 is created.
4. Using this intersegment link, the protein performs an intersegment transfer from chain point 2 to chain point 10 (part of the crankshaft region), which are identical in Euclidean space at the moment.
5. The same DNA chain crankshaft rotates back into its original position, taking the protein along. The corresponding intersegment link between chain point 2 and chain point 10 is destroyed again.
6. The same DNA chain crankshaft rotates even further, thus causing an Euclidean coincidence between its chain point 10 (still occupied by the protein) and chain point 14. Thereby a new intersegment link between these two chain points is generated.
7. The protein performs an intersegment transfer using this intersegment link, thus leaving the DNA chain crankshaft and arriving at chain point 14.
8. The protein slides from chain point 14 to its neighbor chain point 15.

9. A kink of the DNA chain is flipped, thereby separating chain point 14 from chain point 10 in Euclidean space and destroying the intersegment link between these two chain points.

3.2 Protein transport simulations on the DNA chain

3.2.1 Observations

Transport on long DNA chains (here: length N=5000) is simulated in the following way: First an instance of a chain is generated as a random walk on the cubic lattice. Then the trajectory of the protein walker initially sitting at the central point of the chain contour is recorded over a certain number of (Poissonian) time steps, during which the protein as well as the chain are moving. Then another instance of the chain is generated at random, the protein starts again and so on. Collecting the information from all these protein trajectories (16000 trajectories for $K_{\mathrm{DNA}}=0$, 1200 for $K_{\mathrm{DNA}}=500$, 1500 for $K_{\mathrm{DNA}}=2000$ and 900 for $K_{\mathrm{DNA}}=10000$) in chain contour space one obtains the time evolution of the protein density along the chain contour, the so-called probability density function (PDF) of the protein:

$$W(x,t) \tag{3.1}$$

Due to the initial condition, this density is symmetric with respect to the central point of the chain contour where it is typically peaked. As time elapses, this peak becomes broader and broader, until the protein density is equally distributed along the chain contour. A measure of this broadening process is the interquartile distance $\Lambda(t)$, which is the (contour) distance between the first and the last quartile of the PDF and characterizes its width according to the universal scaling form:

$$W(x,t)=\frac{1}{\Lambda(t)}\rho\left(\frac{x}{\Lambda(t)}\right) \tag{3.2}$$

A transport process is called diffusive if the width of the corresponding PDF obeys the following proportionality relation:

$$\Lambda(t)\propto t^{1/2} \tag{3.3}$$

By contrast, subdiffusive transport processes yield $0<\kappa<1/2$, superdiffusive transport processes $\kappa>1/2$ in the generalized proportionality relation

$$\Lambda(t)\propto t^{\kappa}. \tag{3.4}$$

Remarkably, Fig. 3.2 shows that for finite DNA rates (here: $K_{\mathrm{DNA}}=500,\ 2000,\ 10000,\ K_{\mathrm{protein}}=1$) the time evolution of the protein PDF's width exhibits a cross-over between a short-time superdiffusive regime (exponent $\kappa\approx 1.6$, expected value for infinite chain length is $\kappa=2$, deviations emerge only due to finite size effects which reduce the effective exponent) as well as a long-time diffusive regime. The corresponding cross-over time $t_c(K_{\mathrm{DNA}})$ is rate-dependent and increases with increasing

K_{DNA}. This dynamic cross-over between superdiffusive and diffusive protein transport behavior can be expressed for infinite chain length by:

$$\Lambda(t) \propto \begin{cases} t^2 & \text{for } t \ll t_c(K_{DNA}) \\ t^{1/2} & \text{for } t \gg t_c(K_{DNA}) \end{cases} \tag{3.5}$$

That for times even longer the width approaches a stationary value which equals half the contour length of the DNA chain is a mere finite size effect, which can be neglected in this discussion.

On the other hand, protein transport for vanishing and diverging DNA rates does not show such a cross-over: Protein transport on frozen DNA chains ($K_{\mathrm{DNA}}=0$, $K_{\mathrm{protein}}=1$) is diffusive for all times because the cross-over time vanishes, protein transport on infinitely fast moving DNA chains ($K_{\mathrm{DNA}}\to\infty$, $K_{\mathrm{protein}}=1$) is superdiffusive (again: $\kappa\approx 1.6$ for $N=5000$, $\kappa\to 2$ for $N\to\infty$) for all times because the cross-over time diverges. This limit of infinitely fast changing shape of the DNA chain can be achieved by replacing the DNA chain by an entirely new chain instance generated as a random walk trajectory (preserving the protein position in contour space) before each protein move. This has already been shown by Sokolov and co-workers [73]. By contrast, the fast chain limit in Fig. 3.2 is simulated by the equivalent procedure of drawing the jump lengths for the protein from the distribution of intersegment link lengths of the chain before performing a jump. Since this distribution is position-dependent, it incorporates effects of finite chain size correctly. In the fast chain limit, the probability density function of the protein was averaged over 20000 trajectories.

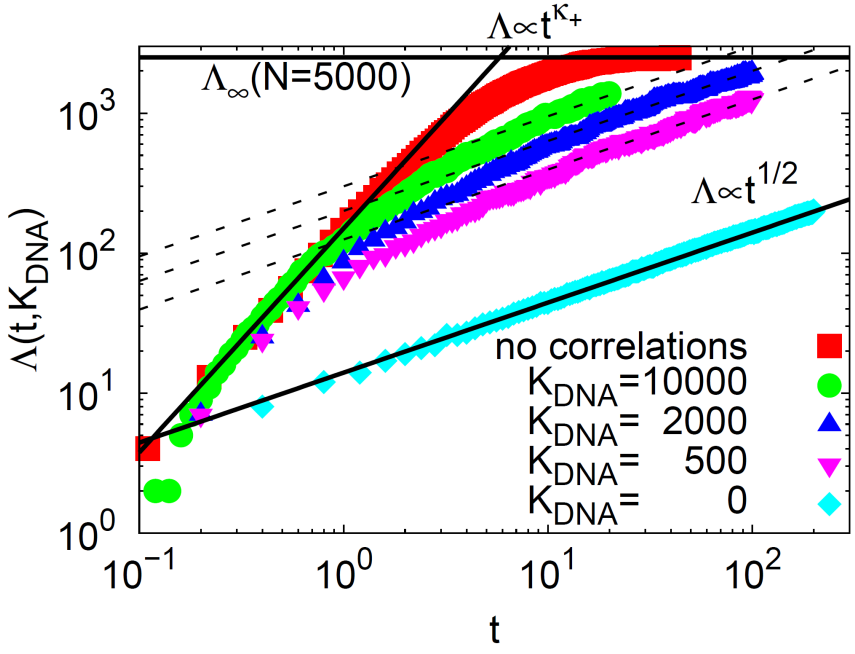


Fig. 3.2: The protein's transport behavior for different DNA chain rates can be characterized by viewing the time evolution of the width of the protein density along the chain contour. For frozen chains highly correlated protein jumps result in purely diffusive transport (cyan). For non-zero but finite chain rates these geometrical correlations are broken dynamically, resulting in a cross-over in transport behavior. The faster the chain moves, the less correlated the protein jumps are, which is reflected in the growing short time regime of superdiffusive transport and the shrinking long time regime of diffusive transport (magenta, blue, green). For infinitely fast moving chains the entirely uncorrelated jumps of a Lévy flight yield purely superdiffusive transport (red).

3.2.2 Theoretical explanation of the observations

Before the intermediate regime of finite DNA rates can be understood, the two opposite limits of frozen and infinitely fast moving DNA chains have to be examined.

The origin of the superdiffusive protein transport behavior in the limit of infinitely fast moving DNA chains is clear: Since before each protein move the protein finds an entirely new DNA chain conformation and – which is even more important – an entirely new intersegment linking pattern, each intersegment transfer which it performs is totally independent from its predecessor. This independence and entire absence of correlations between intersegment transfers is like drawing subsequent protein jump lengths independently from a broad power-law jump length distribution:

$$p(s) \propto s^{-(1+\mu)} \quad \text{with} \quad \mu = \frac{1}{2} \tag{3.6}$$

This relation emerges directly from the distribution of intersegment link lengths and – since every DNA chain conformation is a three-dimensional random walk – is in turn connected closely to the distribution of return times [125] of a random walker in three-dimensional space:

$$w_{\text{return}}(t) \propto t^{-3/2} \tag{3.7}$$

According to the Lévy-Gnedenko generalized central limit theorem [125][126][127][128] [129][130] a continuous time random walker (CTRW, introduced in [131], extended to explain anomalous diffusion in [132]) whose waiting times are drawn from a Poissonian distribution and whose jump lengths are independent and identically distributed random variables drawn from a broad $(0<\mu<2)$ power-law jump length distribution performs a so-called Lévy flight. The corresponding PDF, which measures the probability that the walker is at position x after time t, is a Lévy stable law of index μ [73]:

$$W(x,t) \sim \frac{1}{t^{1/\mu}} L_\mu(x/t^{1/\mu}) \tag{3.8}$$

Obviously, the width of this PDF grows superdiffusively, as observed in the simulation of the fast chain limit:

$$\Lambda(t) \propto t^{1/\mu} = t^2 \tag{3.9}$$

Unlike the fast chain limit, the limit of frozen DNA conformation has very paradoxical properties which cannot be described simply by the Lévy formalism but have to be examined further. Although the PDF in this limit has power-law tails just as expected for a Lévy type PDF, its width grows diffusively as expected for a Gaussian type PDF (remarkably, the PDF is neither of Gaussian nor of Lévy type [73]):

$$\Lambda(t) \propto t^{1/2} \tag{3.10}$$

The cause of this is proposed to be strong geometrical correlations [73] between the intersegment transfers, whose lengths – although identically distributed according to the broad power-law distribution (equation 3.6) – are not independent random variables anymore. The cross-over from superdiffusive to diffusive transport behavior present for finite DNA rates could then be explained by a crossover between an uncorrelated regime

at short times and a correlated regime at long times. The fact that for increasing DNA rate the superdiffusive regime increasingly dominates the transport behavior can be explained by a dynamic destruction of these correlations which is the more effective the faster the chain's intersegment linking pattern changes. Therefore, in order to understand this cross-over it is essential to characterize these correlations and their effects on protein transport in detail.

3.3 Protein transport and correlations

3.3.1 Lévy-type superdiffusion vs. quasi-diffusion and the role of correlations

The faster the DNA chain changes its conformation in three-dimensional space, the faster existing intersegment links break and are recreated elsewhere. The limit in which an infinite number of such conformation changes occurs before the protein performs a new move can be simulated either by creating a new random chain conformation (and setting the protein on it, preserving its contour coordinate) before every protein move or simply by having the protein perform a Lévy-flight on a finite line without intersegment links which is as long as the DNA chain. The jump length distribution from which one has to draw independently before every protein jump in this flight must be chosen equal to the intersegment link length distribution of the DNA chain (equation 3.6). For either method, the resulting distribution of the protein's contour coordinate spreads superdiffusively (equation 3.9).

In the opposite limit, where the chain conformation is frozen in space while the protein is performing its moves, all existing intersegment links are preserved in perpetuity, thus yielding a time-invariant topology of intersegment transfer pathways placed onto the chain contour. Therefore, these pathways can be used by the protein repeatedly in forward as well as in backward direction. This would be impossible in the limit of infinitely fast changing chain conformation with its randomly disappearing and reappearing intersegment links. Different from that limit, one obtains diffusive protein transport along the chain contour for frozen chains (equation 3.10).

Whereas Lévy-type superdiffusive protein transport on infinitely fast moving DNA chains is a result explicable theoretically by the Lévy formalism (cf. chapter 3.2.2), diffusive transport on frozen chains is a paradox result [73], which cannot be explained by this formalism in a straight forward way: Since both limits (infinitely fast changing and frozen chain conformation) feature identical (broad) length distributions of intersegment links (equation 3.6), one would expect identical (superdiffusive) time behavior of the width of the distribution of the protein position on the chain contour as well. However, superdiffusion occurs only in one of these limits. As a consequence, one has to postulate that, different from the simple limit of infinitely fast changing chain conformation which lacks correlations completely, in the limit of frozen chain conformation there are strong correlations which cause transport to be diffusive [73].

Therefore, the relevant questions are: How can these correlations be characterized and why is the resulting transport behavior exactly diffusive?

3.3.2 Spatial vs. temporal correlations

As mentioned above, for frozen chain conformations the resulting pattern of intersegment links in contour space is frozen as well, allowing intersegment transfers of the protein that are highly repetitive and recurrent because they are always reversible. However, the resulting temporal correlations between measured properties of intersegment transfer events (for example jump length, contour coordinate before and after the jump) occurring at different points of time are not sufficient to cause diffusive transport and thus have to be regarded as irrelevant.

To show this, one has to create an intersegment link pattern which obeys the intersegment link length distribution of a real random walk chain conformation and – although kept frozen during the whole transport process – exhibits superdiffusive transport behavior. Of course, the raw intersegment link pattern taken from a random walk chain conformation will still lead to diffusive transport. But if one destroys all geometrical correlations of higher order by shifting the positions of its intersegment links randomly while preserving their link lengths in order not to violate the intersegment link length distribution, the subsequent transport process on the resulting frozen intersegment link pattern is purely superdiffusive. Equal results are obtained for intersegment link patterns created artificially by placing a number of links which is comparable to that which occurs in a real chain conformation on an initially empty contour (cf. Fig. 3.3): First their position x on the chain contour (length N) is chosen at random and then their length s is drawn from a (position-dependent) intersegment link length distribution:

$$p_N(s \mid x) \propto s^{-3/2} \Theta(N-s-x) \tag{3.11}$$

Consequently, only the spatial correlations between subsequent protein intersegment transfers remain as possible cause for diffusive transport. If these correlations are destroyed, sparing only the temporal correlations that exist between these intersegment transfers and necessarily come about due to the frozenness of the intersegment link pattern, transport is superdiffusive again – in spite of the temporal correlations. Therefore, the reason for the fact that an intersegment link pattern representing an infinitely fast moving chain conformation produces superdiffusive protein transport is that the intersegment transfers loose their spatial correlations due to the fast pattern changes. The loss of temporal correlations is only a side-effect and cannot be seen as cause for the fact that transport is no longer diffusive but superdiffusive in this limit. Hence, in order to characterize the correlations further, it is sufficient to concentrate on static intersegment link patterns and their geometric and topological properties.

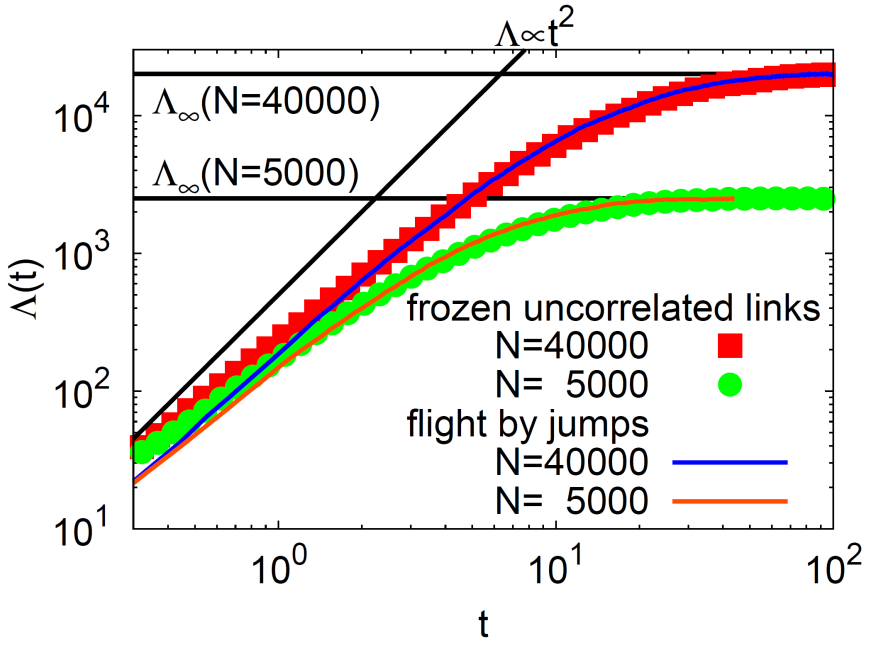


*Fig. 3.3: The width of the protein density along the chain contour grows superdiffusively (before saturating at N/2 due to finite size effects) not only for a Lévy flight, which is equivalent to transport on an **infinitely fast changing** intersegment link pattern and where subsequent protein jumps are entirely uncorrelated (blue and orange lines), but also for transport on **frozen** intersegment link patterns (red and green points), as long as the links are geometrically uncorrelated (here the link lengths are independent, identically and broadly distributed random variables and the link positions are randomized.): "Frozenness" is not sufficient for diffusive transport behavior.*

3.3.3 Geometric properties of static intersegment link patterns

One can distinguish the quality of the geometrical correlations a given artificial intersegment link pattern contains by the type of probability distribution (more exactly: by the number of links not distributed independently there) used in order to generate the pattern by drawing the lengths and positions of the links. Of course, the corresponding probability distribution has to be measured from real random walk chain conformations first. The hope is to find a relatively simple distribution which involves as less correlated intersegment links as possible but brings about all necessary correlations to obtain diffusive transport in a frozen pattern, even if the length distribution for single intersegment links is broad.

For example, one could draw frozen patterns from the – complex but hypothetically measurable – joint probability distribution that contains correlated geometrical information about *all* original intersegment links within a random walk chain

$$p_N(s_1, s_2, \Delta_{12}, s_3, \Delta_{23}, \ldots, s_M, \Delta_{M-1,M}) = ? \,, \tag{3.12}$$

where the s_i denote the lengths of the intersegment links and the $\Delta_{i-1,i}$ the distances between them. Such patterns are expected to show diffusive transport behavior, since literally all geometrical correlations present in the original random walk chain conformations are preserved (maximal correlation). However, this procedure is equivalent to taking the intersegment link pattern simply from the random walk chain itself.

By contrast, only superdiffusive transport is obtained if the frozen pattern is drawn from the length distribution of *single* intersegment links, which is the simplest type of probability distribution one can use and which does not preserve any geometrical correlations between the links (zero correlation):

$$p_N(s_1) \propto {s_1}^{-3/2} \frac{N - s_1}{N} \tag{3.13}$$

One could ask whether the geometrical correlations generated by drawing the frozen pattern from the length (and position) probability distribution of *pairs* of intersegment links are sufficient to make transport diffusive, since the calculation of this pair distribution is theoretically tractable, as will be shown in the following:

3.3.3.1 Generic calculation procedure for arbitrarily complex intersegment link probability distributions

The probability in one dimension to find a random walker (constant step length Δx and time update 1) at time t at position x, if it had been initially in the origin, is given by the Gaussian propagator:

$$G(x|t) = \frac{1}{\sqrt{2\pi(\Delta x)^2 t}} \exp\left(-\frac{x^2}{2(\Delta x)^2 t}\right) \tag{3.14}$$

This result can be easily extended to describe three-dimensional random walks. Since the trajectory of such a walker can be seen as a three-dimensional random walk chain conformation, there is a close correspondence between the time appearing in the Gaussian propagator and the contour coordinate l of the chain. Therefore, the probability that within such a chain there exists a sub-chain with contour length l which connects two points in space with Euclidean distance vector $\boldsymbol{r}$ is given by:

$$g(\boldsymbol{r}|l) \propto l^{-3/2} \exp\left(-\frac{\boldsymbol{r}^2}{2l}\right) \tag{3.15}$$

A given intersegment link geometry can be represented by a product of such probabilities, incorporating the contour lengths and Euclidean distance vectors occurring in that geometry. Then integrating out all remaining spatial degrees of freedom yields a joint probability distribution for all relevant contour lengths the geometry has.

3.3.3.2 Calculating the probability distribution for intersegment link pairs

Although this procedure is capable of determining probability distributions for arbitrarily complex intersegment link geometries, it will be used here for *pairs* of intersegment links, since their probability distribution can easily be measured from real random walk chain conformations and be compared to the theoretical findings. The contour space geometry of such link pairs is best described by three distances (contour length s_1 of the first link, contour length s_2 of the second link and contour distance Δ_{12} between the starting points of the two links), which will be called contour variables. Depending on these distances, all link pairs can be divided into three separate topological classes, namely nested pairs, knotted pairs and independent pairs (cf. Fig. 3.4), which is why the domain of the pair probability distribution separates into three subdomains as well.

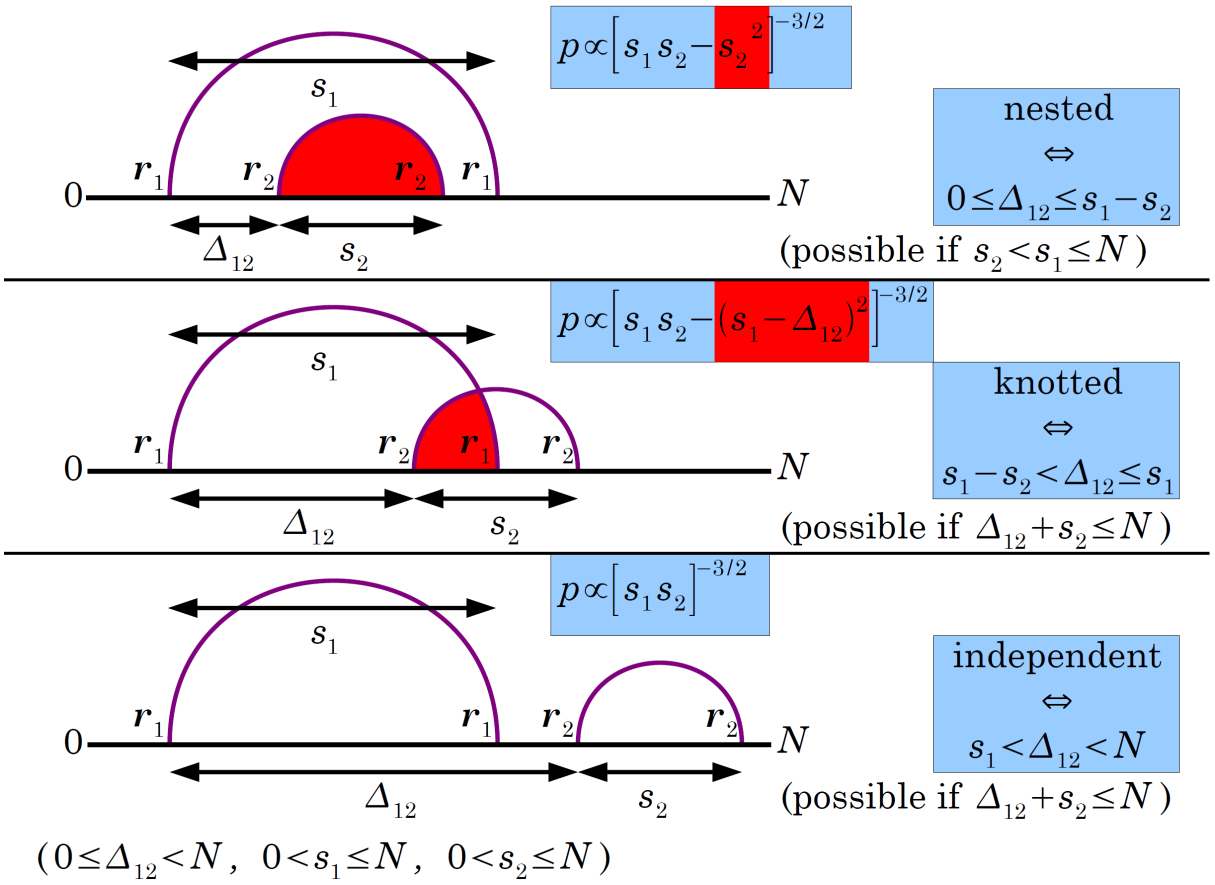


Fig. 3.4: The domain of the intersegment link pair probability distribution $p(s_1, s_2, \Delta_{12})$ is divided in three subdomains, where the two links a intersegment link pair consists of are either nested, knotted or independent in chain contour space. The distribution depends on its arguments (link lengths s_1 and s_2, contour distance between the links Δ_{12}) in a different way for the different subdomains: Only for independent links the distribution factorizes. For nested or knotted links the overlapping region of the chain contour results in a correction term depending on the contour length of the overlap (marked in red).

The **subdomain of nested links** contains only pairs of links which fit into one another:

$$s_2 < s_1 \text{ and } 0 \le \Delta_{12} \le s_1 - s_2 \tag{3.16}$$

In this subdomain the pair probability distribution is given by

$$p(s_1, s_2, \Delta_{12}) \propto \int \mathrm{d}^3(\boldsymbol{r}_2 - \boldsymbol{r}_1)\, g(\boldsymbol{r}_2 - \boldsymbol{r}_1 | \Delta_{12})\, g(\boldsymbol{r}_2 - \boldsymbol{r}_2 | s_2)\, g(\boldsymbol{r}_1 - \boldsymbol{r}_2 | s_1 - \Delta_{12} - s_2) \propto \left[s_2 (s_1 - s_2) \right]^{-3/2} \tag{3.17}$$

for infinite chains and by

$$p_N(s_1, s_2, \Delta_{12}) = p(s_1, s_2, \Delta_{12}) \frac{N - s_1}{N} \tag{3.18}$$

for finite chains of length N. The additional finite size factor results from the fact that for finite chains not all link pair positions are possible anymore because of the size of the link pair (here: s_1).

The **subdomain of knotted links** contains only pairs of links which intersect each other (as if they represented base-pairs forming a pseudoknot):

$$s_1 - s_2 < \Delta_{12} \leq s_1 \tag{3.19}$$

The corresponding pair probability distribution reads

$$p(s_1, s_2, \Delta_{12}) \propto \int \mathrm{d}^3(\boldsymbol{r}_2 - \boldsymbol{r}_1)\, g(\boldsymbol{r}_2 - \boldsymbol{r}_1 | \Delta_{12})\, g(\boldsymbol{r}_1 - \boldsymbol{r}_2 | s_1 - \Delta_{12})\, g(\boldsymbol{r}_2 - \boldsymbol{r}_1 | \Delta_{12} + s_2 - s_1) \propto \left[s_1 s_2 - \left(s_1 - \Delta_{12} \right)^2 \right]^{-3/2} \tag{3.20}$$

for infinite chains and

$$p_N(s_1, s_2, \Delta_{12}) = p(s_1, s_2, \Delta_{12}) \frac{N - \Delta_{12} - s_2}{N} \tag{3.21}$$

for finite chains. Again the finite size factor depends on the pair size (here: $\Delta_{12} + s_2$) in the same way.

All other link pairs can be found in the **subdomain of independent links**, where the links are separated from each other in contour space:

$$s_1 < \Delta_{12} < N \tag{3.22}$$

According to this separation, the pair probability distribution factorizes for infinite chains:

$$p(s_1, s_2, \Delta_{12}) \propto \int \mathrm{d}^3(\boldsymbol{r}_2 - \boldsymbol{r}_1)\, g(\boldsymbol{r}_1 - \boldsymbol{r}_1 | s_1)\, g(\boldsymbol{r}_2 - \boldsymbol{r}_1 | \Delta_{12} - s_1)\, g(\boldsymbol{r}_2 - \boldsymbol{r}_2 | s_2) \propto \left[s_1 s_2 \right]^{-3/2} \tag{3.23}$$

Because the dependence of pair size on the contour variables of the link pair is the same in both the knotted and the independent subdomains, one obtains the same finite size factor:

$$p_N(s_1, s_2, \Delta_{12}) = p(s_1, s_2, \Delta_{12}) \frac{N - \Delta_{12} - s_2}{N} \tag{3.24}$$

Comparing the three function parts the pair probability distribution is composed of for infinite chains, one easily recognizes that they are all power-laws with the same exponent (-3/2), whose arguments are polynomials of second degree in the three contour variables of the link pair (Generally, a link geometry consisting of M links yields polynomials of M-th degree in $2M$-1 contour variables.). Depending on the extent of link intersection within a pair the polynomials in the nested and in the knotted subdomain contain subtractive corrections to the independent case where the polynomial factorizes (cf. intersections and term marked in red in Fig. 3.4).

3.3.3.3 Excursus: Probability distribution for intersegment link M-tupels

Whereas the domain of the probability distribution of link pairs is divided into only three subdomains, the number of subdomains needed for the probability distribution of link M-tupels is so large that this probability distribution shall only be calculated in two very special subdomains:

In the **subdomain of multiply nested links arrays**, which contains all link tupels obeying the conditions

$$s_M < s_{M-1} < \ldots < s_3 < s_2 < s_1 \tag{3.25}$$

and

$$0 \le \Delta_{12} \le s_1 - s_2 \ \wedge \ 0 \le \Delta_{23} \le s_2 - s_3 \ \wedge \ \ldots \ \wedge \ 0 \le \Delta_{M-1,M} \le s_{M-1} - s_M \,, \tag{3.26}$$

the probability distribution on infinite chains can be calculated recursively:

$$p(s_1, s_2, \Delta_{12}, s_3, \Delta_{23}, \ldots, s_M, \Delta_{M-1,M}) \propto \left[s_M (s_{M-1} - s_M) \cdots (s_2 - s_3)(s_1 - s_2) \right]^{-3/2} \tag{3.27}$$

The finite size factor needed to adapt this result to finite chains depends only on the size s_1 of the outermost link of the nested M-tupel:

$$p_N = \frac{N - s_1}{N} \cdot p \tag{3.28}$$

The **subdomain of entirely independent link arrays** contains all link tupels which obey the condition

$$s_1 < \Delta_{12} < N \ \wedge \ s_2 < \Delta_{23} < N \ \wedge \ \ldots \ \wedge \ s_{M-1} < \Delta_{M-1,M} < N \tag{3.29}$$

and shows a completely factorizing probability distribution on infinite chains:

$$p(s_1, s_2, \Delta_{12}, s_3, \Delta_{23}, \ldots, s_M, \Delta_{M-1,M}) \propto \left[s_1 s_2 s_3 \cdots s_{M-1} s_M \right]^{-3/2} \tag{3.30}$$

In order to obtain the probability distribution valid on finite chains, one has to multiply by a finite size factor which again depends on the size of the M-tupel. Here this size is dominated by the sum of all inter-link distances:

$$p_N = \frac{N - (\Delta_{12} + \Delta_{23} + \ldots + \Delta_{M-1,M}) - s_M}{N} \cdot p \tag{3.31}$$

3.3.3.4 Measuring the probability distribution for intersegment link pairs

Comparing the intersegment link pair probability distribution deduced theoretically to histogram measurements performed on frozen intersegment linking patterns obtained from three-dimensional random walk chain conformations one finds perfect congruence. This is demonstrated in Fig. 3.5 for several inter-link distances (i.e. distance between primary and secondary link in an intersegment link pair) by varying the length of the primary link in a pair while keeping the length of the secondary link constant. Clearly in evidence is the tripartite structure of each histogram curve which corresponds directly to the three link pair domains of independent, knotted and nested links.

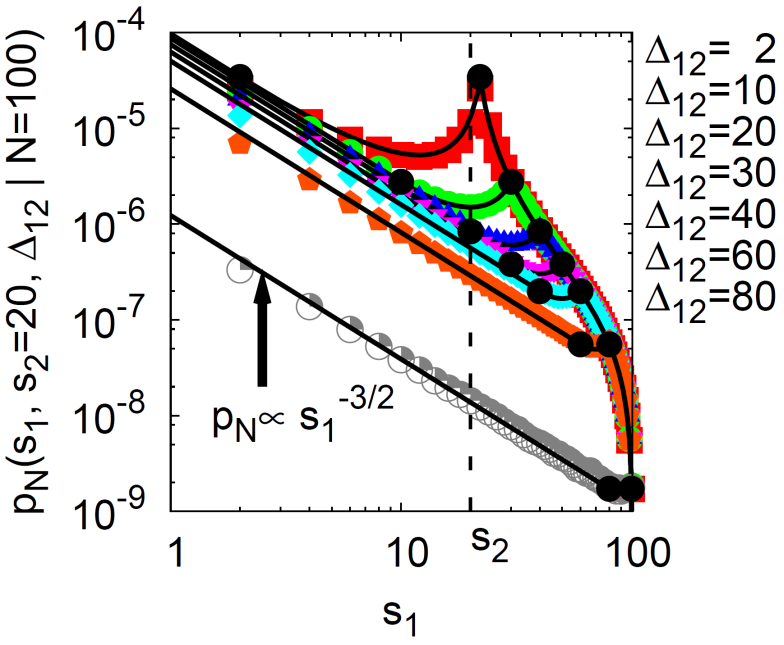


Fig. 3.5: The intersegment link pair probability distribution deduced theoretically (solid black lines) is confirmed exactly by histogram measurements on three-dimensional frozen random walk chain conformations (chain length N=100). One link length is kept fix (s_2=20), the contour distance Δ_{12} between the links parametrizes the array of curves. Each curve consists of three subdomains (1. independent, 2. knotted and 3. nested links). The boundaries between these subdomains are marked by filled black circles. For inter-link distances approaching zero the probability at the knotted/nested boundary diverges since for vanishing distance knotted/nested pairs would consist of "identical" links (cf. vertical asymptote at s_1=s_2), whereas the subdomain of independent links ($p_N \propto s_1^{-3/2}$) disappears entirely.

3.3.3.5 Transport via pairwise correlated intersegment links

Having thus confirmed the correctness of the theoretical form of the intersegment link pair probability distribution, this form can be used in order to study protein transport on frozen intersegment linking patterns lacking all types of geometrical correlations except for those conveyed by this pair distribution. The question is whether these pair correlations are sufficient on their own to make transport diffusive instead of superdiffusive or whether correlations of even higher order are necessary in fact. To answer this question the intersegment linking pattern of a random walk chain, which ab initio contains geometrical correlations of every order, is changed by shifting the positions (not the lengths!) of the intersegment links in contour space by the means of an Metropolis Monte-Carlo algorithm. This algorithm is designed to preserve the pair correlations present in the original intersegment linking pattern while destroying all geometrical correlations of higher order. To achieve this, its probability of move acceptance,

$$P_{\text{shift}} = \begin{cases} \exp[-(c_{t+1}-c_t)] & c_{t+1}-c_t>0 \\ 1 & \text{else} \end{cases}, \tag{3.32}$$

has to be essentially an exponential of the change of the following time-dependent (positive) cost function:

$$c_t = \frac{1}{kT}\int \mathrm{d}s_1 \int \mathrm{d}s_2 \int \mathrm{d}\Delta_{12} \left| n_t(s_1,s_2,\Delta_{12}) - p_N(s_1,s_2,\Delta_{12}) \right| \tag{3.33}$$

This cost function measures the deviation of the link pair density n_t present in the intersegment linking pattern at time t from the theoretical intersegment link pair probability distribution, which can be regarded as an averaged initial link pair density (before shifting links!):

$$\langle n_0(s_1, s_2, \Delta_{12})\rangle_{\text{patterns}} \to p_N(s_1, s_2, \Delta_{12}) \tag{3.34}$$

After a number of such preparing Monte-Carlo annealing steps the intersegment linking pattern is kept frozen and a protein is started on it. Fig. 3.6 shows the width of the protein density along the chain contour as a function of time as obtained by the use of the following simulation parameters:

- 100 distinct intersegment linking patterns
- each taken from a closed random walk chain of length 10000
- variable number of preparation steps per *initial* intersegment linking pattern $(0, 10^2, 10^3, \ldots, 10^7)$, each step performed at temperature $kT = 200$
- 10000 protein runs per *prepared* intersegment linking pattern

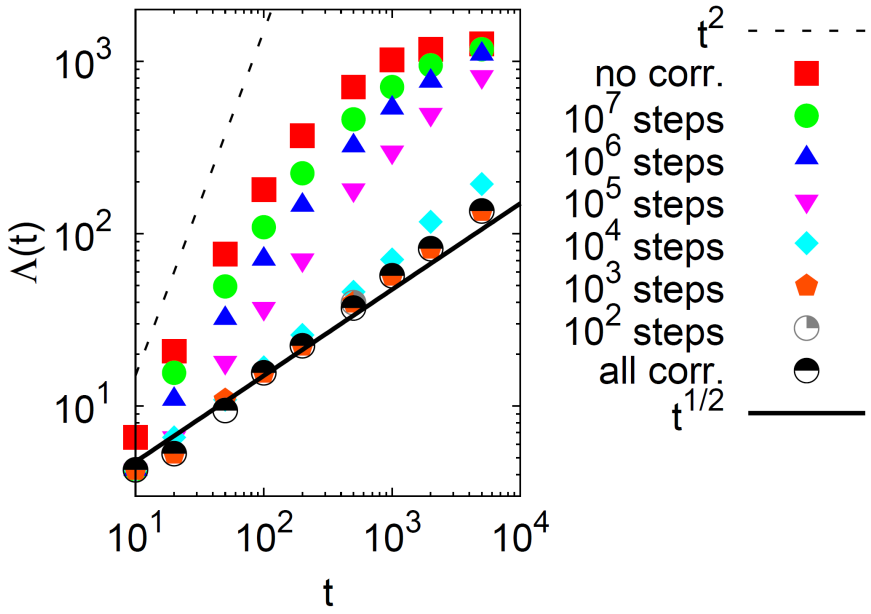


Fig. 3.6: Time evolution of the width of the protein density in contour space, characterizing protein transport performed on frozen intersegment linking patterns (from closed random walk chains of contour length 10000) after gradual elimination of all geometrical correlations of higher order than link pair correlations. This elimination is done by Monte-Carlo steps that shift the intersegment link positions while preserving the initial link pair density on average. Without such steps, transport is diffusive (black circles). The more steps are performed the more superdiffusive transport becomes (gray, orange, ... , green points), until it behaves similar to transport on patterns with entirely randomized intersegment link positions where even the link pair correlations have been eliminated (red squares). Therefore, pair correlations in intersegment linking patterns are not sufficient to make transport diffusive.

Without annealing (0 preparation steps), the width of the protein density grows diffusively, as expected: no geometrical correlations have been destroyed. The longer the annealing process endures (here: up to 10^7 preparation steps), the more thoroughly geometrical correlations between intersegment links of higher order than pair correlations are destroyed and – surprisingly – the more superdiffusive protein transport becomes. Transport on intersegment linking patterns that have been annealed very long (10^7 pre-

paration steps) even exhibits the same superdiffusive value for the transport exponent as transport on patterns with totally randomized intersegment link positions. In other words: Although all pair correlations have been preserved by the annealing process, transport behaves as in the entirely uncorrelated case. Therefore, one has to conclude that not pair correlations, but correlations of higher order than pair correlations cause protein transport on frozen chains to be diffusive. It is not clear whether there exists a minimal order of correlations one has to preserve in an intersegment linking pattern in order to obtain diffusive transport. This makes the problem of characterizing the nature of the geometrical correlations present in random walk chains a difficult one. However, a different approach helps to solve this problem, as will be shown in the following.

3.4 Islands

For a successful characterization of the geometrical correlations that make intersegmental protein transport on stationary DNA conformations diffusive, it is essential to look at the intersegment linking pattern of a frozen random walk chain from a more global point of view. The question is, whether there are any effects induced by the geometrical correlations that are visible on such **native** patterns. And indeed, native patterns exhibit a very striking feature: the divisibility of the random walk chain into an array of adjacent sub-chains each containing an entirely separate set of intersegment links. Since these sub-chains are not connected to each other by intersegment links, they are isolated units in terms of intersegment transfer. That is why these sub-chains are called islands: Whereas the protein can make heavy use of intersegment transfer *within* an island, it can change *between* adjacent islands only via sliding along the chain contour. Fig. 3.7 (above) shows exemplary intersegment linking patterns of frozen random walk chains of length 1000. Clearly in evidence is the high number of islands these native patterns can be divided into.

However, this divisibility is not necessarily the case for an intersegment linking pattern with broad link length distribution, but truly emerges from the strong geometrical correlations present in a frozen random walk chain. This fact can be shown by looking at **artificial** linking patterns, which have been generated by randomly re-positioning the intersegment links of native patterns while preserving the link lengths: As one can see from Fig. 3.7 (below), the resulting artificial patterns cannot be subdivided into smaller islands at all.

Such an *artificial* intersegment linking pattern lacks all types of geometrical correlations. This causes superdiffusive transport even if the pattern is stationary (cf. chapter 3.3.2). On the other hand, stationary *native* patterns have strong geometrical correlations and thus exhibit only diffusive transport. Since *both* diffusive transport *and* the existence of islands seem to be direct effects of these correlations,

$$\text{strong correlations} \Rightarrow \text{islands} \wedge \text{diffusive transport}\,, \tag{3.35}$$

the idea of a causal connection between islands and diffusive transport suggests itself:

strong correlations⇒ islands⇒ diffusive transport (3.36)

It will be shown that this causal connection exists indeed, and that it is possible to coarse-grain the intersegment linking pattern of a frozen random walk chain, regarding it as a chain of independent black-box elements representing the original islands. Due to the process of coarse-graining, these elements ought to be independent from the details of the link geometry in the islands they represent but be described only by a very restricted set of relevant statistical properties. The goal is to demonstrate that transport on a chain consisting of such elements is diffusive as long as they have the same statistical properties as random walk islands. From this proof it would be clear that the relevant characteristic of the strong geometrical correlations present in frozen random walk chains, as far as protein transport behavior is concerned, is the ability to form intersegment link islands with the correct intrinsic statistical properties.

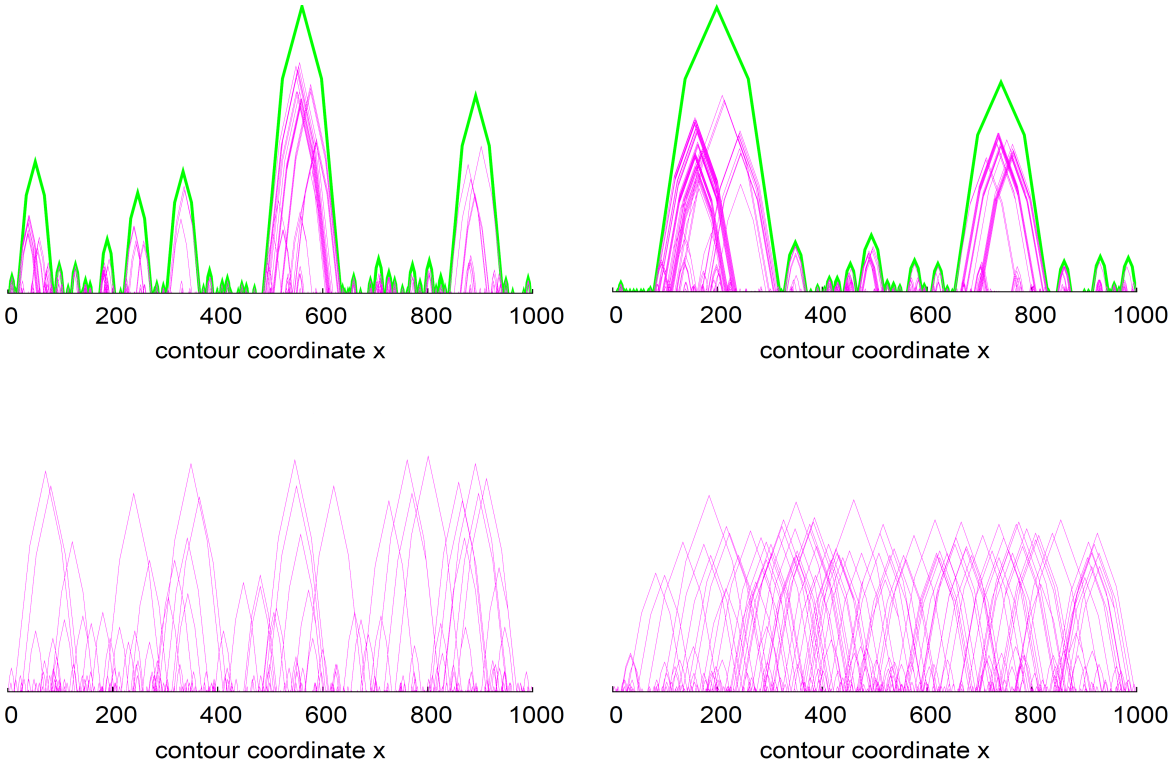


Fig. 3.7: Above: Native intersegment linking patterns originating directly from frozen random walk chains (here: length 1000) can be divided into numerous separate islands (green) of intersegment links (magenta) due to strong geometrical correlations between the links. Below: If these links are set at random positions in contour space in order to destroy all geometrical correlations, the resulting artificial intersegment linking pattern cannot be divided further: there are no islands at all.

But not only transport on frozen chains could be explained with the help of islands then, they would also be useful when describing transport on dynamic chains: Up to now the cause for the fact that protein transport on dynamic DNA conformations is superdiffusive for short times has been attributed to the dynamic destruction of geometrical correlations. Now, this destruction of correlations can be depicted more precisely as a chan-

ging island landscape, where islands continuously appear, move their boundaries, change their size and disappear again. This is shown in Fig. 3.8, where the time evolution of the color-coded intersegment link density of a dynamic chain of length 100 is plotted, by that visualizing the changing island boundaries.

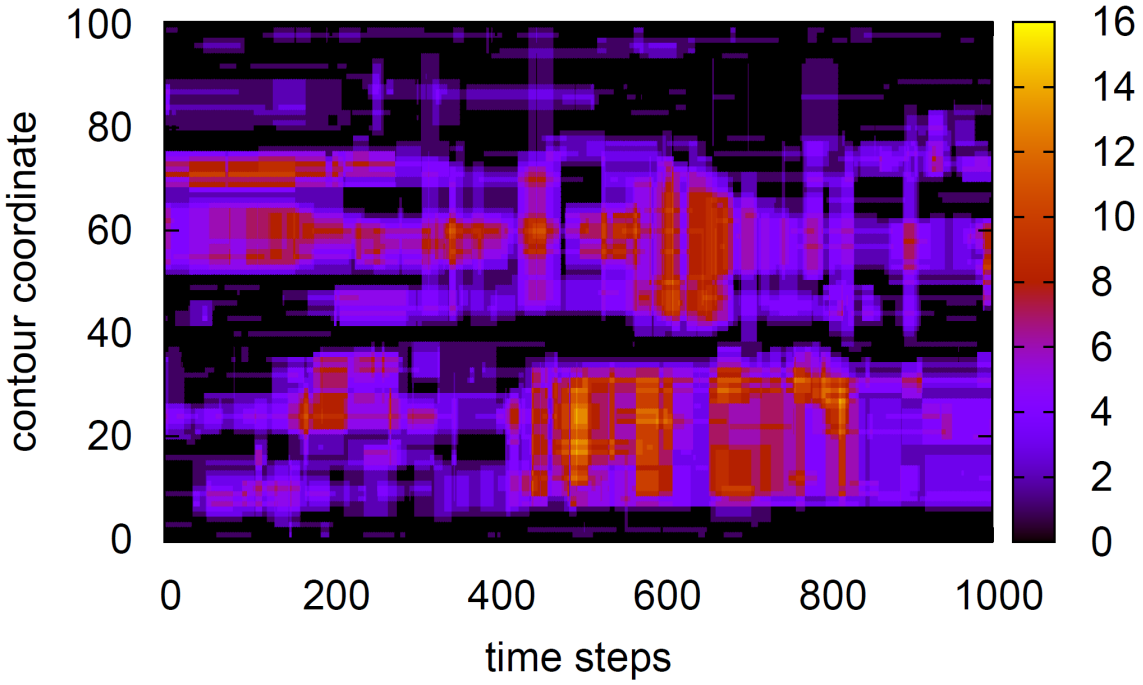


Fig. 3.8: Time evolution of the intersegment link density in contour space for a single dynamic DNA chain of length 100. The colored chain contour regions contain intersegment links and are separated from each other by black linkless regions. The boundaries between colored and black regions are island boundaries. Shape and position of the (colored) islands change over time, because conformational changes of the DNA chain cause changes in its intersegment linking pattern.

3.4.1 Definition

Before the statistical properties that characterize islands and are necessary in order to coarse-grain intersegment linking patterns describing random walk chains can be selected and measured, a mathematically stringent definition of such intersegment link islands and their boundaries, which does not depend on representation, is needed:

Choose a sub-chain M of a random walk chain arbitrarily, by doing so separating the whole chain into a chain region A in front of M, the sub-chain M itself and a chain region B behind M. Then the sub-chain M is called an unconnected island of intersegment links (or short hand "island") if and only if (conditions C_1 - C_4):

1. the boundary points of the sub-chain M are connected to each other or to other chain points belonging to M by intersegment links,

2. there is no intersegment link which connects a chain point belonging to the sub-chain M with a chain point outside the sub-chain M,
3. there is no intersegment link which connects a chain point belonging to region A with a chain point belonging to region B,
4. the sub-chain M does not contain smaller sub-chains obeying these connection conditions (C_1 - C_3) simultaneously.

Fig. 3.9 visualizes this definition in Euclidean representation as well as in contour space representation, especially by showing typical counter-examples.

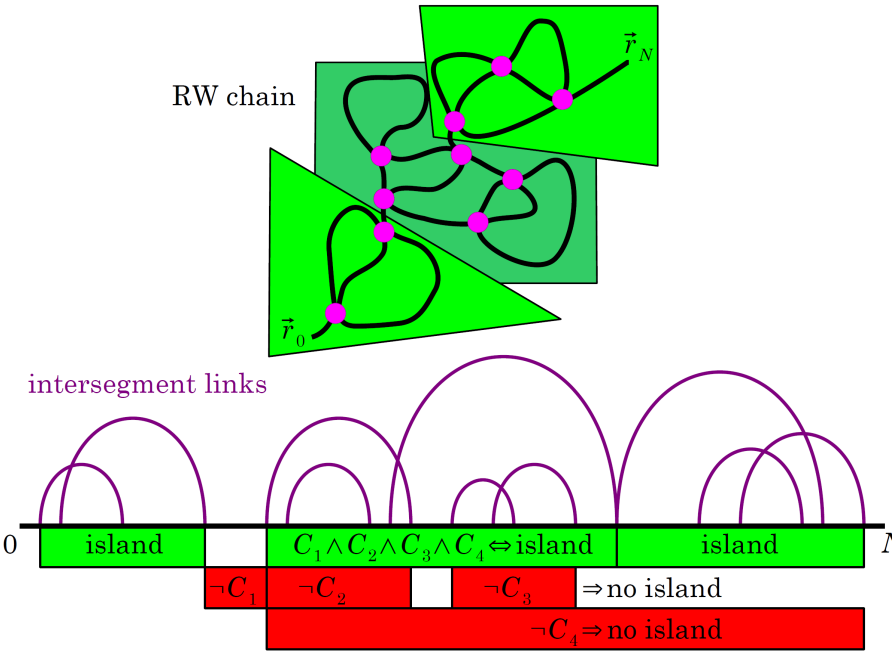


Fig. 3.9: A three-dimensional random walk chain can be divided in sub-chains that are independent in terms of their intersegment link pattern (magenta) and called islands (green). These islands are the shortest sub-chains possible (C_4) for which the following statements are valid: An island is never empty of intersegment links, its boundaries are always link boundaries (C_1). The interior of an island is never connected to its exterior (i.e. the rest of the chain) by intersegment links (C_2). An island is never bridged by an intersegment link connecting chain regions that are exterior (C_3).

3.4.2 Statistical properties of RW islands

The most relevant properties of a *single* island as far as protein transport is concerned are its *length* in contour space and its conformation-*specific mean exit time*. However, since this time strongly depends on the intersegment linking pattern of the island, it is related to the island length itself only weakly and can vary even between equally long islands. Therefore, from a statistical point of view it is useful to look at the *distribution of island lengths* and accordingly at the length-specific but *disorder-averaged mean exit time* which involves averaging the conformation-specific mean exit times over many island conformations of equal contour length.

The **distribution of island lengths** is a power-law, which – not surprisingly – inherits its exponent directly from the intersegment link length distribution:

$$\lambda(s) \propto s^{-3/2} \tag{3.37}$$

The reason for this fact has to be looked for in the self-similarity of intersegment linking patterns: Since the lengths of single links as well as the distances between them are distributed algebraically with exponent -3/2 (cf. for example the joint length distributions for complexes of many intersegment links in chapter 3.3.3.3), the lengths of the islands composed of these links are distributed equally, namely just like the lengths of their components.

Fig. 3.10 shows the power-law nature of the distribution of island lengths: In order to filter out discretization effects and finite size effects, the island length distribution has been measured for several lengths of the random walk chain (from $L=1000$ to $L=100000$) . Each curve exhibits a region of algebraic decay. All curves are congruent within this generic power-law region. The corresponding exponent is -3/2. In order to obtain the island length distribution for a distinct length L of the random walk chain several random walk conformations are generated. However, one randomly selects only one island per conformation. The reason is that islands in the boundary region of a finite random walk chain may behave different from inner islands, which could influence the shape of the island length distribution especially if all islands within each conformation were used.

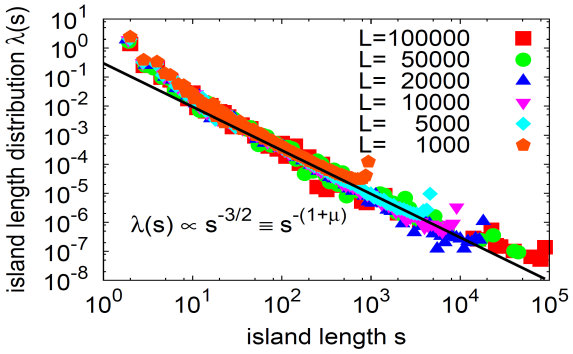


Fig. 3.10: Contour length distribution for intersegment link islands found in random walk conformations (selecting one island per conformation) of several lengths. All curves share a common power-law region. The exponent results from the intersegment link length distribution and corresponds to an island length parameter of $\mu=1/2$.

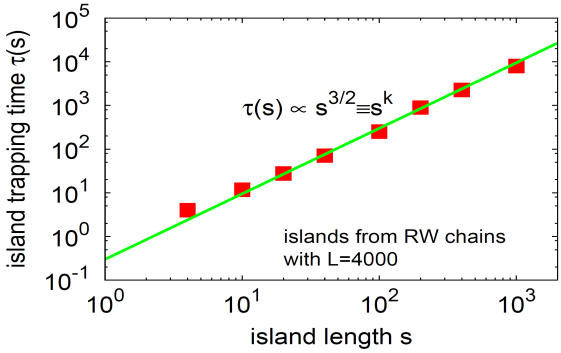


Fig. 3.11: Disorder-averaged (but length-specific) island trapping time. This mean time a protein initially sitting in the center of an island in contour space needs to escape from the island scales superdiffusively with the contour length of the island. The islands have been obtained from random walk conformations of length 4000 by selecting one island per conformation. The scaling exponent results in an island trapping parameter of $k=3/2$.

However, the distribution of island lengths can not explain the paradoxical properties of protein transport on a frozen random walk chain conformation on its own. The complementary statistical determinant of transport is the scaling of the disorder-averaged mean island exit time with the island length, the so-called **island trapping time**. As Fig. 3.11 shows, the island trapping time scales superdiffusively with island length:

$$\tau(s) \propto s^{3/2} \tag{3.38}$$

To find this result the island trapping time has been measured for several island lengths. The island trapping time for one *specific island length* has been obtained by averaging over the mean exit times of several island conformations of this length. This disorder average is necessary because the mean exit times strongly depend on the particular island conformation for which they are determined, even for equal island lengths. The mean exit time is defined here as the time needed on average by a protein walker initially sitting in the center of the contour length interval of a *specific island conformation* to leave this island conformation at one of its boundaries (central mean exit time). As for the island length distribution, all islands needed for this procedure are chosen from distinct random walk conformations (length $L=4000$).

The exponent of the island length distribution and the exponent connecting island length and island trapping time yield together only global information about islands. It is important to clarify the question whether this information is sufficient to explain how the geometric correlations expressing themselves in the formation of islands can make protein transport on frozen random walk chains diffusive. To this end, it is necessary to examine the details of protein transport *within* islands. This will be done by performing first passage processes on intersegment linking patterns of islands.

Generally, the process of escape from a finite size interval of length s with absorbing boundaries at $x=0$ and $x=s$ is characterized by the unconditional mean exit time (without specified escape direction), which in turn depends on the conditional mean exit times for escapes either to the left $(-)$ or to the right $(+)$, weighted by the respective probabilities of escape direction, the so-called splitting probabilities [133]:

$$t_{\text{exit}}(u_0)=P^{-}_{\text{exit}}(u_0)t^{-}_{\text{exit}}(u_0)+P^{+}_{\text{exit}}(u_0)t^{+}_{\text{exit}}(u_0) \tag{3.39}$$

All these are functions of the initial position of the walker within the interval, rescaled by the interval length:

$$u_0=\frac{x_0}{s} \tag{3.40}$$

Both transport processes on directionally unbiased systems (like line intervals for pure one-dimensional diffusion) and transport processes on directionally biased systems (like islands with specific intersegment linking patterns) – if this bias has been eliminated by a disorder average – show a number of **universal symmetries**: Initial positions that are symmetric about the center of the transport interval result in

- equal unconditional mean exit times:

$$t_{\text{exit}}(1-u_0)=t_{\text{exit}}(u_0) \tag{3.41}$$

- equal deviations of the splitting probabilities from equiprobability (0.5):

$$P^{+/-}_{\text{exit}}(1-u_0)+P^{+/-}_{\text{exit}}(u_0)=1 \tag{3.42}$$

- equal conditional mean exit times for opposite exit directions:

$$t^{+/-}_{\text{exit}}(1-u_0)=t^{-/+}_{\text{exit}}(u_0) \tag{3.43}$$

As a consequence, the unconditional and both conditional mean exit times are equal for central initial positions and can be identified with a time τ only depending on interval length, the so-called central mean exit time:

$$t_{\text{exit}}^{+/-}(1/2)=t_{\text{exit}}(1/2)\equiv\tau \tag{3.44}$$

As already known, the **scaling** of this central mean exit time is superdiffusive for island intervals

$$K_{\text{protein}}\tau\propto s^{3/2} \tag{3.45}$$

and diffusive for intervals without intersegment links

$$2K_{\text{protein}}\tau=\frac{1}{4}s^2\,. \tag{3.46}$$

However, this difference in scaling is not the only effect the intersegment linking pattern of an island has on escape behavior. The examination of how unconditional and conditional mean exit times as well as the splitting probabilities **depend on the initial position** will reveal even more striking differences to purely diffusive escape, although both escape from islands and diffusive escape share the same symmetry properties discussed above.

To this end, escape processes are performed on intervals (length $s=50$) with absorbing boundaries, starting 10000 runs for each interval instance and for each of the initial positions. The left part of Fig. 3.12 compares the splitting probabilities for diffusive escape, for escape from one specific island conformation and for an escape process averaged over 1000 different island conformations of equal length (disorder-average). In the right part of the figure the same is done with the unconditional mean exit time. In Fig. 3.13 escape from an island interval (with disorder-average) is compared to diffusive escape in terms of the conditional mean exit times.

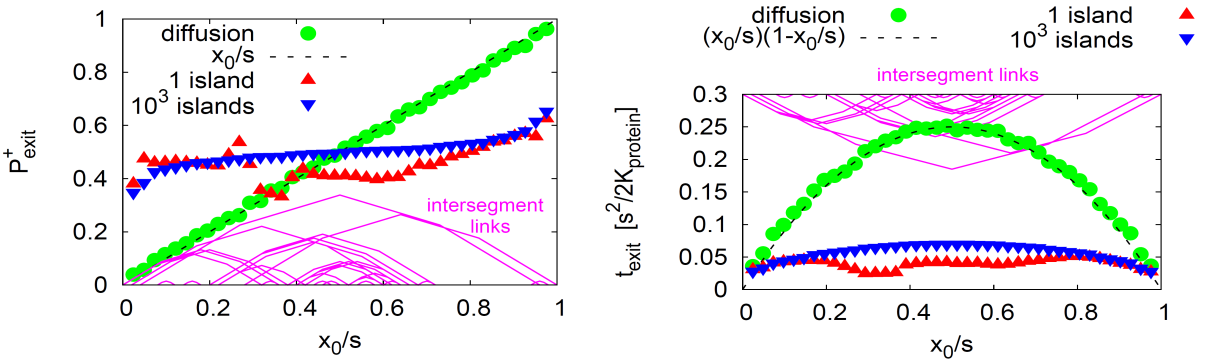


Fig. 3.12: Process of escaping from an interval (length s=50) with and without intersegment linking pattern. Both the probability to leave to the right (left figure) and the unconditional mean exit time (right figure) depend much less from the initial position for disorder-averaged (1000 islands) transport on island intervals (blue) than for purely diffusive transport on intervals without intersegment links (green). Without such a disorder average the island escape probability as well as the island mean exit time (red) are related strongly to the specific shape of the intersegment linking pattern (magenta). 10000 escape processes were averaged per interval instance and initial position. Altogether, different from diffusive transport, intersegment links delocalize the walker over the entire island interval before it can escape.

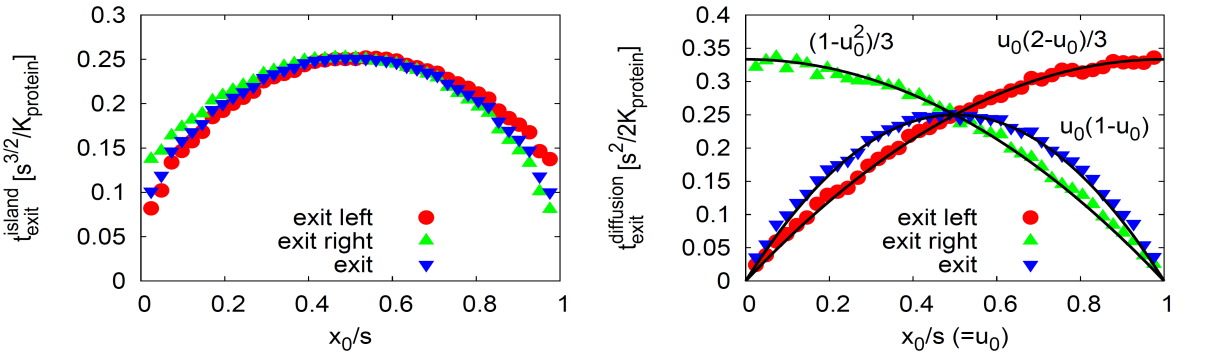


Fig. 3.13: Comparison between escape from island intervals (disorder-averaged over 1000 island instances, left figure) and purely diffusive escape from intervals without intersegment links (right figure) by looking at the unconditional (blue) and the conditional (red, green) mean exit times as functions of the initial position of the walker. Each curve has been obtained by averaging over 10000 escape events per interval instance (length s=50) and initial position. Although both types of escape processes obviously share the same universal symmetries (for details cf. text), they differ appreciably in terms of walker localization: Since diffusive escape is a next-neighbor transport process, the conditional mean exit times are monotonically decreasing functions of the distance to the boundary where the walker is actually absorbed. By contrast, the intersegment links of an island delocalize the walker over the entire interval. Therefore, the conditional mean exit times are not only less dependent on the distance to an absorbing boundary. Moreover, a boundary is reached comparably fast by walkers started at far and near initial positions, as long as they are symmetric about the interval center.

Escape from an interval without intersegment links by pure diffusion is a next-neighbor process. Therefore, the protein is quite localized during the entire escape process, which makes the exit direction depend strongly on the initial position. This is reflected in the fact that the splitting probabilities are linear functions of the initial position:

$$P^{+}_{\text{exit}}(u_0)=u_0 \quad P^{-}_{\text{exit}}(u_0)=1-u_0 \tag{3.47}$$

Another effect of this localization is that the protein reaches the boundary where it is absorbed the faster the smaller the distance between this boundary and its initial position has been. That is why the conditional mean exit times are monotonic functions of the initial position,

$$2K_{\text{protein}}\, t^{+/-}_{\text{exit}}(u_0)=\begin{cases}\frac{1}{3}s^2(2-u_0)u_0\\ \frac{1}{3}s^2(1-u_0^2)\end{cases}, \tag{3.48}$$

unlike the unconditional mean exit time, which is symmetric:

$$2K_{\text{protein}}\, t_{\text{exit}}(u_0)=s^2u_0(1-u_0) \tag{3.49}$$

By contrast, **escape from islands** is by no means a next-neighbor process. On the contrary, the intersegment links an island consists of enhance protein transport by allowing for long-range jumps and thus delocalize the protein position over the entire island interval very quickly, by that causing the memory for the initial position to become lost. This is even the case without disorder average, although then splitting probabilities as well as mean exit times are strongly related to the specific intersegment linking pattern in use. Having performed a disorder average, the delocalization becomes even more evident: First, the exit direction depends then only very weakly on the initial position, which results in splitting probabilities almost constant for initial positions in the interior of the island interval:

$$P^{+/-}_{\text{exit}}(u_0)\approx\frac{1}{2} \quad \text{for} \quad \left|u_0-\frac{1}{2}\right|\ll\frac{1}{2} \tag{3.50}$$

Moreover, unconditional as well as conditional mean exit times depend much weaker on the initial position than in the diffusive case. The most striking feature is, however, that the conditional mean exit times are no longer monotonic functions of the initial position but resemble the unconditional mean exit time, which is symmetric:

$$t^{+/-}_{\text{exit}}(u_0)\approx t_{\text{exit}}(u_0) \tag{3.51}$$

That means: If a protein that gets absorbed at a distinct boundary (e.g. the right-hand boundary) escapes fast, its initial position has not been next to the same boundary necessarily. An initial position next to the opposite boundary would have the same effect, despite the large contour distance to the target boundary. The reason is that due to the presence of intersegment links initial positions which are symmetric about the interval center are almost equivalent with respect to first passage to a distinct boundary.

Intersegment links cause quick delocalization of the protein within the island. This is expressed by the weak dependence of the exit direction and the mean exit time on the initial position, as well as by the equivalence of symmetric initial positions for the mean exit time if the exit direction is specified. Therefore, the protein can be regarded as approximately uniformly distributed over the entire island interval. This suggests that – as far as transport on random walk chain conformations is concerned – it is legitimate to view islands as coarse-grained entities that can be described by two properties, their length and the trapping time corresponding to this length, neglecting any internal features. In this sense, a frozen random walk chain is equivalent to a chain of islands, whose length distribution and length-dependence of trapping times is well-known.

3.5 Toy model

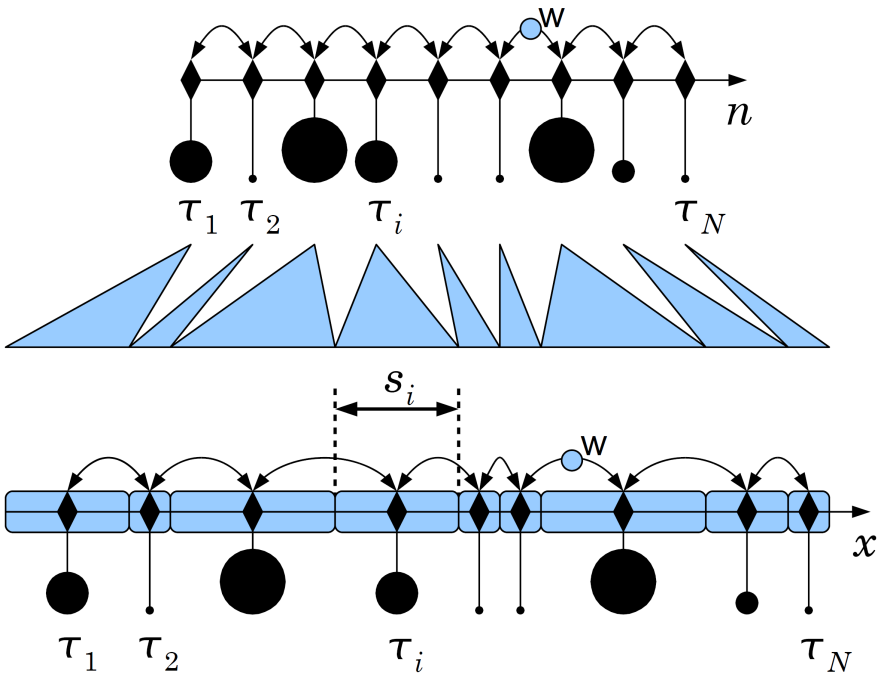


Fig. 3.14: Toy model to study transport on chains of independent entities with known statistical properties such as random walk islands. The model consists of a hopping environment (above) and a leaping environment (below) of sites with nearest neighbor connections. The lengths of the site-surrounding regions in the leaping environment are algebraically distributed, the site-specific characteristic waiting times are powers of these lengths. In the hopping environment the characteristic waiting times are preserved, whereas the region lengths are set to zero for simplicity.

These statistical properties of islands in frozen random walk conformations can be used within a coarse-grained toy model, which is designed to solve the problem of understanding why transport on such conformations is diffusive. The model exploits the fact that random walk islands are independent entities concatenated in contour space.

Correspondingly, the transport environment in this toy model is a chain of N sites P_i (with $1 \le i \le N$) with nearest neighbor connections between them. Each site has a position x_i, is surrounded symmetrically by a region of length s_i and has an intrinsic characteristic waiting time τ_i, which is a power of this length. This is shown in the lower half of Fig. 3.14.

The sites represent the islands of a frozen random walk conformation: The length of the region surrounding a site corresponds to the finite length of an island and the site-specific characteristic waiting time to the island trapping time.

Accordingly, in order to generate an instance of such a coarse-grained transport environment the region lengths are drawn independently from a power-law distribution:

$$\lambda(s_i) \propto s_i^{-(1+\mu)} \tag{3.52}$$

And the characteristic waiting times are calculated from the region lengths using

$$\tau_i = s_i^k \,. \tag{3.53}$$

In other words, lengths and times are distributed according to a joint probability distribution that reads:

$$j(s_i, \tau_i) = \lambda(s_i)\,\delta(\tau_i - s_i^k) \tag{3.54}$$

The *length distribution index* μ and the *waiting time power* k are freely adjustable parameters of the toy model. This fact will be used to characterize transport by deducing the transport exponent as a function of both parameters and – especially – to answer the question whether and why transport is diffusive for $\mu = 0.5$ and $k = 1.5$, which are the parameter values corresponding to islands of a frozen random walk conformation.

The toy model environment described so far will be denoted as *leaping environment*, due to the possibly large distances between adjacent sites. However, by neglecting the lengths of the site-surrounding regions but only keeping the site-specific characteristic waiting times, the leaping environment can be transformed into a so-called *hopping environment*, where the distances between adjacent sites are irrelevant.

According to this, the site positions in the hopping environment are denoted by numbers,

$$n \in \{1, 2, \ldots, N\} \,, \tag{3.55}$$

in contrast to the leaping environment, where contour coordinates are used:

$$x \in \{x_1, x_2, \ldots, x_N\} \,. \tag{3.56}$$

The interrelation between both types of environment is visualized in Fig. 3.14.

Transport on either environment is characterized best by a scaling relation between the size of an interval and the mean time a walker initially sitting in its center needs to leave it for the first time (first passage problem). This relation is called transport law. In the leaping environment one uses contour space intervals with length x which are symmetric about the walker's initial position at $x_{N/2}$. By contrast, the intervals used in the hopping environment are site number intervals which are symmetric about the walker's initial position at $N/2$ and contain n sites.

Given an interval of certain length (x or n) in an environment conformation with N sites, averaging the interval exit time over several walker trajectories yields the mean exit time for this interval length in this conformation. Then this mean exit time is averaged in turn over several independently generated environment conformations (each with N sites), which results in the mean first passage time for the interval length given.

In order to avoid divergences occurring in some parameter regimes, this disorder average is performed logarithmically:

$$T_{\mathrm{mfp}} = \bar{t}_{\text{mean exit}} \quad \Leftrightarrow \quad \ln(T_{\mathrm{mfp}}) = \langle \ln(t_{\text{mean exit}}) \rangle_{\text{conformations}} \tag{3.57}$$

By varying the interval lengths (x or n) the transport laws for each type of environment are obtained:

$$T_{\mathrm{mfp}} \propto x^{\beta} \propto n^{\omega} \tag{3.58}$$

The next question is, how the leaping and hopping transport exponents β and ω depend on the length distribution index μ and the waiting time power k .

3.5.1 Transport on the hopping environment: Simulations

In the hopping environment the transport exponent depends only on the ratio k/μ of both toy model parameters:

$$\omega(\mu, k) = \begin{cases} 2 & \text{for } \frac{k}{\mu} < 1 \\ 1 + \frac{k}{\mu} & \text{for } \frac{k}{\mu} > 1 \end{cases} \tag{3.59}$$

If this ratio is below unity, transport on the hopping environment is diffusive and subdiffusive otherwise.

To get this result the mean first passage time T_{mfp} has been measured on the hopping environment as a function of the interval length n . This function is a power-law for all combinations of parameter values μ and k (data not shown). The corresponding transport exponents $\omega(\mu, k)$ have been obtained by fitting a power-law to each function. Strikingly, the hopping transport exponent does not depend on the waiting time power and the length distribution index separately, but only on their ratio. This fact is shown in the collapse plot of Fig. 3.15, where the hopping transport exponent is plotted against this ratio: Although the exponent as a function of the waiting time power can be represented only by a set of curves, which is parametrized by the length distribution index, all the curves collapse to one single curve if the function variable (k) is rescaled by the curve set parameter (μ) .

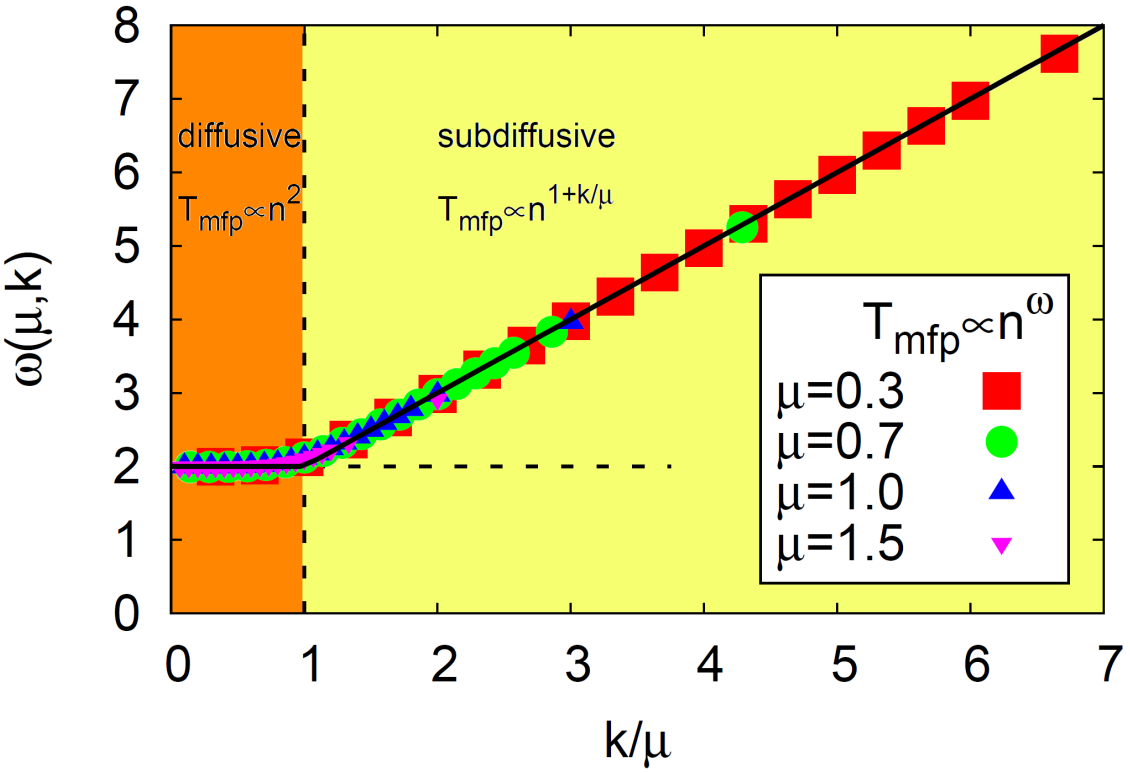


Fig. 3.15: The hopping transport exponent, which relates mean first passage time and interval length to each other in the hopping environment of the toy model, depends only on the ratio of waiting time power and length distribution index: All curves, which are functions of the former and parametrized by the latter, collapse to one single curve when rescaled this way. For parameter ratios below unity, transport on the hopping environment is diffusive (orange region), and subdiffusive otherwise (yellow region).

3.5.2 Transport on the hopping environment: Theory

The first goal is to explain the observed scaling relation between the mean first passage time and the interval length for transport on the hopping environment of the toy model theoretically. To do so, one cannot simply make use of the Lévy and the fractional Fokker-Planck formalisms, although the waiting times occurring during a random walk on the hopping environment are distributed according to a heavy-tailed power-law:

$$w(\tau) \equiv \int_0^\infty j(s,\tau)\, \mathrm{d}s \;\propto\; \tau^{-(1+\mu/k)} \tag{3.60}$$

The transport law that would be obtained if the Lévy formalism was used in spite of the quenched disorder of the characteristic waiting times obviously does not correspond to the observations in the previous chapter:

$$T_{\mathrm{mfp}}^{\mathrm{uncorrelated}} \propto n^{2k/\mu} \tag{3.61}$$

The reason is that the correlations between subsequently drawn waiting times, which result from this quenched disorder, are neglected illicitly by doing so.

Although inappropriate here, the **Lévy formalism** will become useful below when performing the disorder average on the hopping environment and finally when transferring these results back to the leaping environment. Nevertheless, the formalism adequate to accounting for the quenched disorder present in a single conformation of the hopping environment is the generic **mean first passage time formalism**, as described by Gardiner [134]: Given an interval of length m containing $m+1$ sites $(0,1,\ldots,m-1,m)$, the boundary sites 0 and m being absorbing sites, the mean exit time for a nearest neighbor random walk starting from site m_0 with fixed transition rates $k_j^{+/-}$ for hopping from site j to its neighbor sites $j+1$ or $j-1$, respectively, reads:

$$t_{\text{exit}}(m,m_0,\{k_j^{+/-}\})=P_{\text{exit}}^{+}(m_0)\sum_{i=m_0}^{m-1}\Phi_i\sum_{j=1}^{i}\frac{1}{k_j^{+}\Phi_j}+\left[1-P_{\text{exit}}^{+}(m_0)\right]\sum_{i=1}^{m_0}\Phi_i\sum_{j=i}^{m-1}\frac{1}{k_j^{+}\Phi_j} \tag{3.62}$$

The characteristic waiting time before a hopping event in the hopping environment of the toy model only depends on the walker position and not on the hopping direction. Therefore, both hopping rates belonging to a site are equal:

$$k_j^{+/-}=\frac{1}{\tau_j} \tag{3.63}$$

This lack of directional bias leads to rather simple expressions for the products of hopping rate ratios and the probability for being absorbed at the right boundary:

$$\Phi_j\equiv\prod_{i=1}^{j}\frac{k_i^{-}}{k_i^{+}}=1,\quad j\in\{0,1,\ldots,m-1\} \tag{3.64}$$

$$P_{\text{exit}}^{+}(m_0)\equiv\frac{\sum_{j=0}^{m_0-1}\Phi_j}{\sum_{j=0}^{m-1}\Phi_j}=\frac{m_0}{m} \tag{3.65}$$

Inserting these expressions simplifies the mean exit time formula considerably:

$$t_{\text{exit}}(m,m_0,\{\tau_j\})=\frac{m_0}{m}\sum_{i=m_0}^{m-1}\sum_{j=1}^{i}\tau_j+\left(1-\frac{m_0}{m}\right)\sum_{i=1}^{m_0}\sum_{j=i}^{m-1}\tau_j \tag{3.66}$$

In order to compare to the simulation results one has to start each random walk from the center of the interval, which yields:

$$t_{\text{exit}}(m,m/2,\{\tau_j\})=\frac{1}{2}\left[\frac{m}{2}\sum_{j=1}^{m-1}\tau_j+\frac{m}{2}\tau_{m/2}+\sum_{j=1}^{m/2-1}j(\tau_j+\tau_{m-j})\right] \tag{3.67}$$

As this expression shows, the mean exit time for escape from an interval with absorbing boundaries scales approximately with the number $n=m-1$ of the inner sites of the interval multiplied by the sum of the site-specific characteristic waiting times:

$$t_{\text{exit}}\propto n\sum_{j=1}^{n}\tau_j \tag{3.68}$$

Since the characteristic waiting times are independent random variables with known distribution (equation 3.60), this important relation is a valuable input for the calculation of the disorder average: A sum of random variables is a random variable in turn. If

the summands are independent and identically albeit broadly distributed random variables – as it is the case here, the distribution of their sum is calculable by applying the Lévy Khinchin theorem, which generalizes the central limit theorem for broad distributions.

Following these considerations, one has to define the so-called time sum Y as a new random variable representing the sum of the characteristic waiting times belonging to the set of sites of a particular escape interval within a particular conformation of the hopping environment:

$$Y \equiv \sum_{j=1}^{n} \tau_j \tag{3.69}$$

For $k<\mu$ the waiting time distribution $w(\tau)$ is not broad and a mean waiting time, which equals the disorder average over all characteristic waiting times occurring in the entire set of environment conformations, exists:

$$\bar{\tau} = \langle \tau \rangle \equiv \int_0^\infty w(\tau)\tau \, \mathrm{d}\tau \; < \; \infty \tag{3.70}$$

Therefore, the disorder average of the time sum exists in turn:

$$\bar{Y} \propto n\bar{\tau} \tag{3.71}$$

This results in diffusive scaling of the mean first passage time, which is the disorder average of the mean exit time:

$$T_{\mathrm{mfp}} \equiv \bar{t}_{\mathrm{exit}} \equiv n\bar{Y} \propto n^2 \quad \text{for} \quad k<\mu \tag{3.72}$$

Whereas in the previous parameter regime the disorder average of the mean exit time has been calculated straight-forward, one has to make use of the Lévy-Khinchin theorem for $k>\mu$, where the waiting time distribution is broad and accordingly has no mean. Therefore, the normal (non-logarithmic) disorder average of the time sum diverges in this regime, as well. However, a scaling relation between time sum and number of summands is obtained from the fact that the time sum is distributed according to a stable Lévy distribution [135][126]:

$$W(Y,n) = n^{-1/\nu} L_{\nu,1}(n^{-1/\nu} Y) \tag{3.73}$$

The divergent normal disorder average of the time sum can be replaced by its characteristic value, which scales with the number of summands like the width of the stable Lévy distribution:

$$\bar{Y} \propto n^{1/\nu} \tag{3.74}$$

The function $L_{\nu,1}$ in this distribution is defined for Lévy indices $\nu \equiv \mu/k \neq 1$ by its Fourier transform (Since all summands in the time sum are positive, the bias index is 1) [125]:

$$\tilde{L}_{\nu,1} = \exp\left(-|q|^\nu \left[1 - i \tan\left(\frac{\pi}{2}\nu \right) \mathrm{sgn}(q) \right] \right) \tag{3.75}$$

The scaling behavior of the time sum in the regime of diverging mean waiting time results in subdiffusive scaling of the mean first passage time:

$$T_{\mathrm{mfp}} \equiv \bar{t}_{\mathrm{exit}} \equiv n\,\bar{Y} \propto n^{1+k/\mu} \quad \text{for} \quad k > \mu \tag{3.76}$$

On the whole, the previous theoretical deduction confirms and explains the simulation results of chapter 3.5.1 qualitatively as well as quantitatively, so that the characterization of transport on the hopping environment of the toy model is complete:

- The transport exponent is a function of the ratio of both parameters of the toy model there.
- There are two types of transport: diffusion for existing mean waiting time and subdiffusion for diverging mean waiting time.

As random walk islands have a waiting time power greater than their length distribution index, transport on the hopping environment is subdiffusive if this environment represents the island chain of a frozen random walk conformation. However, it is necessary to prove that, if these islands are represented instead by the leaping environment with its site-surrounding regions spanning contour space, transport is diffusive – just like in the contour space of the original random walk conformation. For that purpose, the results obtained for the hopping environment will be used in the following in order to characterize transport on the leaping environment for the whole parameter space of the toy model.

3.5.3 Transport on the leaping environment: Theory

To achieve that, a scaling relation between distances on the hopping environment, where the sites are mere points without surrounding regions, and distances on the leaping environment, which represents contour-space correctly by assigning an algebraically distributed region length to each site, is needed. Then, not only the influence the site-specific characteristic waiting times have on transport is taken into account, but also the influence of the site-specific region lengths.

The distance covered in the leaping environment when n particular adjacent sites are crossed in the hopping environment is given by summing the region lengths of these sites:

$$X = \sum_{j=1}^{n} s_j \tag{3.77}$$

The question is how the characteristic value of this length sum, which results from a disorder average over many possible sets of region lengths, scales with the number of sites. Since the region lengths are independent and identically distributed random variables, the length sum X is a random variable whose distribution can be determined in turn.

If the length distribution index, which governs the exponent of the distribution of region lengths (cf. equation 3.52) is big enough ($\mu > 1$), a mean region length exists:

$$\langle s \rangle \equiv \int_0^\infty \lambda(s)\, s\, \mathrm{d}s \; < \; \infty \tag{3.78}$$

In this case, the characteristic length sum is simply proportional to the number of summands:

$$\bar{X}=\langle X\rangle \propto n \tag{3.79}$$

However, if $\mu<1$, the distribution of region lengths is broad, which means that no mean region length exists. Then, the distribution of the length sum is – again according to the Lévy-Khinchin theorem – a stable Lévy distribution with Lévy index μ and bias index 1:

$$W(X,n)=n^{-1/\mu}L_{\mu,1}(n^{-1/\mu}X) \tag{3.80}$$

In this case the characteristic length sum is proportional to the width of this stable Lévy distribution, which is a power of the number of summands:

$$\bar{X}\propto n^{1/\mu} \tag{3.81}$$

To visualize this, the distribution $W(X,n)$ of the total contour length X occupied by the site-surrounding regions of n adjacent sites in the leaping environment has been measured for several site numbers n in the regime of diverging mean region length ($\mu<1$). Fig. 3.16 shows such sets of contour length distributions obtained for two different length distribution indices μ. The contour length distributions collapse to an n-independent function if rescaled by a power of the corresponding site number n. That means that the distribution width is proportional to that power. Therefore, this power measures the characteristic total contour length for the respective site number. As expected for a stable Lévy distribution, the power-law tail of the collapse function reflects the power-law of the region length distribution.

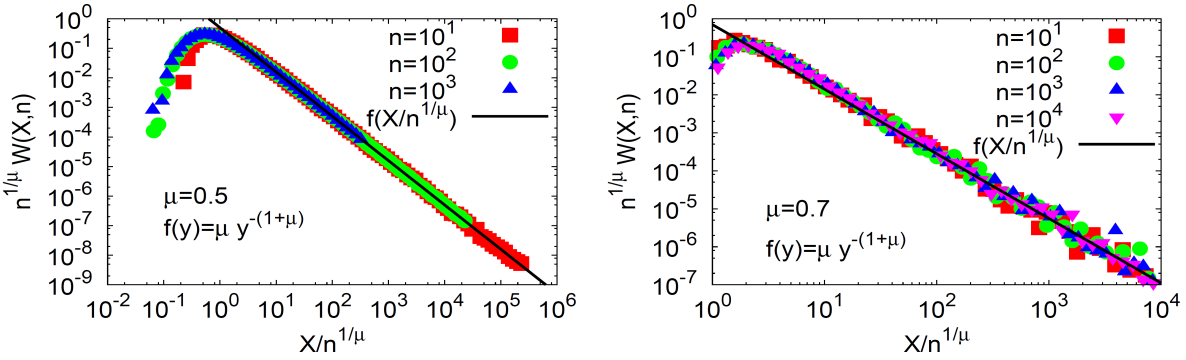


Fig. 3.16: The distributions W of the total contour length covered by a set of n adjacent sites in the leaping environment collapse to a function independent of the number of sites if rescaled by the corresponding distribution widths. For diverging mean length of the site-surrounding regions ($\mu<1$) this distribution width is a μ-dependent power of the number of sites and measures the characteristic total contour length. In this case the collapse function is a stable Lévy distribution with a power-law tail determined by the length distribution index μ.

To summarize results, the scaling relation between distances x on the leaping environment and distances n on the hopping environment is given by:

$$x\propto\begin{cases} n^{1/\mu} & \text{for } \mu<1 \\ n & \text{for } \mu>1 \end{cases} \tag{3.82}$$

Inserting this into the transport law of the hopping environment (equations 3.72 and 3.76) yields the transport law of the leaping environment and the leaping transport exponent β :

$$T_{\mathrm{mfp}} \propto n^{\omega(\mu,k)} \propto \begin{Bmatrix} x^{\mu\omega(\mu,k)} & \text{for } \mu<1 \\ x^{\omega(\mu,k)} & \text{for } \mu>1 \end{Bmatrix} = \begin{pmatrix} x^{2\mu} & \text{superdiffusive} & \text{for } \mu<1 \wedge k<\mu \\ x^{\mu+k} & \text{superd., diff., subd.} & \text{for } \mu<1 \wedge k>\mu \\ x^{2} & \text{diffusive} & \text{for } \mu>1 \wedge k<\mu \\ x^{1+k/\mu} & \text{subdiffusive} & \text{for } \mu>1 \wedge k>\mu \end{pmatrix} = x^{\beta(\mu,k)} \tag{3.83}$$

This mean first passage time scaling relation characterizes transport on the leaping environment of the toy model in the entire parameter space (see appendix B for a comparison with uncorrelated continuous time random walks whose jump lengths and/or waiting times have power-law distributions). Especially, it can be used to predict the scaling exponent for transport in the contour space of a frozen random walk conformation, which is representable by a particular point of the parameter space of the toy model. The transport law of the leaping environment has been deduced theoretically from the transport behavior in the hopping environment, which is already known from simulations that have been confirmed theoretically. However, before it can be applied to the random walk question, the transport law of the leaping environment has to be verified by toy model simulations in the relevant parameter regime.

3.5.4 Transport on the leaping environment: Simulations

To this end, 3000 conformations of the leaping environment of the toy model are generated, each with 10000 region-surrounded sites. The length distribution index μ is set to 0.7 (below 1!) in order to obtain a diverging mean region length when drawing the region lengths from their distribution, just like the mean island length diverges in frozen random walk conformations. Within this regime, the waiting time power k , which relates the characteristic waiting time of an interval site to its region length, is varied over its entire domain in order to explore all sub-regimes. The mean first passage time for an interval length x in contour space results from taking the exit time for one particular interval with this length and performing a disorder average over 3000 such equally long intervals, one for each conformation of the leaping environment.

The results are shown in Fig. 3.17: For each combination of the two toy model parameters μ and k the mean first passage time scales with contour length like a power-law. Moreover, the observed transport exponents follow the theoretical prediction concerning their dependence on μ and k quantitatively:

$$T_{\mathrm{mfp}} \propto \begin{Bmatrix} x^{2\mu} & \text{for } \mu<1 \wedge k<\mu \\ x^{k+\mu} & \text{for } \mu<1 \wedge k>\mu \end{Bmatrix} \tag{3.84}$$

As a consequence of this observation, transport on the leaping environment is superdiffusive for $k<2-\mu$, diffusive for $k=2-\mu$ and subdiffusive for $k>2-\mu$ as long as the mean length of the site-surrounding regions diverges.

The numerical results in the regime of existing mean region length ($\mu > 1$), which is, however, irrelevant when studying transport on frozen random walk chains, confirm the theoretical results as well (data not shown).

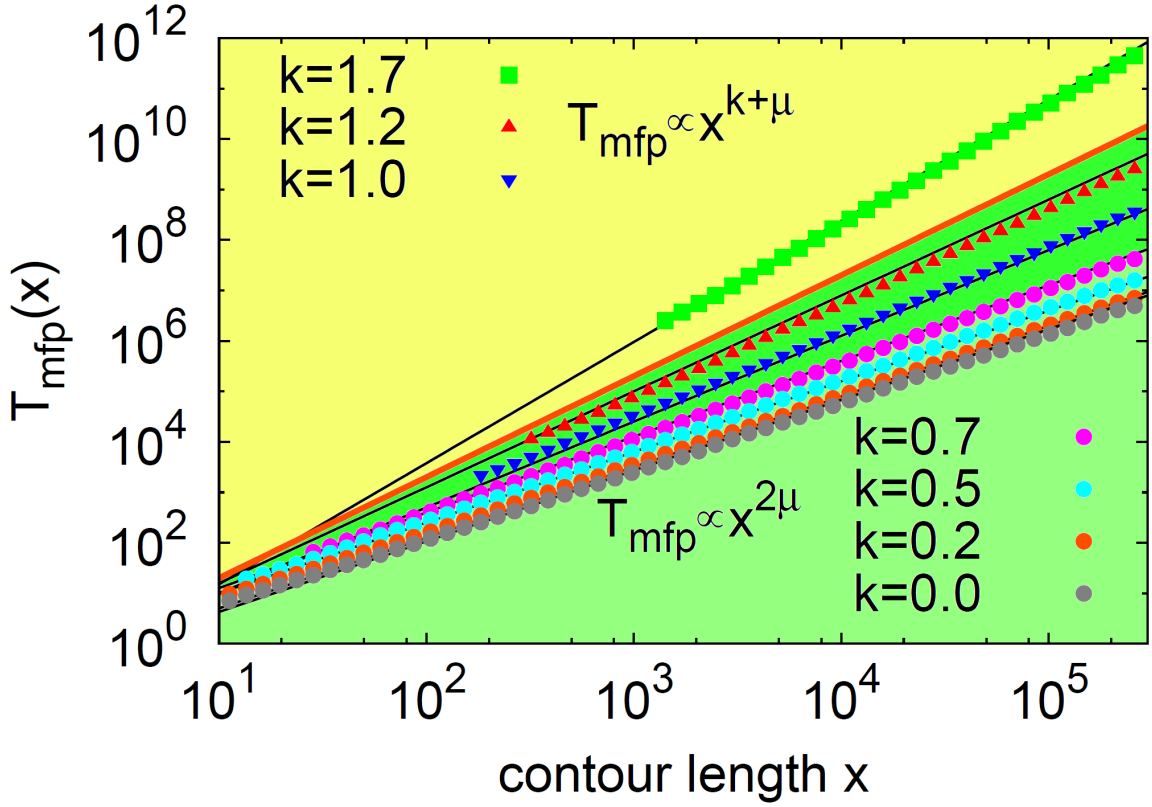


Fig. 3.17: The disorder-averaged mean first passage time for escaping from an interval on the leaping environment of the toy model after having started from the interval center, plotted as a function of the contour length covered by the interval, is a power-law for each parameter combination. The length distribution index is chosen such that the mean length of the regions surrounding the interval sites diverges ($\mu = 0.7 < 1$). For waiting time powers smaller than this value ($k < \mu$) the transport law is independent of the waiting time power (independence regime), otherwise the transport exponent is the sum of waiting time power and length distribution index (dependence regime). The independence regime is purely superdiffusive (light green), the dependence regime consists of a superdiffusive regime (dark green) for $\mu < k < 2 - \mu$ and a subdiffusive regime (yellow) for $k > 2 - \mu$. The orange line represents diffusive scaling (exponent 2).

3.5.5 Parameter space diagrams

Having now characterized both transport environments of the toy model entirely by numerical as well as by theoretical studies, the results can be summarized with the help of two parameter space diagrams (the left part of Fig. 3.18 for the hopping environment and the right part for the leaping environment).

In both the hopping and the leaping environment the transport exponent is independent of the waiting time power if and only if the mean characteristic waiting time

exists, i.e. if the waiting time distribution is narrow and not broad. Therefore, this independence regime is given for $k<\mu$, its counterpart, the dependence regime with broad waiting time distribution and diverging mean characteristic waiting time for $k>\mu$.

In the hopping environment, where there are no site-surrounding regions, only the shape of the waiting time distribution decides on the quality of transport: In the independence regime, where the mean characteristic waiting time exists, transport is normal diffusive whereas in the dependence regime, where the mean characteristic waiting time diverges, it is subdiffusive.

In the leaping environment however, where the length distribution of the site-surrounding regions plays an additional role, this subdivision into a normal diffusive and a subdiffusive transport regime for existing and diverging mean characteristic waiting times respectively is only valid for length distribution indices μ greater than one: Then the region length distribution is narrow and a mean region length exists.

However, if the length distribution index is smaller than one, thus causing a broad region length distribution with diverging mean region length, transport on the leaping environment deviates substantially from transport on the hopping environment:

If – unlike the mean region length – the mean characteristic waiting time exists $(\mu<1 \wedge k<\mu)$, transport is superdiffusive (with an transport exponent independent of k). If both the mean region length and the mean characteristic waiting time diverge $(\mu<1 \wedge k>\mu)$, the resulting regime, whose transport exponent is the sum of μ and k, exhibits tripartite transport behavior, depending on the value of this sum: superdiffusion for values smaller than two, quasi-diffusion for a value of two and subdiffusion for values greater than two.

This manifests oneself in a thin separating line of quasi-diffusion in the parameter space diagram of the leaping environment, which originates from the spacious regime of normal diffusion. Whereas normal diffusive transport needs the existence of both the mean region length and the mean characteristic waiting time $(\mu>1 \wedge k<\mu)$, quasi-diffusive transport results from a subtle balance between the length distribution index and the waiting time power in a regime where both the mean region length and the mean characteristic waiting time diverge $(\mu<1 \wedge k>\mu)$.

This balance can be formulated generally as a so-called balance condition for quasi-diffusive transport on environments consisting of concatenated but independent entities whose lengths are distributed according to a broad power-law (index μ) and whose trapping times are proportional to a power of the corresponding lengths (exponent k):

$$\mu+k=2 \tag{3.85}$$

As the contour space of a frozen random walk chain conformation is an environment consisting of concatenated but independent entities, namely the intersegment link islands, whose lengths are distributed according to the broad power-law

$$\lambda(s_i)\propto s_i^{-3/2} \tag{3.86}$$

and whose trapping times are related to the corresponding lengths by

$$\tau_i\propto s_i^{3/2}, \tag{3.87}$$

the balance condition is fulfilled for frozen random walk chains ($\mu=1/2,\ k=3/2$). Accordingly, the theoretical results deduced with the help of the toy model predict that transport on the contour space of frozen random walk chain conformations is quasi-diffusive.

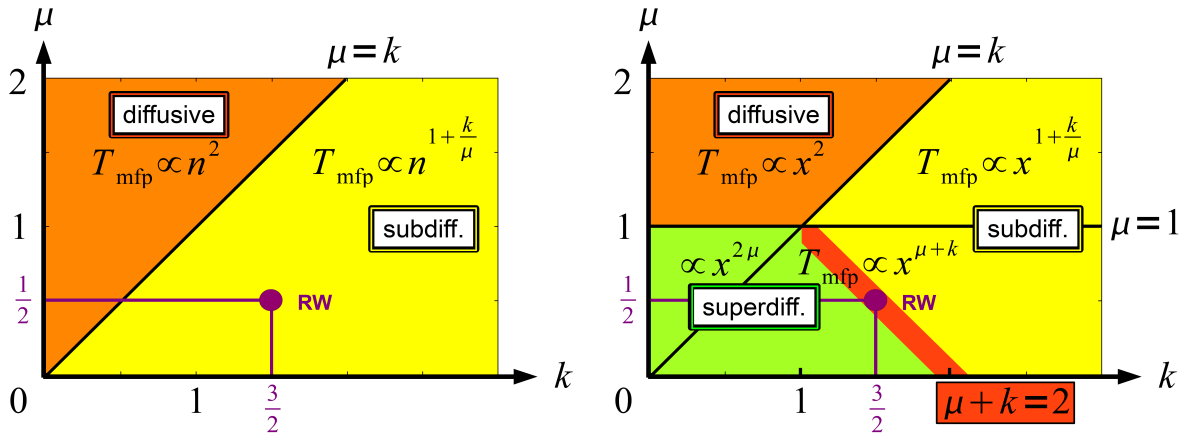


Fig. 3.18: Parameter space diagrams for the hopping (left) and and the leaping environment (right) of the toy model, each spanned by the length distribution index μ and the waiting time power k. Both diagrams are subdivided by the quality of transport: subdiffusion (yellow), diffusion (orange) and superdiffusion (green). Depending on the quotient of the toy model parameters and thus on the existence of the mean characteristic waiting time, transport on the hopping environment is either normal diffusive ($k/\mu<1$) or subdiffusive ($k/\mu>1$). In the leaping environment, this bipartite behavior is observed only for existing mean region length ($\mu>1$). However, if the region length distribution is broad ($\mu<1$), transport on the leaping environment is either superdiffusive ($\mu+k<2$), quasi-diffusive ($\mu+k=2$) or subdiffusive ($\mu+k>2$). The point of parameter space representing transport on frozen random walk conformations (magenta) lies in the subdiffusive transport regime for the hopping environment and on the line of quasi-diffusion between the super- and subdiffusive regimes in the leaping environment. Therefore, diffusive transport in the contour space of frozen random walk conformations, as observed in simulations, can be explained theoretically with the help of the toy model.

3.6 Conclusion

Optimizing the time proteins need to find their specific binding sites on DNA chains is of central importance for the efficiency and performance of cellular functions. The interplay of protein sliding and protein intersegment transfer with simultaneous conformational changes of the DNA is of special interest in this optimization problem:

Protein sliding is needed for thorough exploration of short intervals of the DNA contour and eventually allows the protein to arrive at its target site. Protein intersegment transfer, on the other hand, denotes long-range jumps necessary for quickly changing between remote regions on the DNA contour. And finally, conformational changes of the DNA connect and disconnect such remote regions dynamically. Implementing the com-

bination of these types of protein and DNA movement within a coarse-grained lattice model allows to study protein transport in contour space of a DNA chain varying its shape.

This model can reproduce the known limits of protein transport behavior, which have already been studied numerically and theoretically [73]:

- Superdiffusion is observed for infinitely fast conformational changes of the DNA and caused by uncorrelated long jumps. This limit has been predicted using the Lévy-Khinchin theorem.
- By contrast, without conformational changes transport is purely diffusive, in spite of the long jumps. The postulated cause for this paradoxical behavior are strong geometrical correlations.

Beyond the reproduction of these limits, the model is also able to represent the intermediate transport regime in between, due to a freely adjustable rate for conformational changes. Interestingly, this intermediate regime exhibits a dynamical cross-over between superdiffusion for short times and diffusion for long times.

In order to obtain a theoretical starting point to explain this cross-over, it is necessary to develop a theory of diffusive protein transport in contour space of frozen DNA chains in spite of a broad jump length distribution and poissonian time steps. This theory should verify Sokolov's postulate of geometrical correlations and explain the cause of their existence and their effect on transport in detail. This is the goal of these studies.

To this end, at first all relevant types of correlations that may occur in systems with broadly distributed jumps and poissonian time steps have been analyzed with respect to their ability to make transport diffusive. The results are as follows:

- Without spatial and temporal correlations between the jumps (annealed disorder), transport is superdiffusive and one obtains a Lévy-flight.
- Eliminating all spatial correlations but keeping the temporal correlations by freezing the resulting intersegment linking pattern (quenched disorder without spatial correlations) still yields superdiffusive transport.
- By contrast, transport on frozen *native* intersegment linking patterns is diffusive, because all spatial correlations have been preserved (quenched disorder with spatial correlations).

Therefore, frozenness alone is not sufficient for diffusive transport. In fact, one needs temporally invariant spatial correlations to observe diffusive transport in spite of long jumps. These can be identified with the strong geometrical correlations mentioned in Sokolov's postulate.

A further characterization of these geometrical correlations has been attempted by studying their order in terms of the number of participating intersegment links: It exceeds the order of pair correlations. And due to the self-similar nature of native in-

tersegment linking patterns probably all orders contribute to the correlations necessary for diffusive transport.

However, the geometrical correlations manifest themselves in another feature, which can be characterized more easily: They cause the formation of separate and independent islands in the intersegment linking pattern of frozen chain conformations. These islands have interesting properties:

- As they consist of intersegment links arranged in a self-similar way, their length distribution in contour space equals the length distribution of intersegment links.
- The disorder-averaged mean time needed for symmetric escape from the contour space interval covered by an island scales superdiffusively with the island length.
- The exact position of the protein within an island plays only a negligible role, since the intersegment linking pattern the island consists of quickly delocalizes the protein over the entire island range.

All these facts allow to coarse-grain islands as black-box entities without internal structure and in turn frozen DNA conformations as chain of such entities. Each entity is defined by its length and its intrinsic trapping time only. The lengths of the entities representing islands are distributed according to the measured island length distribution, their trapping times are calculated from the corresponding lengths by making use of the measured transport law for symmetric escape from islands.

In order to study transport on chains of independent black-box entities with freely adjustable length distribution and escape time scaling law in general, a toy model has been created. This toy model is especially able to represent chains of islands, i.e. the random walk chains frozen DNA molecules are modeled by. The numerical and theoretical characterization of transport on this toy model allows the following conclusions concerning transport on frozen DNA conformations:

- Diffusion in the contour space of frozen DNA conformations is not normal, but in fact a kind of quasi-diffusion, which is rendered possible by the existence of islands.
- The particular shapes of island length distribution and island escape time scaling law together are necessary conditions for this subtle balance between super- and subdiffusion.

Therefore, the goal to establish a theory of diffusive transport on frozen DNA conformations which verifies Sokolov's postulate has been accomplished:

The intersegment linking pattern of any DNA conformation (i.e. a native pattern) contains strong geometrical correlations in all orders of link-tupels. If the conformation is frozen, these correlations of the linking pattern are preserved. They manifest themselves in the formation of separate islands, which form a chain covering the contour space of the conformation. The local length and escape time properties of these islands create a subtle balance which makes next-neighbor transport on the island chain quasi-diffusive.

The dynamical cross-over between superdiffusive transport for short times and diffusive transport for long times then can be explained as follows: The faster the DNA conformation changes with respect to the protein's motility, the faster the strong geometrical correlations in the intersegment linking pattern are destroyed and the faster islands change their size, shift, split, merge or even appear and disappear and the larger – as a consequence – the superdiffusive regime of uncorrelated protein jumps becomes.

Appendix A: Brownian dynamics algorithm

The new position (time $t+\Delta t$) of each of the N beads the RNA model chain consists of is calculated from the corresponding old position at time t by using the forward Euler-Maruyama integration method:

$$\boldsymbol{r}_k(t+\Delta t)=\boldsymbol{r}_k(t)+\frac{\Delta t}{\zeta_k}(\boldsymbol{F}_{\text{total}})_k\left(\boldsymbol{r}_1(t),\boldsymbol{r}_2(t),\ldots,\boldsymbol{r}_N(t)\right)+\sqrt{2kT\frac{\Delta t}{\zeta_k}}\boldsymbol{\eta}_k(t)\quad\text{for}\quad 1\leqslant k\leqslant N \tag{A.1}$$

Thermal fluctuations are included by the random vectors $\boldsymbol{\eta}_k(t)$, whose components, which are uncorrelated and obey a normal distribution with zero mean and unity variance, are calculated from random variables $X_{k\alpha}$ uniformly distributed over the unit interval by using the Box Muller method:

$$\eta_{k\alpha}(t)=\sqrt{-2\ln\left(1-X_{k\alpha}(t)\right)}\sin\left(2\pi X_{k\alpha}(t)\right) \tag{A.2}$$

The total conservative force consists of forces calculated from potentials by differentiation:

$$(\boldsymbol{F}_{\text{total}})_k=(\boldsymbol{F}_{\text{pore}}+\boldsymbol{F}_{\text{bond}}+\boldsymbol{F}_{\text{LJ}}+\boldsymbol{F}_{\text{torsion}}+\boldsymbol{F}_{\text{bending}}+\boldsymbol{F}_{\text{tilt}})_k \tag{A.3}$$

The resulting contributions to the conservative force will be enlisted in what follows:

- Every bead of the RNA model chain interacts with the barrier and the drift channel of the pore:

$$\begin{aligned}(\boldsymbol{F}_{\text{pore}})_k\equiv-\nabla_k U_{\text{pore}}=&-\frac{E_{\text{barrier}}}{d_{\text{pore}}^2}\left[(\boldsymbol{r}_k)_x\hat{\boldsymbol{e}}_x+(\boldsymbol{r}_k)_y\hat{\boldsymbol{e}}_y\right]\exp\left(-\frac{(\boldsymbol{r}_k)_x^2+(\boldsymbol{r}_k)_y^2}{2d_{\text{pore}}^2}\right)\exp\left(-\frac{(\boldsymbol{r}_k)_z^2}{2l_{\text{pore}}^2}\right)\\&+\frac{E_{\text{barrier}}}{l_{\text{pore}}^2}(\boldsymbol{r}_k)_z\hat{\boldsymbol{e}}_z\left[1-\exp\left(-\frac{(\boldsymbol{r}_k)_x^2+(\boldsymbol{r}_k)_y^2}{2d_{\text{pore}}^2}\right)\right]\exp\left(-\frac{(\boldsymbol{r}_k)_z^2}{2l_{\text{pore}}^2}\right)\\&-\frac{E_{\text{drift}}}{l_{\text{pore}}}\hat{\boldsymbol{e}}_z\cosh^{-2}\left(\frac{2(\boldsymbol{r}_k)_z}{l_{\text{pore}}}\right)\end{aligned} \tag{A.4}$$

- Beads adjacent along the chain backbone interact by harmonic forces:

$$\begin{aligned}(\boldsymbol{F}_{\text{bond}})_k\equiv-\nabla_k U_{\text{bond}}=&\theta_{1\leqslant k\leqslant N-1}K_{\text{bond}}\left(r_{k,k+1}-r_0\right)\hat{\boldsymbol{r}}_{k,k+1}\\&-\theta_{2\leqslant k\leqslant N}K_{\text{bond}}\left(r_{k-1,k}-r_0\right)\hat{\boldsymbol{r}}_{k-1,k}\end{aligned} \tag{A.5}$$

- The shape of the binding interaction within an arbitrary pair of bases depends on the compatibility of the bases:

$$(\boldsymbol{F}_{\mathrm{LJ}})_k \equiv -\nabla_k U_{\mathrm{LJ}} = \sum_{j=k+1+\Delta}^{N} \hat{\boldsymbol{r}}_{kj} \frac{\partial}{\partial r_{kj}} \left[c_{kj} \Phi_{\mathrm{LJ}}^{\mathrm{comp}}(r_{kj}) + (1-c_{kj}) \Phi_{\mathrm{LJ}}^{\mathrm{incomp}}(r_{kj}) \right] - \sum_{i=1}^{k-1-\Delta} \hat{\boldsymbol{r}}_{ik} \frac{\partial}{\partial r_{ik}} \left[c_{ik} \Phi_{\mathrm{LJ}}^{\mathrm{comp}}(r_{ik}) + (1-c_{ik}) \Phi_{\mathrm{LJ}}^{\mathrm{incomp}}(r_{ik}) \right] \tag{A.6}$$

Binding interactions between two compatible bases are mediated by Lennard-Jones forces for all distances:

$$\frac{\partial}{\partial r} \Phi_{\mathrm{LJ}}^{\mathrm{comp}}(r) = \frac{12 E_{\mathrm{LJ}}}{R} \left[\left(\frac{R}{r} \right)^{13} - \left(\frac{R}{r} \right)^{7} \right] \tag{A.7}$$

Binding forces between two incompatible bases become harmonic for small distances, where the interaction is already repulsive:

$$\frac{\partial}{\partial r} \Phi_{\mathrm{LJ}}^{\mathrm{incomp}}(r) = \begin{cases} \dfrac{12 E_{\mathrm{LJ}}^{\mathrm{incomp}}}{R} \left[\left(\dfrac{R}{r} \right)^{13} - \left(\dfrac{R}{r} \right)^{7} \right] & \text{for} \quad r \geq \dfrac{R}{2} \\ 2ar + b & \text{for} \quad 0 \leq r < \dfrac{R}{2} \end{cases} \tag{A.8}$$

- Torsion forces involve angular interaction between two strong base-pairs, which causes two pairs of sums, each containing products of two binding functions:

$$(\boldsymbol{F}_{\mathrm{torsion}})_k \approx \sum_{j=k+3+\Delta}^{N} \hat{\boldsymbol{D}}_k \left[H_k^j H_{k+1}^{j-1} U_{\mathrm{torsion}}^{kj} \right] + \sum_{i=1}^{k-3-\Delta} \hat{\boldsymbol{D}}_k \left[H_i^k H_{i+1}^{k-1} U_{\mathrm{torsion}}^{ik} \right] + \sum_{j=k+2+\Delta}^{N} \hat{\boldsymbol{D}}_k \left[H_{k-1}^j H_k^{j-1} U_{\mathrm{torsion}}^{k-1,j} \right] + \sum_{i=1}^{k-2-\Delta} \hat{\boldsymbol{D}}_k \left[H_i^{k+1} H_{i+1}^k U_{\mathrm{torsion}}^{i,k+1} \right] \tag{A.9}$$

- Bending and inclination forces involve angular interaction between three strong base-pairs, which causes three pairs of sums, each containing products of three binding functions:

$$\begin{aligned} (\boldsymbol{F}_{\mathrm{bending}} + \boldsymbol{F}_{\mathrm{tilt}})_k \approx & \sum_{j=k+5+\Delta}^{N} \hat{\boldsymbol{D}}_k \left[H_k^j H_{k+1}^{j-1} H_{k+2}^{j-2} (U_{\mathrm{bending}}^{kj} + U_{\mathrm{tilt}}^{kj}) \right] + \sum_{i=1}^{k-5-\Delta} \hat{\boldsymbol{D}}_k \left[H_i^k H_{i+1}^{k-1} H_{i+2}^{k-2} (U_{\mathrm{bending}}^{ik} + U_{\mathrm{tilt}}^{ik}) \right] \\ & + \sum_{j=k+4+\Delta}^{N} \hat{\boldsymbol{D}}_k \left[H_{k-1}^j H_k^{j-1} H_{k+1}^{j-2} (U_{\mathrm{bending}}^{k-1,j} + U_{\mathrm{tilt}}^{k-1,j}) \right] + \sum_{i=1}^{k-4-\Delta} \hat{\boldsymbol{D}}_k \left[H_i^{k+1} H_{i+1}^k H_{i+2}^{k-1} (U_{\mathrm{bending}}^{i,k+1} + U_{\mathrm{tilt}}^{i,k+1}) \right] \\ & + \sum_{j=k+3+\Delta}^{N} \hat{\boldsymbol{D}}_k \left[H_{k-2}^j H_{k-1}^{j-1} H_k^{j-2} (U_{\mathrm{bending}}^{k-2,j} + U_{\mathrm{tilt}}^{k-2,j}) \right] + \sum_{i=1}^{k-3-\Delta} \hat{\boldsymbol{D}}_k \left[H_i^{k+2} H_{i+1}^{k+1} H_{i+2}^k (U_{\mathrm{bending}}^{i,k+2} + U_{\mathrm{tilt}}^{i,k+2}) \right] \end{aligned} \tag{A.10}$$

Due to the product-type structure of the angular potentials (torsion, bending and inclination potential) force calculation by numerical differentiation is – from a computational perspective – much more efficient than analytical differentiation. Therefore all angular forces are obtained applying the difference quotient operator instead of the nabla operator:

$$\boldsymbol{D}_k \left[f(\boldsymbol{r}_k) \right] = \frac{1}{h_{\mathrm{diff}}} \sum_{\alpha=1}^{3} \hat{\boldsymbol{e}}_\alpha \left[f\left(\boldsymbol{r}_k + \frac{h_{\mathrm{diff}}}{2} \hat{\boldsymbol{e}}_\alpha \right) - f\left(\boldsymbol{r}_k - \frac{h_{\mathrm{diff}}}{2} \hat{\boldsymbol{e}}_\alpha \right) \right] \quad \text{with} \quad \frac{h_{\mathrm{diff}}}{r_0} = 10^{-2} \ll 1 \tag{A.11}$$

All other conservative forces are calculated by analytical differentiation.

Appendix B: Fractional calculus and CTRWs

There are two types of CTRWs: CTRWs whose jump lengths are correlated with the corresponding waiting times (inter-correlated CTRWs), and CTRWs without such length-time correlations (normal CTRWs). The latter type is intensively treated in [135], which shall be summarized here in order to understand the special properties of the inter-correlated CTRW's playing an important role in this dissertation.

Regardless of its type, a CTRW is characterized by the joint distribution of jump length and waiting time (joint jump distribution):

$$j(x,t) \tag{B.1}$$

Integrating over all waiting times yields the jump length distribution:

$$\lambda(x)=\int_0^\infty \mathrm{d}\,t\, j(x,t) \tag{B.2}$$

Integrating over all jump lengths yields the waiting time distribution:

$$w(t)=\int_{-\infty}^\infty \mathrm{d}\,x\, j(x,t) \tag{B.3}$$

The Fourier-Laplace transform of the walker probability density function $W(x,t)$ (walker PDF) can be calculated from the Laplace transform of the waiting time distribution and the Fourier-Laplace transform of the joint jump distribution [136]:

$$W_{F,L}(k,u)=\frac{1-w_L(u)}{u}\frac{1}{1-j_{F,L}(k,u)} \tag{B.4}$$

In some cases the walker PDF is the solution of a differential equation containing fractional derivatives, a so-called fractional diffusion equation. Temporal fractional derivatives are expressed by the Riemann-Liouville fractional operator,

$${}_0D_t^{1-\alpha}=\frac{\partial}{\partial t}\ {}_0D_t^{-\alpha} \quad \text{for} \quad 0<\alpha<1 \tag{B.5}$$

spatial ones by the Weyl fractional operator:

$${}_{-\infty}D_x^{\mu}=\begin{Bmatrix} \dfrac{\partial}{\partial x}\ {}_{-\infty}D_x^{\mu-1} & \text{for} \quad 0<\mu\leqslant 1 \\ \dfrac{\partial^2}{\partial x^2}\ {}_{-\infty}D_x^{\mu-2} & \text{for} \quad 1<\mu\leqslant 2 \end{Bmatrix} \tag{B.6}$$

Both operators are defined via the Riemann-Liouville fractional integral [137][138][139] [140][141]:

$$ {}_{y_0}D_y^{-\beta} f(y) = \frac{1}{\Gamma(\beta)} \int_{y_0}^{y} \mathrm{d}y' \frac{f(y')}{(y-y')^{1-\beta}} \quad \text{for} \quad \mathrm{Re}(\beta)>0 \tag{B.7} $$

B.1. Normal diffusion

A jump length distribution which is narrow (i. e. whose second moment exists) and its Fourier transform:

$$ \lambda(x) = \left(4\pi\sigma^2\right)^{-1/2} \exp(-x^2/4\sigma^2) \qquad \lambda_F(k) \sim 1-\sigma^2 k^2 \tag{B.8} $$

A waiting time distribution which is narrow (i. e. whose first moment exists) and its Laplace transform:

$$ w(t) = \tau^{-1} \exp(-t/\tau) \qquad w_L(u) \sim 1-\tau u \tag{B.9} $$

The resulting walker probability density function – a Gaussian, its likewise Gaussian-shaped Fourier transform and its Fourier-Laplace transform:

$$ W(x,t) = \frac{1}{\sqrt{4\pi(\sigma^2/\tau)t}} \exp\left(-\frac{x^2}{4(\sigma^2/\tau)t}\right) \tag{B.10} $$

$$ W_F(k,t) = \exp\left[-(\sigma^2/\tau)k^2 t\right] \qquad W_{F,L}(k,u) = \frac{u^{-1}}{1+(\sigma^2/\tau)u^{-1}k^2} \tag{B.11} $$

The walker probability density function solves the diffusion equation, which is a consequence of the central limit theorem [142][126][130]:

$$ \frac{\partial}{\partial t} W(x,t) = (\sigma^2/\tau) \frac{\partial^2}{\partial x^2} W(x,t) \tag{B.12} $$

The characteristic length scale of the walker PDF grows normal diffusively:

$$ x_{\mathrm{char}}(t) \sim t^{1/2} \tag{B.13} $$

B.2. Long rests: subdiffusion

Again the narrow jump length distribution and its Fourier transform:

$$ \lambda(x) = \left(4\pi\sigma^2\right)^{-1/2} \exp(-x^2/4\sigma^2) \qquad \lambda_F(k) \sim 1-\sigma^2 k^2 \tag{B.14} $$

Now a waiting time distribution which is broad (i. e. which lacks its first moment: $0<\alpha<1$) and its Laplace transform [143][144][145]:

$$ w(t) \sim A_\alpha \tau^{-1} (t/\tau)^{-(1+\alpha)} \qquad w_L(u) \sim 1-(\tau u)^\alpha \tag{B.15} $$

The resulting walker probability density function, which is a Fox H-function [146][147] [148][149][150][151], its Fourier transform, which is a Mittag-Leffler function [152] [153], and its Fourier-Laplace transform:

$$W(x,t)=\frac{1}{\sqrt{4\pi(\sigma^2/\tau^\alpha)t^\alpha}}H_{1,2}^{2,0}\left(-\frac{x^2}{4(\sigma^2/\tau^\alpha)t^\alpha}\right)\left[\begin{matrix}1-\alpha/2 & \alpha\\ (0,1) & (1/2,1)\end{matrix}\right] \tag{B.16}$$

$$W_F(k,t)=E_\alpha\left[-(\sigma^2/\tau^\alpha)k^2t^\alpha\right] \qquad W_{F,L}(k,u)=\frac{u^{-1}}{1+(\sigma^2/\tau^\alpha)u^{-\alpha}k^2} \tag{B.17}$$

The walker probability density function solves a fractional diffusion equation containing the Riemann-Liouville operator:

$$\frac{\partial}{\partial t}W(x,t)=(\sigma^2/\tau^\alpha)\ {}_0D_t^{1-\alpha}\frac{\partial^2}{\partial x^2}W(x,t) \tag{B.18}$$

The characteristic length scale of the walker PDF grows subdiffusively:

$$x_{\mathrm{char}}(t)\sim t^{\alpha/2} \tag{B.19}$$

B.3. Long jumps: Lévy flights

A jump length distribution which is broad (i. e. which lacks its second moment: $0<\mu<2$) and its Fourier transform:

$$\lambda(x)\sim A_\mu\sigma^{-1}|x/\sigma|^{-(1+\mu)} \qquad \lambda_F(k)\sim 1-\sigma^\mu|k|^\mu \tag{B.20}$$

A narrow waiting time distribution and its Laplace transform:

$$w(t)=\tau^{-1}\exp(-t/\tau) \qquad w_L(u)\sim 1-\tau u \tag{B.21}$$

The resulting Fox-type walker probability density function [154][155], its exponential Fourier transform and its Fourier-Laplace transform:

$$W(x,t)=\frac{1}{\mu|x|}H_{2,2}^{1,1}\left(-\frac{|x|}{[(\sigma^\mu/\tau)t]^{1/\mu}}\right)\left[\begin{matrix}(1,1/\mu) & (1,1/2)\\ (1,1) & (1,1/2)\end{matrix}\right] \tag{B.22}$$

$$W_F(k,t)=\exp\left[-(\sigma^\mu/\tau)t|k|^\mu\right] \qquad W_{F,L}(k,u)=\frac{u^{-1}}{1+(\sigma^\mu/\tau)u^{-1}|k|^\mu} \tag{B.23}$$

The walker probability density function solves a fractional diffusion equation containing the Weyl operator [156][157]:

$$\frac{\partial}{\partial t}W(x,t)=(\sigma^\mu/\tau)\ {}_{-\infty}D_x^\mu W(x,t) \tag{B.24}$$

The characteristic length scale of the walker PDF grows superdiffusively:

$$x_{\mathrm{char}}(t)\sim t^{1/\mu} \tag{B.25}$$

B.4. Competition between long rests and long jumps

A broad ($0<\mu<2$) jump length distribution and its Fourier transform:

$$\lambda(x)\sim A_\mu\sigma^{-1}|x/\sigma|^{-(1+\mu)} \qquad \lambda_F(k)\sim 1-\sigma^\mu|k|^\mu \tag{B.26}$$

A broad ($0<\alpha<1$) waiting time distribution and its Laplace transform:

$$w(t) \sim A_\alpha \tau^{-1} (t/\tau)^{-(1+\alpha)} \qquad w_L(u) \sim 1-(\tau u)^\alpha \tag{B.27}$$

The resulting Fourier-Laplace transform of the walker probability density function:

$$W_{F,L}(k,u) = \frac{u^{-1}}{1+(\sigma^\mu/\tau^\alpha)u^{-\alpha}|k|^\mu} \tag{B.28}$$

The walker probability density function solves a fractional diffusion equation containing both the Riemann-Liouville operator and the Weyl operator:

$$\frac{\partial}{\partial t} W(x,t) = (\sigma^\mu/\tau^\alpha) \; {}_0D_t^{1-\alpha} \; {}_{-\infty}D_x^\mu W(x,t) \tag{B.29}$$

The characteristic length scale of the walker PDF can grow subdiffusively ($2\alpha<\mu$), quasi-diffusively ($2\alpha=\mu$) or superdiffusively ($2\alpha>\mu$):

$$x_{\text{char}}(t) \sim t^{\alpha/\mu} \tag{B.30}$$

B.5. Competition in inter-correlated CTRWs

In the idealized cases seen up to now (normal CTRWs), the joint jump distribution factorized:

$$j(x,t) = \lambda(x)\, w(t) \tag{B.31}$$

However, if waiting time correlates with jump length (inter-correlated CTRWs), as in the present work, where the walker explores a frozen intersegment link pattern, the competition between long rests and long jumps cannot be described by the use of fractional diffusion equations. In this quenched case the joint jump distribution reads:

$$j(x,t) \propto x^{-(1+\mu)} \delta(t-x^k) \tag{B.32}$$

Integration yields the jump length distribution, which looses its first moment for $\mu<1$,

$$\lambda(x) \propto x^{-(1+\mu)} \tag{B.33}$$

and the waiting time distribution, which is broad for $\mu<k$:

$$w(t) \propto t^{-(1+\mu/k)} \tag{B.34}$$

Since random walks on frozen intersegment link patterns are not only inter-correlated (i. e. jump length and waiting time within one jump are not independent) but also geometrically correlated (resulting in successive jump lengths which are not independent), it is clear that the fractional diffusion approach must fail. Nevertheless, there is again a competition between long rests and long jumps which exhibits tripartite transport behavior, however in a manner very different from the uncorrelated competition in normal CTRWs: subdiffusion ($\mu+k>2$), quasi-diffusion ($\mu+k=2$) or superdiffusion ($\mu+k<2$):

$$t_{\text{char}}(x) \propto x^{\mu+k} \tag{B.35}$$

Bibliography

[1] Francis Crick, Nature **227**, 561-563 (1970)

[2] George Rice, http://serc.carleton.edu/microbelife/research_methods/genomics/index.html

[3] Dr. Ebert-May, https://www.msu.edu/course/isb/202/ebertmay/images/HIV%20virus.png

[4] George Rice, http://serc.carleton.edu/microbelife/research_methods/genomics/ribosome.html

[5] H. Kontos, S. Napthine, and I. Brierley, Mol. Cell. Biol. **21**, 8657–8670 (2001)

[6] C. Tu, T. H. Tzeng, and J. A. Bruenn, Proc. Natl. Acad. Sci. USA **89**, 8636-8640 (1992)

[7] P. Somogyi, A. J. Jenner, I. Brierley, and S. C. Inglis, Mol. Cell. Biol. **13**, 6931-6940 (1993)

[8] J. W. Harger, A. Meskauskas, and J. D. Dinman, Trends Biochem. Sci. **27**, 448-454 (2002)

[9] David P. Giedroc, Carla A. Theimer, and Paul L. Nixon, J. Mol. Biol. **298**, 167-185 (2000)

[10] Thomas M. Hansen, S. Nader S. Reihani, Lene B. Oddershede, and Michael A. Sørensen, Proc. Natl. Acad. Sci. USA **104**, 5830-5835 (2007)

[11] Jonathan D. Dinman, Yeast **11**, 1115–1127 (1995)

[12] Ewan P. Plant and Jonathan D. Dinman, Nucl. Acids Res. **33**, 1825-1833 (2005)

[13] O. Namy, S. J. Moran, D. I. Stuart, R. J. Gilbert, and I. Brierley, Nature **441**, 244–247 (2006)

[14] Ewan P. Plant, Kristi L. Muldoon Jacobs, Jason W. Harger, Arturas Meskauskas, Jonathan L. Jacobs, Jennifer L. Baxter, Alexey N. Petrov, and Jonathan D. Dinman, RNA **9**, 168-174 (2003)

[15] Jihong Ren, Baharak Rastegari, Anne Condon, and Holger H. Hoos, RNA **11**, 1494-1504 (2005)

[16] Jan Pieter Abrahams, Mirjam Van den Berg, Eke Van Batenburg, and Cornelis Pleij, Nucl. Acids Res. **18**, 3035 (1990)

[17] Elena Rivas and Sean R. Eddy, J. Mol. Biol. **285**, 2053-2068 (1999)
[18] Ye Ding and Charles E. Lawrence, Nucl. Acids Res. **31**, 7280-7301 (2003)
[19] A.P. Gultyaev, Nucl. Acids Res. **19**, 2489-2494 (1991)
[20] R. Bundschuh and T. Hwa, Phys. Rev. Lett. **83**, 1479-1482 (1999)
[21] R. Bundschuh and T. Hwa, Phys. Rev. E **65**, 031903 (2001)
[22] R. Bundschuh and T. Hwa, Europhys. Lett. **59**, 903–909 (2002)
[23] Ulrich Gerland, Ralf Bundschuh, and Terence Hwa, Biophys. J. **81**, 1324–1332 (2001)
[24] Ulrich Gerland, Ralf Bundschuh, and Terence Hwa, Biophys. J. **84**, 2831–2840 (2003)
[25] Ulrich Gerland, Ralf Bundschuh, and Terence Hwa, Phys. Biol. **1**, 19–26 (2004)
[26] Ralf Bundschuh and Ulrich Gerland, Phys. Rev. Lett. **95**, 208104 (2005)
[27] J. S. McCaskill, Biopol. **29**, 1105-1119 (1990)
[28] Nan-Jie Deng and Piotr Cieplak, J. Chem. Theory Comput. **3**, 1435-1450 (2007)
[29] Guanghong Zuo, Wenfei Li, Jian Zhang, Jin Wang, and Wei Wang, J. Phys. Chem. B **114**, 5835–5839 (2010)
[30] A. Xayaphoummine, T. Bucher, and H. Isambert, Nucl. Acids Res. **33**, W605-W610 (2005)
[31] J.A. Abels, F. Moreno-Herrero, T. van der Heijden, C. Dekker and N.H. Dekker, Biophys. J. **88**, 2737 - 2744 (2005)
[32] Zev Bryant, Michael D. Stone, Jeff Gore, Steven B. Smith, Nicholas R. Cozzarelli, and Carlos Bustamante, Nature **424**, 338-341 (2003)
[33] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Clarendon Press, Oxford, 1986)
[34] Patrick S. Doyle and Patrick T. Underhill, *Brownian dynamics simulations of polymers and soft matter* (Springer Netherlands, 2005)
[35] Amit Meller, Lucas Nivon, and Daniel Branton, Phys. Rev. Lett. **86**, 3435-3438 (2000)
[36] Alexis F. Sauer-Budge, Jacqueline A. Nyamwanda, David K. Lubensky, and Daniel Branton, Phys. Rev. Lett. **90**, 238101 (2003)
[37] Jérôme Mathé, Hasina Visram, Virgile Viasnoff, Yitzhak Rabin, and Amit Meller, Biophys. J. **87**, 3205–3212 (2004)
[38] Olga K. Dudko, Jérôme Mathé, Attila Szabo, Amit Meller, and Gerhard Hummer, Biophys. J. **92**, 4188–4195 (2007)
[39] Ben McNally, Meni Wanunu, and Amit Meller, Nanolett. **8**, 3418-3422 (2009)
[40] Gary M. Skinner, Michiel van den Hout, Onno Broekmans, Cees Dekker, and Nynke H. Dekker, Nanolett. **9**, 2953-2960 (2009)
[41] F. Jacob and J. Monod, J. Mol. Biol. **3**, 318–356 (1961)
[42] S. Bourgeois, M. Cohn and L. E. Orgel, J. Mol. Biol. **14**, 300–302 (1965)

[43] Isabelle Bonnet, Andreas Biebricher, Pierre-Louis Porté, Claude Loverdo,Olivier Bénichou, Raphael Voituriez, Christophe Escudé, Wolfgang Wende, Alfred Pingoud and Pierre Desbiolles, Nucl. Acids Res. **36**, 4118–4127 (2008)
[44] B. van den Broek, M. A. Lomholt, S.-M. J. Kalisch, R. Metzler, and G. J. L. Wuite, Proc. Natl. Acad. Sci. USA **105**, 15738-15742 (2008)
[45] G. D. Stormo and D. S. Fields, Trends Biochem. Sci. **23**, 109 (1998)
[46] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. USA **99**, 12015 (2002)
[47] R. F. Bruinsma, Physica A **313**, 211 (2002)
[48] D. Brockmann and T. Geisel, Phys. Rev. Lett. **91**, 048303 (2003)
[49] M. Slutsky and L. A. Mirny, Biophys. J. **87**, 4021 (2004)
[50] M. Coppey, O. Benichou, R. Voituriez, and M. Moreau, Biophys. J. **87**, 1640 (2004)
[51] M. A. Lomholt, T. Ambjörnsson, and R. Metzler, Phys. Rev. Lett. **95**, 260603 (2005)
[52] T. Hu, A. Y. Grosberg, and B. I. Shklovskii, Biophys. J. **90**, 2731 (2006)
[53] T. Hu and B. I. Shklovskii, Phys. Rev. E **76**, 051909 (2007)
[54] Z. Wunderlich and L. A. Mirny, Nucleic Acids Res. **36**, 3570 (2008)
[55] Michael Sheinman and Yariv Kafri, Phys. Biol. **6**, 016003 (2009)
[56] S. E. Halford and J. F. Marko, Nucleic Acids Res. **32**, 3040 (2004)
[57] Y. Harada, T. Funatsu, K. Murakami, Y. Nonoyama, A. Ishihama, and T. Yanagida, Biophys. J. **76**, 709–715 (1999)
[58] Y. M. Wang, R. H. Austin, and E. C. Cox, Phys. Rev. Lett. **97**, 048302 (2006)
[59] A. Graneli, C. C. Yeykal, R. B. Robertson, and E. C. Greene, Proc. Natl. Acad. Sci. USA **103**, 1221–1226 (2006)
[60] P. C. Blainey, A. M. van Oijen, A. Banerjee, G. L. Verdine, and X. S. Xie, Proc. Natl. Acad. Sci. USA **103**, 5752–5757 (2006)
[61] J. Elf, G.-W. Li, and X. S. Xie, Science **316**, 1191 (2007)
[62] A. D. Riggs, M. S. Bourgeois, and M. Cohn, J. Mol. Biol. **53**, 401 (1970)
[63] G. Adam and M. Delbrück, *Structural Chemistry and Molecular Biology* (W.H. Freeman, 1968)
[64] P. H. Richter and M. Eigen, Biophys. Chem. **2**, 255 (1974)
[65] O. G. Berg and M. Ehrenberg, Biophys. Chem. **15**, 41–51 (1982)
[66] O. G. Berg, R. B. Winter, and P. H. von Hippel, Biochemistry **20**, 6929 (1981)
[67] R. B. Winter and P. H. von Hippel, Biochemistry **20**, 6948 (1981)
[68] R. B. Winter, O. G. Berg, and P. H. von Hippel, Biochemistry **20**, 6961 (1981)
[69] P. H. von Hippel and O. G. Berg, J. Biol. Chem. **264**, 675–678 (1989)
[70] O. Benichou, M. Coppey, M. Moreau, P.-H. Suet, and R. Voituriez, Phys. Rev. Lett. **94**, 198101 (2005)
[71] Ying Kao-Huang, Arnold Revzin, Andrew P. Butler, Pamela O'Conner, Daniel W. Noble, and Peter H. Von Hippel, Proc. Natl. Acad. Sci. USA **74**, 4228-4232 (1977)
[72] R. Metzler and J. Klafter, Phys. Rep. **339**, 1 (2000)

[73] I. M. Sokolov, J. Mai, and A. Blumen, Phys. Rev. Lett. **79**, 857 (1997)
[74] S. Arnott, D. W. L. Hukins, and S. D. Dover, Biochem. Biophys. Res. Commun. **48**, 1392-1399 (1972)
[75] S. Arnott, D. W. L. Hukins, S. D. Dover, W. Fuller, and A. R. Hodgson, J. Mol. Biol. **81**, 107-122 (1973)
[76] Stephen R. Holbrook and Sung-Hou Kim, Biopolymers **44**, 3-21 (1997)
[77] Changbong Hyeon and D. Thirumalai, Biophys. J. **90**, 3410-3427 (2006)
[78] Thomas E. Ouldridge, Iain G. Johnston, Ard A. Louis, and Jonathan P. K. Doye, J. Chem. Phys. **130**, 065101 (2009)
[79] A. C. Brańka and D. M. Heyes, Phys. Rev. E **58**, 2611–2615 (1998)
[80] Hiroshi Taketomi, Yuzo Ueda, Nobuhiro Gō, Int. J. Pept. Protein Res. **7**, 445–459 (1975)
[81] Ignacio Tinoco Jr, and Carlos Bustamante, J. Mol. Biol. **293**, 271-281 (1999)
[82] Noel L. Goddard, Grégoire Bonnet, Oleg Krichevsky, and Albert Libchaber, Phys. Rev. Lett. **85**, 2400–2403 (2000)
[83] Grégoire Bonnet, Oleg Krichevsky, and Albert Libchaber, Proc. Natl. Acad. Sci. USA **95**, 8602-8606 (1998)
[84] S. B. Smith, L. Finzi, and C. Bustamante, Science **258**, 1122 - 1126 (1992)
[85] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, Science **265**, 1599 - 1600 (1994)
[86] Jean-Francois Allemand, David Bensimon, and Vincent Croquette, Curr. Op. Struct. Biol. **13**, 266-274 (2003)
[87] Philippe Cluzel, Anne Lebrun, Christoph Heller, Richard Lavery, Jean-Louis Viovy, Didier Chatenay, François Caron, Science **271**, 792 - 794 (1996)
[88] Anne Lebrun, and Richard Lavery, Nucl. Acids Res. **24**, 2260-2267 (1996)
[89] Joost van Mameren, Peter Gross, Geraldine Farge, Pleuni Hooijman, Mauro Modesti, Maria Falkenberg, Gijs J. L. Wuite, and Erwin J. G. Peterman, Proc. Natl. Acad. Sci. USA **106**, 18231-18236 (2009)
[90] Ioulia Rouzina and Victor A. Bloomfield, Biophys. J. **80**, 882-893 (2001)
[91] Ioulia Rouzina and Victor A. Bloomfield, Biophys. J. **80**, 894-900 (2001)
[92] M. Rief, H. Clausen-Schaumann, and H. E. Gaub, Nature Struct. Biol. **6**, 346-349 (1999)
[93] Hongxia Fu, Hu Chen, John F. Marko, and Jie Yan, Nucl. Acids Res. **Advance Access**, 1-7 (2010)
[94] A-DNA, http://en.wikipedia.org/wiki/A-DNA
[95] P. Mangeol, D. Côte, T. Bizebard, O. Legrand and U. Bockelmann, Eur. Phys. J. E **19**, 311 - 317 (2006)
[96] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner, J. Mol. Biol. **288**, 911-940 (1999)

[97] Steven B. Smith, Yujia Cui, and Carlos Bustamante, Science **271**, 795-799 (1996)
[98] K. Rietveld, R. van Poelgeest, C. W. Pleij, J. H. van Boom, and L. Bosch, Nucl. Acids Res. **10**, 1929-1946 (1982)
[99] K. Rietveld, C. W. A. Pleij, and L. Bosch, EMBO J. **2**, 1079-1085 (1983)
[100] D. S. McPheeters, G. D. Stormo, and L. Gold, J. Mol. Biol. **201**, 517-535 (1988)
[101] T. Jacks, K. Townsley, H. E. Varmus, and J. Majors, Proc. Natl. Acad. Sci. USA **84**, 4298-4302 (1987)
[102] T. Jacks, H. D. Madhani, F. R. Masiarz, and H. E. Varmus, Cell **55**, 447-458 (1988a)
[103] T. Jacks, M. D. Power, F. R. Masiarz, P. A. Lucius, P. J. Barr, and H. E. Varmus, Nature **331**, 280-283 (1988b)
[104] N. M. Wills, R. F. Gesteland, and J. F. Atkins, Proc. Natl. Acad. Sci. USA **88**, 6991-6995 (1991)
[105] Y.-X. Feng, H. Yuan, A. Rein, and J. G. Levin, J. Virol. **66**, 5127-5132 (1992)
[106] T. Powers, and H. F. Noller, EMBO J. **10**, 2203-2214 (1991)
[107] N. Nameki, B. Felden, J. F. Atkins, R. F. Gesteland, H. Himeno, and A. Muto, J. Mol. Biol. **286**, 733-744 (1999)
[108] E. S. Hass, J. W. Brown, C. Pitulle, and N. R. Pace, Proc. Natl. Acad. Sci. USA **91**, 2527-2531 (1994)
[109] E. Jabri, S. Aigner, and T. R. Cech, Biochemistry **36**, 16345-16354 (1997)
[110] A. R. Ferré-D'Amaré, K. Zhou, and J. A. Doudna, Nature **395**, 567-574 (1998)
[111] D. Gilley and E. H. Blackburn, Proc. Natl. Acad. Sci. USA **96**, 6621-6625 (1999)
[112] E. B. ten Dam, K. Pleij, and D. Draper, Biochemistry **31**, 11665-11676 (1992)
[113] C. W. A. Pleij, Curr. Opin. Struct. Biol. **4**, 337–344 (1994)
[114] I. Brierley, N. J. Rolley, A. J. Jenner, and S. C. Inglis, J. Mol. Biol. **220**, 889-902 (1991)
[115] M. Chamorro, N. Parkin, and H. E. Varmus, Proc. Natl. Acad. Sci. USA **89**, 713-717 (1992)
[116] E. ten Dam, P. W. G. Verlaan, and C. W. A. Pleij, RNA **1**, 146-154 (1995)
[117] C. W. A. Pleij, K. Rietveld, and L. Bosch, Nucl. Acids Res. **13**, 1717-1731 (1985)
[118] L. X. Shen and I. Tinoco Jr, J. Mol. Biol. **247**, 963-978 (1995)
[119] D. E. Draper, Curr. Opin. Cell Biol. **2**, 1099-1103 (1990)
[120] Sawsan Napthine, Jan Liphardt, Alison Bloys, Samantha Routledge, and Ian Brierley, J. Mol. Biol. **288**, 305-320 (1999)
[121] X. Chen, M. Chamorro, S. I. Lee, L. X. Shen, J. V. Hines, I. Tinoco Jr, and H. E. Varmus, EMBO J. **14**, 842–852 (1995)
[122] Lisa Green, Chul-Hyun Kim, Carlos Bustamante, and Ignacio Tinoco Jr, J. Mol. Biol. **375**, 511-528 (2008)
[123] P. H. Verdier and W. H. Stockmayer, J. Chem. Phys. **36**, 227 (1962)

[124] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2000)
[125] B. D. Hughes, *Random Walks and Random Environments, vol. 1: Random Walks* (Oxford University Press, 1995)
[126] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990)
[127] B. J. West, W. Deering, Phys. Rep. **246**, 1 (1994)
[128] P. Lévy, *Calcul des Probabilités* (Gauthier-Villars, 1925)
[129] P. Lévy, *Théorie de l'addition des variables Aléatoires* (Gauthier-Villars, 1954)
[130] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Random Variables* (Addison-Wesley, 1954)
[131] E. Montroll and G. H. Weiss, J. Math. Phys. **6**, 167 (1965)
[132] H. Scher and E. Montroll, Phys. Rev. B **12**, 2455 (1975)
[133] Sidney Redner, *A Guide to First-Passage Processes* (Cambridge University Press, 2001)
[134] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences* (Springer, 2004)
[135] R. Metzler and J. Klafter, Phys. Rep. **339**, 1 (2000)
[136] J. Klafter, A. Blumen, M. F. Shlesinger, Phys. Rev. A **35**, 3081 (1987)
[137] S. G. Samko, A. A. Kilbas, O. I. Marichev, *Fractional Integrals and Derivatives - Theory and Applications* (Gordonand Breach, 1993)
[138] K. B. Oldham, J. Spanier, *The Fractional Calculus* (Academic Press, 1974)
[139] B. Ross (Ed.), *Fractional Calculus and its Applications, Lecture Notes in Mathematics, Vol. 457* (Springer, 1975)
[140] K. S. Miller, B. Ross, *An Introduction to the Fractional Calculus and Fractional Differential Equations* (Wiley, 1993)
[141] R. Hilfer (Ed.), *Applications of Fractional Calculus in Physics* (World Scientific, 1999)
[142] P. Lévy, *Processes stochastiques et mouvement Brownien* (Gauthier-Villars, 1965)
[143] E. W. Montroll, J. SIAM **4**, 241 (1956)
[144] E. W. Montroll, G. H. Weiss, J. Math. Phys. **10**, 753 (1969)
[145] E. W. Montroll, H. Scher, J. Stat. Phys. **9**, 101 (1973)
[146] H. M. Srivastava, K. C. Gupta, S. P. Goyal, *The H-Functions of One and Two Variables with Applications* (South Asian Publishers, 1982)
[147] C. Fox, Trans. Amer. Math. Soc. **98**, 395 (1961)
[148] B. L. J. Braaksma, Compos. Math. **15**, 239 (1964)
[149] A. M. Mathai, R. K. Saxena, *The H-Function with Applications in Statistics and Other Disciplines* (Wiley Eastern, 1978)
[150] H. M. Srivastava, B. R. K. Kashyap, *Special Functions in Queuing Theory and Related Stochastic Processes* (Academic Press, 1982)

[151] S. Wolfram, *The Mathematica Book, 4th Edition* (Cambridge University Press, 1999)

[152] G. Mittag-Leffler, Acta Math. **29**, 101 (1905)

[153] A. Erdélyi (Ed.), *Tables of Integral Transforms, Bateman Manuscript Project, Vol. I* (McGraw-Hill, 1954)

[154] B. J. West, P. Grigolini, R. Metzler, T. F. Nonnenmacher, Phys. Rev. E **55**, 99 (1997)

[155] S. Jespersen, R. Metzler, H. C. Fogedby, Phys. Rev. E **59**, 2736 (1999)

[156] A. Compte, Phys. Rev. E **53**, 4191 (1996)

[157] R. Metzler, J. Klafter, I. Sokolov, Phys. Rev. E **58**, 1621 (1998)

Danksagung

Mein Dank gilt besonders meinem Betreuer Ulrich Gerland, den ich als Wissenschaftler und Mensch sehr zu schätzen gelernt habe, für die offenen, überaus konstruktiven und inspirierenden physikalischen Diskussionen, für seinen unerschöpflichen Ideenreichtum und für seine Fähigkeit stets neue interessante Sichtweisen zu vermitteln und mich für immer neue Herausforderungen zu begeistern.

Dem Lehrstuhlinhaber Erwin Frey bin ich in höchstem Maße dankbar dafür, dass ich an seinem Lehrstuhl, einem motivierenden Ort der in wissenschaftlicher hinsicht unbegrenzten Möglichkeiten, meine Doktorarbeit anfertigen durfte.

Allen Lehrstuhlmitgliedern danke ich für die hervorragende Arbeitsatmosphäre, insbesondere für unsere wertvollen Gespräche, die immerwährende Diskussionsbereitschaft bei fachlichen Problemen aller Art und für die kollegiale Zusammenarbeit in der Studentenbetreuung. Insbesondere die fachlichen Diskussionen mit Richard waren immer sehr hilfreich.

Dirk Brockmann und Ralf Metzler, die über Diskussionen mein Interesse an den mathematischen Feinheiten von Transportprozessen in Physik und Biologie weckten, bin ich dankbar für ihre Expertise auf diesem Gebiet.

Ebenfalls möchte ich mich bei unserer Lehrstuhlsekretärin Caroline Lesperance bedanken, unser aller geschätzten Ansprechpartnerin in den Wirrungen der Bürokratie, die weiß, was die Lehrstuhlwelt im Innersten zusammenhält. Dank schulde ich auch dem Systemadministrator Ralph Simmler und Frederik Wagner für ihren kompetenten und unermüdlichen Einsatz für unsere IT-Infrastruktur, die meine umfangreichen Simulationsrechnungen überhaupt erst möglich machte. Schließlich danke ich dem Sonderforschungsbereich 486 "Manipulation von Materie auf der Nanometerskala" für die finanzielle Unterstützung meiner Promotion.

Zu größtem Dank verpflichtet bin ich meinen Eltern, Florentine und ihrer Familie dafür, dass sie immer an mich geglaubt und mich unterstützt haben.

Target search on a dynamic DNA molecule

Thomas Schötz,[1] Richard A. Neher,[2] and Ulrich Gerland[1,*]

[1] *Arnold Sommerfeld Center for Theoretical Physics and Center for Nano Science, University of Munich, Theresienstr. 37, D-80333 München, Germany*

[2] *Max-Planck-Institute for Developmental Biology, Spemannstr. 35, 72076 Tübingen, Germany*

(Dated: November 9, 2010)

DNA-binding proteins find their targets by a combination of 3D diffusion in solution, 1D diffusion along the DNA, and "jumping" between sites distant along the DNA contour but close in 3D space. Depending on the timescale of the protein jumps relative to the equilibration time of the DNA contour, the search dynamics is qualitatively affected by temporal and spatial correlations in the jumps. We characterize these correlations, their effect, and their breakdown using a target search model with explicit DNA dynamics. In particular, we rationalize the previously described paradoxical quasi-diffusive transport regime as a subtle compensation between long jumps and long trapping times of the protein in "islands" within the random configurations of polymers in solution.

PACS numbers: 123

The quantitative characteristics of proteins searching for their specific target sites on long DNA molecules has become a paradigmatic question of biological physics [1–6]. The question is of considerable biological interest, since search processes of this type are key steps in cellular functions. For instance, in signal transduction, a protein belonging to the large class of transcription factors conveys an external signal and triggers the appropriate genetic response by binding to specific operator sites on the genomic DNA. Similarly, restriction enzymes, used by bacteria to fight invading viruses, search for cleavage sites marked by specific DNA sequences. It is generally assumed that the target search mechanism has been optimized by evolution, due to a significant selective pressure for fast signaling and rapid responses in competitive environments. From the physics perspective, the protein-DNA target search is a complex but tractable stochastic process that combines basic aspects of Brownian motion, polymer physics, and information theory [7–17]. Experimentally, the search process can be probed on the single-molecule level *in vitro* [18], and even *in vivo* [19].

Early *in vitro* experiments [2] indicated that the association rate of *lac* repressor to its specific binding site embedded in short pieces of DNA is faster than the diffusion limit, $k_a = 4\pi Db$, for a direct binding reaction (D denotes the diffusion constant of the protein in solution and b the reaction radius for protein-DNA binding). Inspired by Adam and Delbrück's idea that reduction of dimensionality is a generic way to enhance reaction rates [20], Richter and Eigen [3] interpreted these experiments with a two-step mechanism where three-dimensional (3D) diffusion and non-specific association to DNA (idealized as a cylinder) is followed by one-dimensional (1D) diffusive sliding into the target site. In a seminal series of papers [4–6], Berg, Winter, and von Hippel then established much of what is known today about the protein-DNA search kinetics. Experimentally, they varied the strength of non-specific binding via the ion concentration, clearly identifying an optimum where association to the specific target site is fastest. Theoretically, their analysis within the framework of physical chemistry explained the observed dependence.

Recent work reformulated the problem using the theory of stochastic processes [10–13, 17], which led to a number of further insights. For instance, the existence of the optimum reflects a generic tradeoff in search processes for hidden targets [21]: A stochastic local search is exhaustive but redundant; interrupting the search by phases of rapid movement to new territory is a time investment that pays off by significantly reducing the redundancy. The optimal fraction of time spent in each of the two "modes" depends on the statistical characteristics of the search mechanism. The simplest scenario, where proteins slide diffusively along the DNA, dissociate spontaneously, and then randomly reattach at an uncorrelated position, leads to an optimal non-specific binding strength such that, on average, only half of the proteins are bound somewhere on the DNA and the other half is free in solution [6]. Physically, this optimality condition is best understood [11] in terms of the typical dwell times of a protein in the sliding mode, τ_s, and in the dissociated state, τ_d. The latter should be regarded as a fixed parameter, set by the size and composition of the cells, whereas τ_s can easily be adapted by molecular evolution of the DNA-binding domain of the protein (to adjust the non-specific affinity). If $\tau_s < \tau_d$, the protein spends too

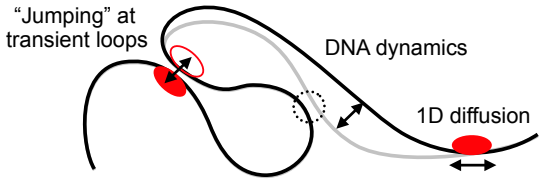


FIG. 1: Illustration of the target search by sliding (1D diffusion) and jumping on a dynamic polymer.

little time searching, while if $\tau_s > \tau_d$, the search is too redundant; the mean search time is minimal when they are equal.

However, in bacterial cells, well studied transcription factors are bound to the genome $\gtrsim 90\%$ of the time [7]. This fact has drawn attention to the 'intersegment transfer' [4, 13, 15, 17] of a protein within the same DNA molecule, between sites close in space but distant along the contour. Potentially, this process can destroy the redundancy of the 1D search without the price of interrupting it by long excursions into the solvent. The term intersegment transfer was introduced for proteins with two DNA-binding domains and refers to a process during which the protein never detaches from the DNA molecule; a similar transfer but with a brief period in an unbound state is referred to as 'hopping' [4]. In both cases, the essential difference to the uncorrelated random reattachment discussed above is the correlated nature of the process: Transfer does not occur with equal probability to every site on the DNA, but to "linked" sites. Here, we simply refer to both processes as 'jumping'.

As illustrated in Fig. 1, the interplay of protein sliding and jumping leads to intricate search dynamics. Notably, the physics of this interplay depends crucially on the dynamics of the DNA, which has so far never been considered explicitly. An elegant analytical study [13] considered the effect of long-range jumps on the level of the fractional Fokker-Planck equation [22]. However, by design this framework assumes that consecutive jumps are uncorrelated, i.e. that the DNA configuration completely randomizes between two jumps of the protein. In contrast, a numerical study of sliding and jumping on a random but frozen contour [23] showed that the correlations between jumps drastically alter the dynamics, leading to "paradoxical" quasi-diffusive behavior instead of super-diffusion along the contour. These findings, and the fact that the dynamics of real DNA is neither frozen nor annealed over the relevant range of μs to s timescales [6], calls for an analysis of target search on a dynamic DNA.

Here, we study the crossover between the two regimes and clarify the mechanism whereby correlated jumps create the paradoxical behavior. We take the dynamics of the polymer and the TF into account and explicitly show the crossover from the search on annealed to the search on quenched polymer configurations. We study the search dynamics of a TF on a frozen polymer chain in detail and show that the quasi-diffusive behavior is due to a competition between long sojourn times in some stretches of the polymer, interspersed by long jumps. The long waiting times are caused by a hierarchical organization of possible jumps, causing a maze from which the TF takes long to escape.

Model.— To make the problem tractable, we describe the DNA contour as a path of L segments on a simple cubic lattice, and generate its conformational dynamics with a kinetic Monte Carlo scheme. The scheme is based on a generalized Verdier-Stockmayer move set [24] with moves for kinks, chain ends, and crankshafts, as illustrated in Fig. S1. These moves, carried out at the rate k_{D}, implement Rouse dynamics on a lattice for an ideal chain (no self-avoidance). We describe a protein as a point particle on the lattice, which diffuses along the DNA contour at rate k_{p}. If another DNA segment passes through the same point, the protein can randomly jump to it (at the same rate k_{p}, for simplicity). As initial condition, we use a random DNA configuration with the protein on the central segment. Of course, the configuration of the DNA inside a bacterial cell is far from random, due to genome packaging and confinement, but a random configuration is an interesting starting point for exploration of the physical principles, and mimics the situation of *in vitro* experiments.

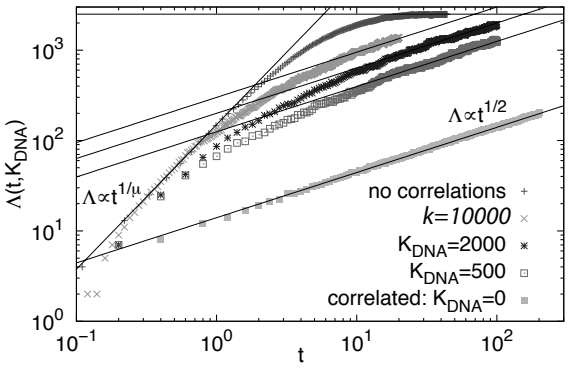


FIG. 2: The width Λ of the protein distribution as a function of time for different kinetic ratios $k = k_{\mathrm{D}}/k_{\mathrm{p}}$ (chain length $L = 5000$; for each ratio, $10^3 - 10^4$ simulation runs were performed, each with a different initial DNA configuration). In the annealed limit, the DNA dynamics is not simulated explicitly, but instead after each jump of the protein, a new link set is randomly drawn. For finite k, a crossover from super-diffusive dynamics to quasi-diffusive dynamics is clearly visible.

Simulation results.— To characterize how a protein explores the search space, we study the time evolution of its probability distribution $P(s,t)$ along the DNA contour ($0 \le s \le L$). Fig. 2 plots the width $\Lambda(t)$, defined as the interquartile range [23] of $P(s,t)$, for different kinetic ratios $k = k_{\mathrm{D}}/k_{\mathrm{p}}$. Here, $P(s,t)$ is obtained by averaging over many trajectories and initial DNA configurations with $L = 5000$, see caption. In the 'quenched limit' $k \to 0$ (squares), the protein moves on a frozen contour, and the width grows quasi-diffusively with time, $\Lambda \sim t^{1/2}$, despite the long-range jumps along the contour and a heavy tail of the distribution $P(s,t)$ at a fixed time [23]. In the opposite 'annealed limit' $k \to \infty$ (crosses), the distribution initially spreads super-diffusively along the contour, $\Lambda \sim t^{\alpha}$ (here: $\alpha \approx 1.7$). The width saturates at $\Lambda \to L/2$ as the protein explores the entire DNA. In the regime of intermediate k, which is relevant in most

experimental situations, $\Lambda(t)$ displays a crossover from super- to quasi-diffusive scaling. The curves for different k show that the crossover timescale τ_c increases with k.

Annealed limit.— For large k, the connectivity of the DNA meshwork on which the protein moves changes rapidly, such that successive protein jumps are uncorrelated, since they occur on a different set of links. It is then sufficient to describe the dynamics by the average jump probability $P(s, s')$ from site s to site s', which is physically determined by the DNA looping probability. For a random ideal chain, this probability decays as $|s - s'|^{-3/2}$ for large loops, before it is cut off by the finite DNA length. When successive jump lengths are independently drawn from this distribution, the typical distance Λ from the initial position is dominated by the largest jump, which grows with the number of jumps (and the time t) as $\Lambda(t) \sim t^2$ [25]. Indeed, our numerical exponent α defined above approaches 2 at large L (data not shown). What is the effect of this super-diffusive exploration of the DNA contour on the target search? Clearly, without a guiding "funnel", no search process can be faster than linear exploration, i.e. faster than linear spreading along the DNA leads to "sloppy search" [13] where typically patches distributed over the entire contour are explored before the target is located. This implies that in the annealed limit, the jumping process efficiently destroys the redundancy of simple 1D diffusion, even without interrupting the search by phases of 3D diffusion.

Crossover timescale.— The crossover between super-diffusive and quasi-diffusive dynamics is associated with a timescale that is characteristic for the interplay of the DNA dynamics with that of the protein. A simple scaling argument captures much of the physics of this interplay: The Rouse dynamics equilibrates a DNA contour segment of length ℓ on a time scale $\tau \sim \ell^2$. Within a time τ after a protein docks onto the DNA and starts exploring, it will typically visit a DNA stretch of size $\Lambda(\tau)$. During this time, a DNA segment of size $\ell \sim (k_D\tau)^{1/2}$ equilibrates. Superdiffusive protein dynamics results as long as $\Lambda(\tau) < \ell$, however the fast growing $\Lambda(t) \sim (k_p t)^\alpha$ quickly outruns the "equilibration blob", and the passing point marks the crossover to the quasi-diffusive regime. With $\alpha = 2$, this crossover timescale t_c then depends on the kinetic ratio k as $k_p t_c \sim k^{1/3}$. Our simulations cannot explore a wide range of k values due to computational cost, and hence do not allow a precise numerical determination of this scaling. However, the scaling exponent that best describes our limited data deviates only by 0.08 from the expected value 1/3 (see Fig. S2).

Quenched limit.— Ironically, the seemingly simple, quasi-diffusive scaling at $k = 0$ is theoretically the most challenging phenomenon. When first reported [23], it was described as arising from geometrical correlations. However, the question remains what the nature of these correlations is and how they render the long-range jumping process quasi-diffusive. To address this question, our first step is to distinguish between two types of correlations, which we refer to as *temporal* and *spatial*. Neither type can be included in a fractional Fokker-Planck description. Temporal correlations arise given any static set of links, where proteins can jump back and forth between distant parts of the polymer by using the same link multiple times. Additionally, the positions of different links are spatially correlated, since an existing link strongly enhances the probability to find another link nearby (e.g. a loop in the DNA favors further contacts within the loop).

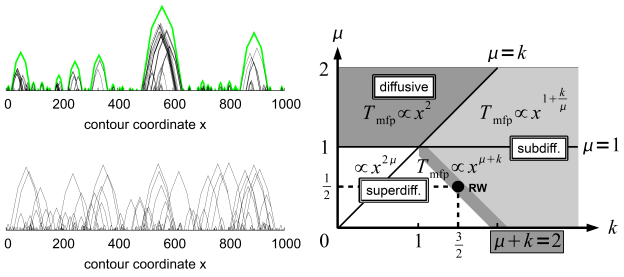


FIG. 3: Link diagrams originating from random DNA conformations are separable into islands (A, top). Transport on such link diagrams is quasi-diffusive due to strong geometrical correlations. Destroying all correlations by reshuffling links destroys islands (A, bottom). Dynamical phase diagram (B) of our transport model, which explains the quasi-diffusive protein dynamics by a cancellation of the effect of traps and long-range jumps.

To clearly separate the effect of temporal and spatial correlations, we can destroy the latter by choosing a new random starting point for each link while conserving its arc length $|s - s'|$. The protein dynamics on such reshuffled link sets is super-diffusive as revealed by simulations shown in Fig. S3. Hence temporal correlations alone are not sufficient to cause the quasi-diffusive behavior. A simple argument makes this observation plausible: If the region visited by the protein grows super-diffusively as $\Lambda(t) \sim t^2$, the protein visits only a fraction $\sim 1/t$ of the sites within Λ. Since it sees each site $\mathcal{O}(1)$ times, it mostly uses novel links and temporal correlations become unimportant.

Islands.— A striking consequence of the spatial correlations is revealed in Fig. 3A, where all links in a typical DNA configuration are depicted as arcs. The arcs cluster into "islands" with many internal links but no links between islands. These islands disappear when the same links are randomly placed on the DNA. Intuitively, it is clear that the existence of islands slows the exploration of the DNA, since the protein can move from one island to another only by sliding. In fact, if the islands had a well-defined typical size $\bar{s}$, the protein dynamics would be diffusive on long scales $s \gg \bar{s}$. However, the problem is more intricate, since the distribution of island sizes

has the same heavy tail $p(s) \sim s^{-3/2}$ as the link length distribution, see Fig. S4. Nevertheless, the existence of islands is a crucial clue; we show below that it leads to a dynamics that can be described by a 1D transport model with traps and long-distance jumps. To this end, we first note two essential transport properties of islands: (i) Due to the internal links, the position of a protein is rapidly randomized within an island, such that for most starting positions within an island, it leaves the island with nearly the same probability to each side, see Fig. S5. (ii) The typical trapping time within an island scales as $\tau \sim s^{3/2}$ with the island size, see Fig. S6.

Given these properties, we now consider protein transport on an array of islands with sizes s_i drawn from the distribution $p(s)$. Each island has an associated trapping time $\tau_i(s_i)$. It will be instructive to allow for adjustable exponents μ and κ in the scaling behavior,

$$p(s) \sim s^{-1-\mu} \qquad \text{and} \qquad \tau \sim s^{\kappa} \,. \tag{1}$$

Combining these relations, we obtain a distribution of trapping times

$$w(\tau) \sim \tau^{-1-\mu/\kappa} \,, \tag{2}$$

since $w(\tau)\mathrm{d}\tau = p(s)\mathrm{d}s$. The transport behavior of the protein in the space of islands is then entirely determined by the ratio of the two exponents: Using the first passage time calculus [26], the typical time needed to move over n islands is

$$T \sim n \sum_{i=1}^{n} \tau_i \sim \begin{cases} n^{1+\frac{\kappa}{\mu}} & \text{for} \quad \kappa > \mu \\ n^2 & \text{for} \quad \kappa < \mu \end{cases} \,, \tag{3}$$

where the sum is dominated by the largest term for the case $\kappa > \mu$ (leading to sub-diffusive dynamics in island space) while a typical trapping time exists for $\kappa < \mu$ (diffusive dynamics in island space). To map the dynamics in island space back onto the DNA, we note that the total DNA length S of n islands scales as

$$S(n) \sim \begin{cases} n^{1/\mu} & \text{for} \quad \mu < 1 \\ n & \text{for} \quad \mu > 1 \end{cases} \,, \tag{4}$$

where the S is dominated by the largest island for $\mu < 1$. Combining (3) and (4) yields the transport behavior along the DNA, i.e. the typical time to travel a given distance. Fig. 3B shows the phase diagram spanned by the exponents μ and κ. It exhibits four different regimes. For $\mu > 1$, the distribution of island sizes has a well defined mean and no super-diffusion can occur, but sub-diffusive dynamics results when the trapping time distribution has a sufficiently heavy tail ($\mu < \kappa$). If $\mu < 1$, the dynamics is super-diffusive unless long trapping times in islands compensate for long jumps. In particular, $t \sim s^{\kappa+\mu}$ for $\mu < \kappa$, which includes the case of interest here, where the two exponents precisely add up to 2, explaining the quasi-diffusive dynamics in the quenched limit. Within our more general island model, a whole line of points exists where the dynamics is quasi-diffusive.

Discussion.— We found that protein jumping (including intersegment transfer) is an effective mechanism to destroy the redundancy of a diffusive 1D search only if the DNA dynamics is sufficiently fast (compared to the timescale between consecutive protein jumps). We have characterized the crossover from annealed to quenched DNA dynamics. Furthermore, we have explained the paradoxical quasi-diffusive dynamics in the quenched limit [23] as a subtle cancellation of the effect of traps and long-distance jumps in a random DNA configuration. An interplay between traps and jumps in 1D transport has been studied before in statistical models [27]. Here, we have identified a physical system where such an interplay occurs naturally, and surprisingly is self-tuned to a critical point of the dynamical phase diagram.

Acknowledgements. — We acknowledge useful discussions with Yariv Kafri, Joseph Klafter, Ralf Metzler, Igor Sokolov and funding by the DFG via NIM.

* Electronic address: gerland@lmu.de

[1] S. E. Halford and J. F. Marko, Nucleic Acids Res. **32**, 3040 (2004).
[2] A. D. Riggs, M. S. Bourgeois, and M. Cohn, J. Mol. Biol. **53**, 401 (1970).
[3] P. H. Richter and M. Eigen, Biophys. Chem. **2**, 255 (1974).
[4] O. G. Berg, R. B. Winter, and P. H. von Hippel, Biochemistry **20**, 6929 (1981).
[5] R. B. Winter and P. H. von Hippel, Biochemistry **20**, 6948 (1981).
[6] R. B. Winter, O. G. Berg, and P. H. von Hippel, Biochemistry **20**, 6961 (1981).
[7] G. D. Stormo and D. S. Fields, Trends Biochem. Sci. **23**, 109 (1998).
[8] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. USA **99**, 12015 (2002).
[9] R. F. Bruinsma, Physica A **313**, 211 (2002).
[10] D. Brockmann and T. Geisel, Phys. Rev. Lett. **91**, 048303 (2003).
[11] M. Slutsky and L. A. Mirny, Biophys. J. **87**, 4021 (2004).
[12] M. Coppey, O. Bénichou, R. Voituriez, and M. Moreau, Biophys. J. **87**, 1640 (2004).
[13] M. A. Lomholt, T. Ambjörnsson, and R. Metzler, Phys. Rev. Lett. **95**, 260603 (2005).
[14] T. Hu, A. Y. Grosberg, and B. I. Shklovskii, Biophys J. **90**, 2731 (2006).
[15] T. Hu and B. I. Shklovskii, Phys. Rev. E **76**, 051909 (2007).
[16] Z. Wunderlich and L. A. Mirny, Nucleic Acids Res. **36**, 3570 (2008).
[17] M. Sheinman and Y. Kafri, Phys. Biol. **6**, 016003 (2009).
[18] B. van den Broek, M. A. Lomholt, S.-M. J. Kalisch, R. Metzler, and G. J. L. Wuite, Proc. Natl. Acad. Sci. USA **105**, 15738 (2008).
[19] J. Elf, G.-W. Li, and X. S. Xie, Science **316**, 1191 (2007).
[20] G. Adam and M. Delbrück, in *Structural Chemistry and*

Molecular Biology, edited by A. Rich and N. Davidson (W.H. Freeman, San Francisco, CA, 1968), p. 198.
[21] O. Bénichou, M. Coppey, M. Moreau, P.-H. Suet, and R. Voituriez, Phys. Rev. Lett. **94**, 198101 (2005).
[22] R. Metzler and J. Klafter, Phys. Rep. **339**, 1 (2000).
[23] I. M. Sokolov, J. Mai, and A. Blumen, Phys. Rev. Lett. **79**, 857 (1997).
[24] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2000).
[25] J.-P. Bouchaud and A. Georges, Phys. Rep. **195**, 127 (1990).
[26] C. Gardiner, *Handbook of Stochastic Methods* (Springer, 2004).
[27] G. Zumofen and J. Klafter, Phys. Rev. E **51**, 1818 (1995).

Publication:

Parts of the material of this thesis have been prepared for publishing in the following work:

Thomas Schötz, Richard A. Neher, and Ulrich Gerland:
Target search on a dynamic DNA molecule.
In preparation, contains some of the results of chapter 3 of this thesis.

The article manuscript has been submitted to Physical Review Letters and is based on the manuscript presented on pages 131-135 of this thesis.

A current version of the article has been made available in 2011 at http://arxiv.org (Quantitative Biology > Biomolecules).
The official arXiv identifier of the article is: [arXiv:1109.3211]