NTNU – Trondheim
Norwegian University of
Science and Technology

The dynamics of DNA denaturation

Oda Dahlen

Nanotechnology
Submission date: June 2013
Supervisor: Titus van Erp, IKJ

Norwegian University of Science and Technology
Department of Chemistry

The dynamics of DNA denaturation

Oda Dahlen

June 16, 2013

Acknowledgements

This thesis was carried out under the supervision of professor Titus Sebastiaan van Erp at the department of chemistry. First and foremost, I would like to thank Titus S. van Erp for giving me the opportunity to work with this thesis, and for all the great help and guidance throughout these last six months. I appreciate very much the help he provided me with for understanding and gaining insight into the theory behind this work, dealing with problems related to the simulations, evaluating the results obtained and much more.
I would also like to thank senior engineer Egil Holvik at the deparment of scientific data processing for helping me run the simulations on the NTNU server system, and for answering questions when problems occured.
Last, but not least, I want to thank Susanne, Camilla, Monika and Kristina for making my days at NTNU unforgettable.

Oda Dahlen
Trondheim, June 16, 2013

Abstract

Gaining knowledge and understanding of the structure and function of DNA, our genetic material, is crucial for dealing with diseases related to DNA. In 2004, the mapping of the complete human genome was accomplished, leading to an enormous progress in treating DNA-related diseases. The attention was for a long time directed at the DNA structure, specifically the sequence. However, the knowledge of the structure of DNA is not sufficient to understand biological processes. For this, we also need to understand how this structure affects the equilibrium properties and the dynamics of the DNA molecule. In fundamental genetic processes, such as transcription and replication, DNA must undergo dynamical changes. Both processes are highly complex, and due to the lack of detail insight, a satisfactory descriptive model is difficult to design. However, since these processes require a local opening of the DNA molecule, they resemble DNA denaturation, or DNA melting, which is a considerably simpler process to study theoretically and experimentally. Studying DNA denaturation is, besides that it is interesting by itself, also considered a well-grounded step towards the full comprehension of the mechanisms involved in transcription and replication.

In this work, the denaturation of DNA was explored by computer simulations applying the Peyrard-Bishop-Dauixous (PBD) model. DNA chains consisting of 33 % AT base pairs and 66 % GC base pairs were investigated, as well as a key secondary structure of DNA and RNA, the hairpin, which is involved in many important processes of DNA and RNA. Although DNA dynamics has gained increased interest during the last decades, there is still a need of more insight and knowledge within this field. This work contains the first quantitative study in which dynamical data, like denaturation rate constants, of the well-known PBD mesoscopic model has been compared with experiments. Our work will be valuable for improvements of these mesoscopic models.

Contents

1 Introduction

DNA is the carrier of our genetic information, which is responsible for the development and functioning of all living organisms as well as several types of viruses. If the structure or function of DNA is impaired, the result can be a serious illness possibly followed by a fatal outcome. To understand and prevent these outcomes, a key factor is to gain knowledge of DNA and how it functions. It was formerly believed that it was mainly the structure of DNA that accounted for its functional properties, however, it is now a known fact that the dynamics is essential for many genetic processes[1][2]. Indeed, the structure does play a major role in determining the DNA functionality. For one thing, the structure contains the codes for the proteins produced upon DNA transcription. Also, even a single base substitution in the DNA structure can have serious consequences, such as cancer or inability to reproduce.
The complete mapping of the human genome was completed in 2004[3], and this insight lead to several advancements. It expanded the understanding of human evolution as well as the understanding of DNA structure-related diseases. Also, by genotyping viruses (identifying the difference between types of viruses), it became easier to find the appropriate treatment for a person infected with a specific virus. These examples are just a few of many and one can easily understand the importance of acquiring information related to DNA.
As mentioned above, the dynamics of DNA also plays a major role in the function of DNA, however, while the DNA structure is thorughly examined and described, the knowledge of DNA dynamics is rather limited. The transcription and replication of DNA, two central and vital processes for sustaining life, involves a dynamical alteration of the molecule. In order to expose the DNA bases to chemical reactions, the DNA double helix must locally open. Due to their high degree of complexity, the details of transcription and replication are not adequately understood, and a suitable model for these processes is very difficult to develope. Thermal DNA denaturation, a process in which the DNA molecule melts and completely opens, initiates with the formation of local "denaturation bubbles" in the same way as in transcription and replication. Due to this similarity, DNA denaturation is seen as a valid approach on the way of understanding the mechanism of transcription and replication. Besides the biological relevance, DNA denaturation is a very interesting process by itself. A well known mesoscopic model for describing DNA denaturation is the Peyrard-Bishop-Dauixous (PBD) model[2].

Understanding the dynamics of DNA denaturation is now also becoming vital to the field of nanotechnology. DNA is a suitable material for construction of nanostructures and functional nanodevices[4]. It has for example been proposed that altering the ionic strength to trigger a structural change or increasing the temperature to denaturate a DNA hairpin can be used as engines to drive nanodevices[5][6], or for storing molecular memory[7]. For the reasons stated above, the aim of this thesis is to contribute to the field of theoretical and computational models that try to describe DNA dynamics.

2 Theory

2.1 DNA denaturation and a theoretical model

Denaturation or melting of DNA means complete opening of the DNA chain, where the double strand separates into two single strands. The single strands are connected to each other by hydrogen bonds between two complementary bases, either A and T or G and C. A and T is linked with two hydrogen bonds, G and C with three bonds.
The denaturation process is influenced by a number of parameters, both surrounding conditions and characteristics of the chain. Surrounding factors includes among others temperature, the salt concentration(ionic strength), pH and other proteins or drugs in the solution. The characteristics of the chain influencing the denaturation is the length of the chain, the ratio of AT base pairs (bps) to GC bps, as well as the order of the bps, i.e. the specific sequence. Due to the strong dependence on the surroundings, the dynamic denaturation process must be described by a nonlinear system. Also, as the processes operates at relatively long length and time scales, having a model at a atomistic level would result in a extremely large computational cost, more or less impossible to obtain. Thus, this problem is avoided by using a coarse-grained model, a model which uses reduced representation and thus moves from the microscopic to the mesoscopic level.
In this work, the Peyrard-Bishop-Dauxois (PBD) model was used to describe the DNA chains. The PBD model describes the DNA molecule as a connected chain of one-dimensional particles, each representing the separation of a specific base pair. This effectively reduces the huge number of degrees of freedom one would have if each atom in the system were to be described at detail level. This model have been used to study the occurence of DNA bubbles, which are simply (long-lived) local openings of the DNA molecue occuring at temperatures below the melting temperature. The size and total number of DNA bubbles in a strand increases with increasing temperature, in the end resulting in complete denaturation of the dsDNA strand.
Several papers has shown that after fitting of the parameters the PBD model can reproduce denaturation experiments reasonably well[8][9]. The sharp phase transition seen in the melting curves of long homopolymers, are reproduced by the PBD model due to the nonlinear stacking interaction[10][2]. The model has also been shown able to reproduce specific features seen in experiments, such as bubble formation in DNA strands and the dependence on the sequence and temperature[11].

2.2 The PBD model

The Peyrard-Bishop (PB) model was introduced in 1989[12], a simple lattice model describing the denaturation of the DNA double helix. In 1993, the model was improved[2], and is now known as the Peyrard-Bishop-Dauxois (PBD) model. Prior to the establishment of the PBD-model, the statistics of DNA were computed using Ising-like models. Ising-like models calculates the ratio of open basepairs by giving the value 0 to those who are open, and 1 to those who are closed. Due to the discreteness of Ising-like models, they are not able to describe the dynamics of DNA, and the PBD model was thus an

continous analogue of the Ising model for describing nonlinear dynamics. The computational cost is considerably lower for the Ising model than for the PBD-model, but as the PBD-model operates at the mesoscopic level, its computational cost is not too high, and the efficiency of the model is satisfying.

The PBD model simplifies the geometry of the DNA double helix by assuming that the strands can be described as a one-dimensional chain. Each base pairs is represented as a point-mass along the DNA chain. The longitudinal amplitudes are significantly smaller than the tranverse ones[13], and longitudinal displacements are thus discarded. The interaction between two neighbouring bases at the same strand are described by a stacking potential. The interaction between two bases in a basepair (on opposite strands) are described by a Morse potential, and it considers the hydrogen bonds, the repulsion of phosphate groups as well as the screening by the surrounding solvent. The transverse displacements of the nucleotides from equlibrium are denoted y_i,i being the ith base pair. The potential energy of the system can then be represented by the following energy function:

$$U(y^N) = V_1(y_1) + \sum_{i=2}^{N} V_i(y_i) + W(y_i, y_{i-1}) \tag{1}$$

Notice that the equation contains solely y_i-terms, that are the base pair separation minus the equilibrium base pair separation.
V_k is the Morse potential, describing the repulsive interaction due to the hydrogen bonds between the two bases of a basepair on opposite strands

$$V_i(y_i) = D_i(e^{-a_i y_i} - 1)^2 \tag{2}$$

Where the depth of the potential is represented by D_i and the width by a_i. There are two possible values for D_i an a_i depending on whether i is a weak AT basepair or a strong GC basepair. As the GC bps is connected by three hydrogen bonds and the AT with two, the depth and the width of the V_i potential are larger for the strong GC bps than the weak AT bps.
This model treats A and T bases as identical particles, and it also does not distinguish between G and C bases. W represents the stacking potential, the interaction between neighboring bps, and includes a harmonic and a nonlinear expression.

$$W(y_i, y_{i-1}) = \frac{1}{2}K(1 + \rho e^{-\alpha(y_i + y_{i-1})})(y_i - y_{i-1})^2 \tag{3}$$

The term $\rho e^{-\alpha(y_i+y_{i-1})}$ was introduced in the improvement of the original PB-model, leading to the PBD model[14]. By removing $\rho e^{-\alpha(y_i+y_{i-1})}$, from equation (3), the remaining expression is the original harmonic PB model expression. It was discovered that that the stacking energy was a characteristic of the base pair, not the single bases. As an example, if one or both of the base pair opens (or merely stretches), $(y_i + y_{i-1})$ increases and $\rho e^{-\alpha(y_i - y_{i-1})}$ decreases and also W decreases. This shows the fact that the stacking interaction with neighboring bases is reduced when stretching and breaking a base pair, which is due to the altered electronic distribution occuring when the hydrogen

bonds of a base pair breaks. This gives the near-by bases more freedom to move, and increases the entropy. It becomes almost like a domino-effect, and this is the reason why in experiments, sharp phase transitions are often observed. Also, when a base pair is close to a open site, its vibrational frequencies is lower, and thus, its contribution to free energy is lower.
Along with introducing the PB model and later on the PBD model, the authors introduced parameters for long homogenous chains[12][14]. Applying these parameters for the PBD model produced satisfying melting curves, in particular when the parameters were improved in reference[14]. In 1998 and 1999, Campa and Giansanti presented parameters suitable for short heterogeneous chains (between 21 and 42 bps)[15][16], found by matching the model with experimental curves of DNA melting. A decade later, Theodorakopoulos introduced parameters for much longer heterogenous chains (between 1695 and 159662 bps)[9], found by matching the model to denaturation curves as well as dependence of salt content. The parameters given from Campa and Giansanti is as follows: $K = 0.025eV/\text{Å}^{\ 2}$, $\rho = 2$, $\alpha = 0.35\text{Å}^{\ -1}$, $D_w = 0.05eV$, $D_s = 0.075eV$, $a_w = 4.2\text{Å}^{\ -}1$ and $a_s = 6.9\text{Å}^{\ -1}$ The parameters given from Theodorakopoulos is $K = 0.00045eV/\text{Å}^{\ 2}$, $\rho = 50$, $\alpha = 0.20A^{-1}$, $D_w = 0.1255eV$, $D_s = 0.1655eV$, $a_w = 4.2\text{Å}^{\ -}1$ and $a_s = 6.9$ $\text{Å}^{\ -1}$.
Here, the subscript w or s refers to the specific weak AT and strong GC parameters, respectively. The nonlinear stacking parameter, ρ, is 25 times larger in the Theodorakopoulos than in the Campa and Giansanti parameter set. This large difference in the ρ effectively means if one base pairs moves out of stack, the rest of the base pairs feels this base pair less in the simulation with the Theodorakopoulos parameters. The K-value for Campa and Giansanti is more than 50 times larger than the K-value for Theodorakopoulos. If we look at K and ρ together, the term $K(1+p)$ in the Morse potential will be 0.075 for the Campa and Giansanti and 0.02295 for the Theodorakopoulos, where the term is 3.3 timer larger for the Campa and Giansanti parameter set, so perhaps the large difference in ρ and K, respectively, is effectively not so large after all. Both parameter sets are in use today.

2.3 Limitations of the PBD model

There are a number of limitations that should be considered when working with the PBD-model. First of all, the stacking potential does not distinguish between the weak AT and the strong GC base pairs. Although the parameter D in the Morse potential distinguishes between weak AT and strong GC base pairs, it treats the A and T bases as identical, and the G and C bases as identical. Another limitation of the PBD model lies in the simplicity of the model. It does not contain helicical stress, the stress involved in a double helix. There exists is a more complicated adaptation of the PBD model that includes helicity[17], but this model is much more expensive, and for short chains this helicical stress is not considered to be very important[18]. Also, the PBD model does not use explicit solvents, but includes solvent effects only approximatly via the effictive interaction parameters. Last, it is important to remember that for very large base pair

separations, the model becomes less trustable due to its one-dimensional character.

2.4 Molecular dynamics

Molecular Dynamics is a powerful numerical and computational method used to simulate the movement of atoms and molecules and from this, estimate physical quantities. In an MD simulation, the movements of the components of the system is computed by integrating the equations of motion. However plain MD can not be used to simulate most types of reactions, as it only reaches the microscopic time. A solution to this problem was pursuited by Wigner and Eyring already in the 1930s, and they introduced the Transition State theory (TST), and within this theory, the principle of the Transition State (TS).

2.5 Transition State Theory

Transition State Theory (TST) is used to explain how chemical reactions take place, and provides an expression for the reaction rates of the chemical reaction. Consider a very simple reversible reaction:

$$A \rightleftharpoons B \tag{4}$$

where A and B are the reactants and products, respectively.
In order for the reaction to be carried out, the reactants must have sufficient energy to overcome the energy activation barrier. At this barrier, TST hypotesizes a transititon state. At this hypotetical transitition state, the reactants have formed an unstable activated complex high in energy and ready to proceed to become products or go back to the reactant state. The activated complex is in a quasi-equilibrium with the reactants, meaning that for a equilibrated system, the activated complexes will also be at equilibrium with the reactants, see figure 1.

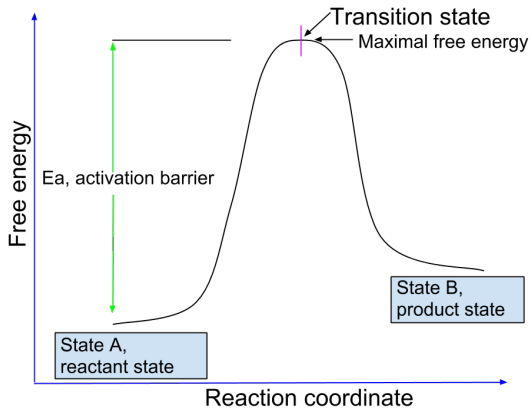


Figure 1: Illustration of the free energy profile along the reaction coordinate of the simple reaction $A \rightleftharpoons B$ with the transition state at the maximum of the free energy profile.

The rate constant of such a reaction usually follows the Arrhenius law.

$$k = A \cdot e^{-E_a \beta} \tag{5}$$

where k is the rate constant, A is the frequency factor and E_a is the activation energy, the difference in energy between the transition state and the reactant state. β is given by $\frac{1}{RT}$, where R is the universal gas constant and T the temperature in Kelvin. The rate constant has units $\frac{1}{s}$. The denaturation rate represents the fraction of closed DNA molecules that denaturate per units of time. By taking the natural logarithm on each side of the equation, one gets

$$\ln k = -\frac{E_a}{RT} + \ln A \tag{6}$$

Therefore, if a reaction obeys Arrhenius behaviour, the activation energy can be found from the slope of a plot of $\ln k$ versus $\frac{1}{T}$, and the frequency factor from the interception with the y-axis.

2.5.1 TST rate constant and reaction coordinate

Before we present the TST expression for the rate constant, let us define some conditions and introduce a new variable, the reaction coordinate(RC). First of all, we are operating with an dynamical system, having two stable states A and B on each side of the reaction energy barrier. As long as this barrier is adequately high, the system will show an exponential relaxation, and the forward and backward rate constants are easily described.
Let us assign $\lambda(x)$ as the reaction coordinate, which is a function x, being all the coordinates of the system. Let $P(\lambda')$ be the probability density that the reaction coordinate $\lambda(x)$ takes a certain value λ'. The free energy along the reaction coordinate can then be expressed as $G(\lambda') = -k_B T \ln P(\lambda') + const.$ This constant is gauged to have have $G = 0$ at the bottom of the reactant well. This free energy profile has a minimum at the reactant and product state, and a local maxima at the transition state, where $\lambda(x) = \lambda^*$. Hence, λ^* is the surface dividing states A and B, at the free energy profile.
When $A < B$, we know that if $\lambda(x) < \lambda^*$, the system is in A and if $\lambda(x) > \lambda^*$, the system is in B. One important assumption in TST is that a trajectory coming from state A and crosses λ^* will stay in state B for a long period of time. Hence, in TST, all pathways heading towards the product site B are assumed to stay in B for a very long time.
The TST expression for the forward rate constant can be expressed as follows:

$$k_f^{TST} = \frac{1}{2} \langle \mid \dot{\lambda} \mid \rangle \frac{e^{-\beta G(\lambda^*)}}{\int_{-\infty}^{\lambda^*} e^{-\beta G(\lambda)} d\lambda} \tag{7}$$

$\dot{\lambda}$ is the time derivative of the reaction coordinate, given as $\lambda = \frac{\partial \lambda}{\partial x} \frac{\partial x}{\partial t}$ and $\langle ... \rangle$ is the ensemble averages at equilibrium. β is given by $\frac{1}{RT}$, where R is the universal gas constant and T the temperature in Kelvin. G is the free energy along the reaction coordinate, λ, and the height of the barrier corresponds to the activation energy in equation 5, which means that term $\frac{1}{2} \frac{\langle |\dot{\lambda}| \rangle}{\int_{-\infty}^{\lambda^*} e^{-\beta G(\lambda)} d\lambda}$ corresponds to the frequency factor in equation 5.

2.5.2 Limitations of TST

TST is a technique in widespread use among researchers. However, there are some limitations in the TST that must be commented upon. As the TST expression for the rate constant strongly depends on the activation energy E_a, it is crucial for the calculation to know the exact value for E_a. This stresses the importance of an exact choice of the reaction coordinate for which recrossings do not occur, which is a very diffucult task. One should keep in mind that the TST expression of the rate constant and the free energy barrier can depend strongly on the choice of the reaction coordinate. This is because, even if a molecule reaches the top of the free energy barrier, it is not certain that it will go to the product state, it might go back to the reactant state. For this reason, a wrong choice of the reaction coordinate would lead to an amplified rate constant.

2.6 The dynamical transmission coefficient

Due to the limitiations of TST stated in section 2.5.2, Keck established the calculation of the transmission coefficient[19].
The transmission coefficient is a dynamical factor that corrects for fast correlated recrossings[20]. This effectively avoids the problem of overestimating the rate constant. By multiplying the transimission coefficient by the TST rate constant, one obtains the true rate constant, given as

$$k_f = k_f^{TST} \cdot \kappa \tag{8}$$

Thus, finding the rate constant now consist of two basic operations, calculation of the free energy and calculation of the dynamical transmission coefficient. This approach is also called the reactive flux approach.

There exist several methods to calculate both types of quantities. In this work, for calculation of the free energy, the direct numerical integration method was used, and for calculation the transmission coefficient, the effective positive flux (EPF) method was used.

2.7 Calculation of the transmission coefficient by the EPF formalism

There exists several methods of calculating the transmission coefficient within the RF method, as was summarized in reference[21]. We do not comment upon these various formalisms, but turn to the effective positive flux (EPF) formalism, that is shown to be the most effective[21]. By including the transmission coefficient, rather than counting every positive crossing event as in TST, the RF method count only effective crossing events. The EPF calculates the transmission coefficient κ by counting for each crossing which one goes in the right direction and which ones are "effective". With effective, the EPF approach means that a crossing must be a first-time crossing of the transition dividing surface λ^* since leaving state A and must finally go to B without revisiting state A, see figure 2.

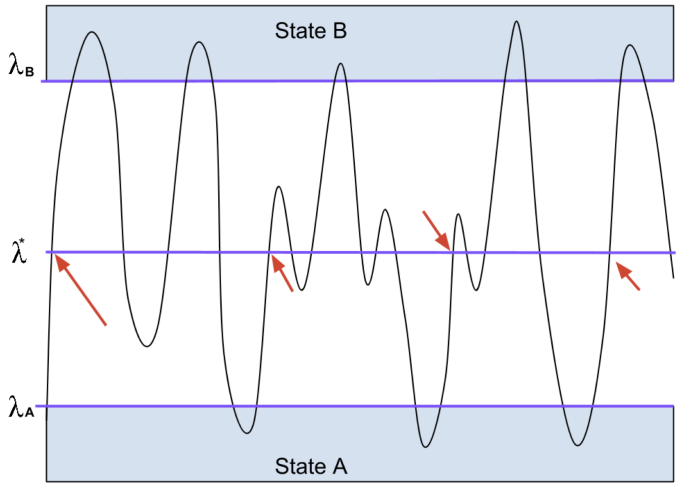


Figure 2: Illustration of the counting of the effective positive flux EPF formalism for calculating the transmission coefficient. The red arrows points at each effective crossing, at which we assign the value 1 to χ. All other crossings have $\chi = 0$.

The transmission coefficient can be expressed as

$$\kappa = \frac{\langle \dot{\lambda}\chi \rangle^{TS}}{\langle \dot{\lambda}\theta(\dot{\lambda}) \rangle^{TS}} \tag{9}$$

χ is equal to 1 if it is a effective positive crossing, and 0 otherwise.
To perform this calculation, it implies that you generate a lot of configurations having $\lambda(x) = \lambda^*$, then randomize the velocity according to the Maxwell-Boltzmann distribution, and from this point we go backwards and forwards in time using molecular dynamics until we know if it is an effective positive crossing or not. In practice, this implies that we first go backwards in time until reaching state A or recrosses the TS surface. In the latter case, we know it is not an effective positive crossing, and we can assign $\chi = 0$ for this point. In the first case we will continue this configuration forward in time until reaching state A or B. Only if it reaches state B, we will assign $\chi = 1$ for this point. Otherwise, $\chi = 0$ again. After this, we sample a new configuration with $\lambda(x) = \lambda^*$, and perform this procedure numerous times, to obtain sufficient data. In our case, we typically ran 10^6 trajectories.
If we mulitply the obtained transmission coefficient with the TST expression, $k = k^{TST}\kappa$, we get the exact value for the reaction rate. In the RF formalism, both the transmission coefficient and the height of the free energy barrier depends on the choice of the reaction coordinate. However, the final rate constant that is calculated in the RF method depends on both quantities in such a way that it becomes independent of the choice of the reaction coordinate. The method is even exact if the TS dividing surface is not taken at the maximum of the free energy barrier. With exact we mean that it should give the same answer as an infinitely long MD simulation, but orders of magnitude faster. In fact, for our PBD-model, we assume the system is in state B when the transition

state surface is crossed. Here, $\lambda^* = \lambda_B = \xi$, see figure 2. It is important to notice that a proper choice for the RC in TST is neccesary to obtain valid results, and in the RF method, a good choice of the RC is critical to the efficiency of the method. An alternative to the RF method are the Transition Path Sampling (TPS) based methods, that are even more insensitive to the reaction coordinate. Transition Interface sampling (TIS) improved upon the original TPS, and its efficiency has shown to be very stabe irrespective to the choice of the reaction coordinate[21]. However, for our system, the RF method is extremely fast, as we can use a very efficient method for the calculation of the free energy, as explained in the next section.

2.8 Free energy calculation using direct numerical integration method(DNIM)

Suppose a dsDNA should fully denature, all the base pairs must open, that is, they must reach the Morse potential plateau. If only one base pair has not opened, it can pull all the others back and avoid dissociation. By choosing the reaction coordinate as $\lambda(y^N) \equiv min[\{y_i\}]$, both the state of full dissociation and association can be represented. The opening threshold value, denoted as ξ, was chosen to be 1 Å . The reaction coordinate chosen here is only chosen for computational effectiveness. Naturally, if the aim was to find the exact reaction mechanism, a variable related to the positions of all the entities in the system would be the most suitable. As this was not a goal for this work, the choice made above is appropriate. When simulating close to room temperature, denaturation is rarely occuring, and gives an excellent opportunity to calculate the rate constant.
Alternative to calculating the free energy by umbrella sampling and the termodynamic integration method, is the direct numerical integration method. This method can only be applied for specific models, but if it can be used, it is much more efficient than the others. This specific application of the method was developed by the authours of reference[**?**]. In statistical physics, most properties like average energy or the probability density can be expressed as the ratio of two multi-dimensional integrals. Solving these multi-dimensional integrals numerically, is normally not possible, and it is therfore that we need to use molecular dynamics (MD) or Monte Carlo(MC) methods. However, in some cases, these integrals can be solved, which is the direct aim of the DNIM approach. DNIM exploits specific features of the mathematical model to express the multi-dimensional integrals into products of low-dimensional integrals that can be solved by iteration. In specific, an N-dimensional integral can be solved by the numerical integration method(DNIM) method, if the integrant is of a factorizable form as shown below.

$$Z = \int dy^N a^{(N)}(y_N, y_{N-1}) ... a^{(3)}(y_3, y_2) a^{(2)}(y_2, y_1) \tag{10}$$

Now, we can solve this integration using the following iterative scheme

$$\begin{aligned} z^{(2)}(y_2) &= \int dy_1 a^{(2)}(y_2, y_1), \\ z^{(3)}(y_3) &= \int dy_2 a^{(3)}(y_3, y_2) z^{(2)}(y_2) \end{aligned} \tag{11}$$

and so on, until one reaches

$$z^{(N)}(y_N) = \int dy_{N-1} a^{(N)}(y_N, y_{N-1}) z^{(N-1)}(y_{N-1}) \tag{12}$$

which gives

$$Z = \int dy_N z^{(N)}(y_N) \tag{13}$$

Now, given that we know $z^{(k-1)}$, in order to find $z^{(k)}$ for a discrete set of n_{grid} values y_i, only ${n^2}_{grid}$ (function) evaluations is needed. Thus, for the calculation of all $z^{(k)}$, $k \in \{1 : N\}$, a total number of $N \cdot {n^2}_{grid}$ function evaluations are required. This trick reduces the ${n^N}_{grid}$ function evaluations originally required greatly. However, the calculations can be further optimized by applying convenient cutoffs for the integration boundaries.

2.9 DNIM method applied to the PBD-model

The probability $P(\lambda_B)$ can be expressed in the following way:

$$P(\lambda') = \frac{\int dy^N \delta[\lambda(y^N) - \lambda'] e^{-\beta U(y^N)}}{\int dy^N \theta[\lambda^* - \lambda(y^N)] e^{-\beta U(y^N)}]} \tag{14}$$

Since each base pair i can be the minimum, we can rewrite the above probability as the sum of sub-probabilities. Each sub-probability corresponds to the probability that particle i is exactly at λ', while the other particles j has y_j-values larger than λ'.

$$P(\lambda') = \frac{\sum_i \int dy^N \delta[y_i - \lambda'] \prod_{j \neq i} \theta(y_j - \lambda') e^{-\beta U(y^N)}}{\int dy^N [1 - \prod_k \theta(y_k - \xi)] e^{-\beta U(y^N)}} \tag{15}$$

Since $U(y^N) = V_1(y_1) + \sum_{i=2}^{N} V_i(y_i) + W(y_i, y_{i-1})$, all the integrals in the above equation can be solved by DNIM, since they are of the specific factorizable form.
In this work, we applied numerical integration boundaries so that the integration stops when the weigth $e^{-\beta U}$ of a specific arrangement is very low. Thus we confine y_i to lie between L and R, and in addition, we stopped the integration whenever $\mid y_i - y_{i+1} \mid \leq \sqrt{\frac{2|\ln \epsilon|}{\beta K}}$. L and R are defined as

$$\begin{aligned} L &= -(1/\alpha_{AT}) ln[\sqrt{|\ln \epsilon| \beta D_{AT}} + 1] \\ R &= y_0 + \sqrt{N} d \end{aligned} \tag{16}$$

For ϵ, the value applied is 10^{-40}, which in fact is considerable lower than whats required for our calculations. Now, all arrangement giving a contribution of $e^{-\beta V(yi)}$ smaller than ϵ will be disregarded. The numerical integration is carried out using Simpson's rule with an integration step d_y.
d_y should ideally be set as small as possible, to obtain a minimal systematic error. The

smaller the integration step, the more accurate the results, but the larger the computational cost. The cpu cost scales as $(1/d_y)^3$. The integration step must therefore be set to a value than balances the computational cost and the error. Due to the low value of K in the parameter set of Theodorakopoulos, 0,00045 compared to Campa and Giansanti, which is 0,025, the computational cost is more than 50 times larger for the Theodorakopoulos parameter set. Thus, one should keep in mind that when applying the Theodorakopoulos parameter set, a larger integration step might be required to reduce the computational cost.

Normally, one would apply a integration step of 0.05, 0.025 or 0.01 Å, depending on the time and computational space available.

2.10 Langevin dynamics

Langevin is a very commonly used technique to describe the effect of a solvent that is not explicitly described in terms of actual solvent molecules. Langevin assumes that molecules moving through this solvent feel a certain friction due to the viscosity of this solvent. In addition, molecules are constantly perturbed by random forces describing the collisions between the solute and solvent molecules. In Langevin dynamics, the frequency of collisions and the friction coefficient are related in a such a way that they do not change the statistics of the system. Hence, static equilibrium properties, such as denaturation curves will not change when changing the friction coefficient, but the dynamical properties such as the rate constant, will change. In reference[22], the relationship between γ and κ was examined. The authors found that for a large friction coefficient($\gamma > \frac{10}{ps}$), the relationship could be described by Kramers behaviour $\kappa = \frac{1}{\gamma}$.

2.11 The melting temperature of DNA chains

In reference [**?**], van Erp et al. obtained melting curves for pure AT and GC-sequences at various chain lengths, see figure 3. We can see from figure 3 that for short chains, the

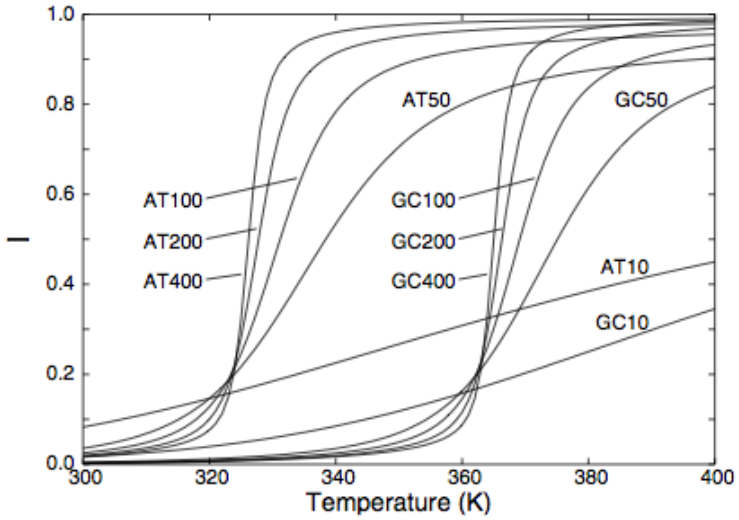


Figure 3: Melting curves of pure AT and GC chains obtained by the PBD model using free boundary conditions. The y-scale is the fraction of open base-pairs given that the molecules are in the double stranded state. This picture is taken from reference[**?**].

melting occurs approximately linear to the temperature, and the longer the chain, the sharper the melting transition. This sharp transition is due to non-linear character of the PBD stacking potential. These sharp curves is found experimentally, and is seen as one of the very few examples of a quasi one-dimensional phase transition.

2.12 DNA hairpins

Until now, we have for the most part discussed the primary structure of DNA, the double stranded DNA. However, the interest of the dynamics of DNA denaturation is not limited to its primary structure; rather, it has been of great interest to researchers to obtain knowledge of the secondary structure. One of the structures of particular interest is the DNA hairpin. A hairpin or a hairpin-loop is a secondary structure of DNA/RNA, and is found most frequently in RNA, where it is one of the entities of many RNA configurations. Hairpins are highly dynamic structures, and in a simplified definition, they fluctuate between two main states, the open state and the closed state. A hairpin consists of a base pair stem and a base loop. The stem is a double stranded DNA chain, while in the loop, the bases are free. In the closed state, see figure 4, the base pairs of the stem are paired, and therefore, the enthalpy of the system is low. In the open state, see figure 5, the base pair stem has opened, and due to the numerous configurations achievable for a single DNA/RNA strand, the open hairpin is in a state of high

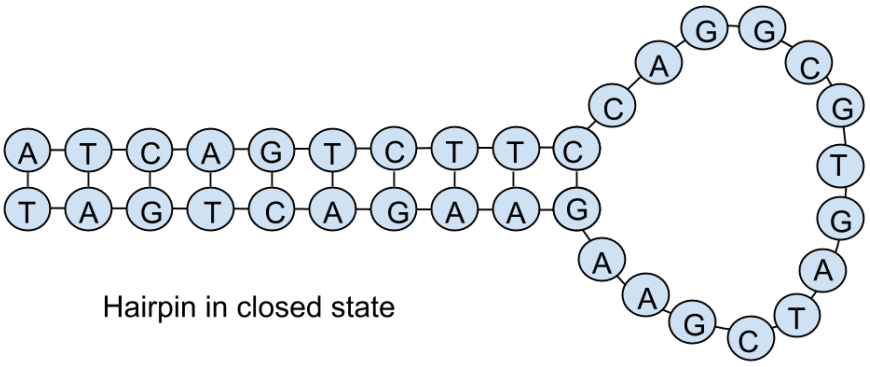


Figure 4: Illustration of a DNA hairpin with a 15 bps stem and a 14 bps loop in the closed state.

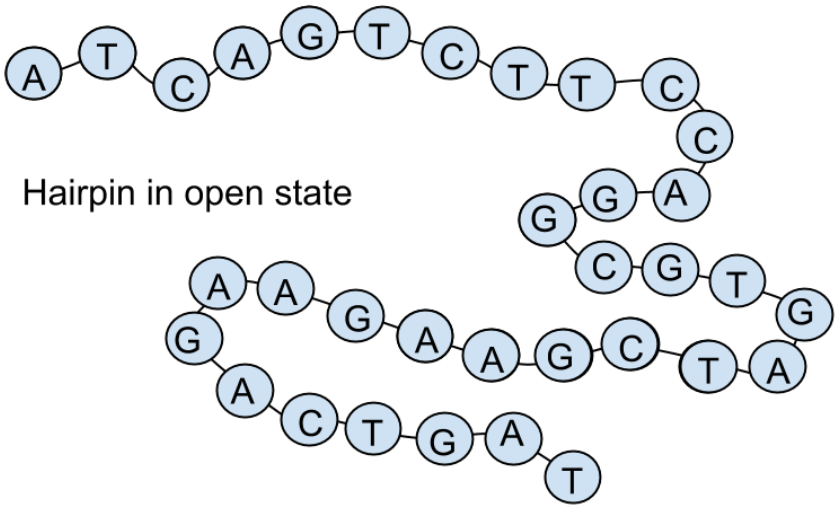


Figure 5: Illustration of a DNA hairpin with a 15 bps stem and a 14 bps loop in the open state.

entropy. Hairpins are vital for several of the functions of DNA and RNA. They are involved in regulation of the transcription by binding to proteins[23], the regulation of gene expression[27], and intermediary hairpin structures participate in replication and recombination[24][25][26]. DNA/RNA hairpins has been proposed suitable for storing molecular memory[7], and as engines to drive nanodevices[5] [6].
The PBD model represents a double stranded DNA, and to simulate hairpins, one needs to somehow mimic the effect of the loop of the hairpin. Errami et al[28] used a com-

bination of two models to describe the hairpin, the PBD model for describing the stem and the Kratky-Porod model for describing the loop. Reference [29] used a much more crude approach for describing the loop, by simply applying a square potential of $y = 50$ Å (from personal correspondence with the authors of reference [29]) to the base pair at the end of the stem (which would be connected to the loop). Since our integration method requires a very simple model, we adapted a similar approach, where we applied an exponential potential at the last base pair of the stem. We used an exponential expression because it was more convenient for MD trajectories. We can adjust at which base pair separation the hairpin effect should start. We define the 'Cutoff' value as this separation. The expontential potential can then be written as:

$$V(y_N) = \begin{cases} (y_N - \text{cutoff})^6 & \text{when} \quad y_N > \text{cutoff} \\ 0 & \text{when} \quad y_N < \text{cutoff} \end{cases} \tag{17}$$

In our simulations, for each hairpin we simulated, we applied a range of cutoff values to see how the cutoff applied influenced the results. By applying a potential at the last base pair to create a hairpin effect, we can not say anything about the characteristics of the loop. However, for calculations of the opening rate constants, this is not such a big problem, since it has been reported that the opening rate is only slightly affected by alteration of the loop[30], while the closing rate is much more affected by changing the loop characteristics.

3 Simulation details

The simulations were carried out by a fortan code, written by Titus van Erp. For each specific sequence, the simulations gives the free energy along the reaction coordinate, the transmission coefficient and the rate constant. The transmission coefficient were obtained by running 10^6 trajectories, using a timestep of 1 fs and bp masses of 300 amu. In all simulations, Langevin dynamics were applied. The friction coefficient, γ, was in these experiments set to 0.50902 in atomic units or $50\frac{1}{ps}$, which is considered as suitable value for the friction of water[22].
For sequences less than 50 bps, we used an integration step of 0.01 for the Campa and Giansanti parameter set, and 0.05 for Theodorakopoulos parameter set. For sequences more than 50 bps, the computational cost is so large that we used an integration step of 0.05 regardless of the parameter set.
We ran simulations for two main types of DNA structures, double stranded DNA (ds-DNA) and DNA hairpins.
For the ds DNA chains, we used the Campa and Giansanti parameter set, and obtained results showing details of the the behaviour of the PBD model, as well as comparing sequences with various orders of the basepairs to each other. For the DNA hairpins, we ran simulations for both the Campa and Giansanti and the Theodorakopoulos parameter set. We compared the results to experimentally obtained values found in the literature, as well as comparing the two parameter sets to each other.

4 Results and Discussion

4.1 Investigation of double stranded DNA chains consisting of 33 % AT-bps and 66 % GC-bps

We did numerous simulations of double stranded DNA chains all having 33 % AT bps and 66 % GC bps, where we looked at various lengths of the chains (ranging from 6-99 bps) and various placements of the base pairs, i.e. the specific sequence. For all these simulations, we used the Campa and Giansanti parameter set. To ease all further explanations, we identify four main sequences by numbers. For further simplicity, frow now on, we only write the sequence for one of the double strands, so the sequence AAAGGG represents a DNA double strand of three AT bps and three GC bps.
Sequence 1: The sequence having all the AT bps at one end of the chain, and all the GC bps at the other end. For N=6 basepairs it would be AAGGGG, for N=12 it would be AAAAGGGGGGGG and so on.
Sequence 2: having all the GC bps in the middle of the chain, and half of the AT bps at each end of the chains, also called the strong block in the middle-sequence. For example, for N=6, it would be AGGGGA, for N=12 it would be AAGGGGGGGGAA, for N=18 it would be AAAGGGGGGGGGGGGAAA and so on.
Sequence 3: The exact opposite of sequence 2. All the AT bps in the middle of the chain, and GC bps at each end, also called the weak block in the middle-sequence. For N=6 it would be GGAAGG, N=12 would have GGGGAAAAGGGG and so on.
Sequence 4: The alternating sequence GAGGAGGAG..., for N=6 it would be GAGGAG, for N=12 GAGGAGGAGGAG and so on.

First, we present the rate constants and the transmission coefficients for sequence 1-4 as a function of the chain length at temperatures ranging from 300-400 Kelvin.

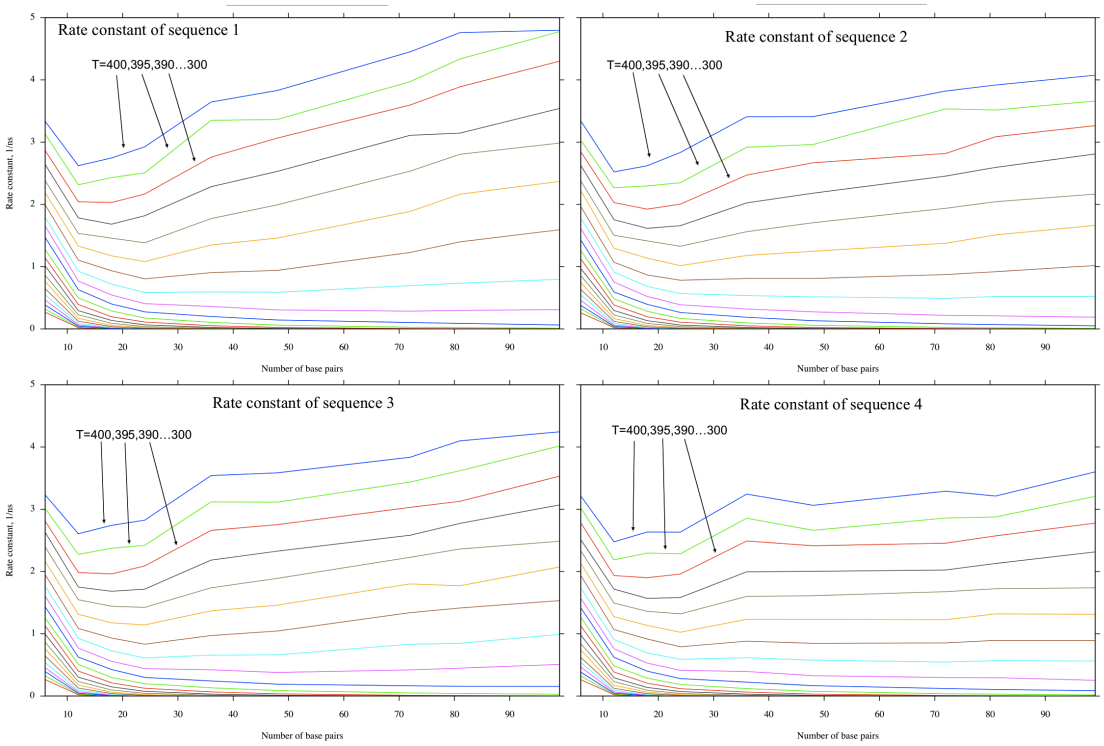


Figure 6: Plots of the rate constants for sequence 1-4 as a function of the chain length, at temperatures 300-400 K. All chains containing 33 % AT and 66 % GC bps.

We see in figure 6 that for all sequences, below a certain temperature, the rate constant decreases exponentially as a function of the chain length. Above this specific temperature, the rate constant decreases at first, for short chains, and then, at longer chains again starts to increase. The higher the temperature, for shorther sequences the onset of the increase. This 'onset' temperature is named the critical denaturation temperature[22]. For sequence 1 and 3, this temperature is 360 K, while for sequence 2 and 4, this temperature is 355 K. In reference [22], the authors found that for a pure AT chain, the critical denaturation temperature is approximately 325 K. This corresponds to the melting curve in figure 3. Thus, it seems like the melting temperature for chains containing 33 % AT and 66 % GC bps is close to 360-365 K.

One should also notice from figure 6 that all four sequences has more or less the same rate constants for short chains and below the critical denaturation temperature. However, for long chains above the the critical denaturation temperature, we can see that sequence 1 has larger rate constants than the other sequences, sequence 3 has the second largest rate constans followed by sequence 2, and sequence 4 has the lowest rate constants.

Plots of the transmission coefficients of sequence 1-4 as a function of chain length is shown in figure 7.

4.1.1 The normalized transmission coefficient

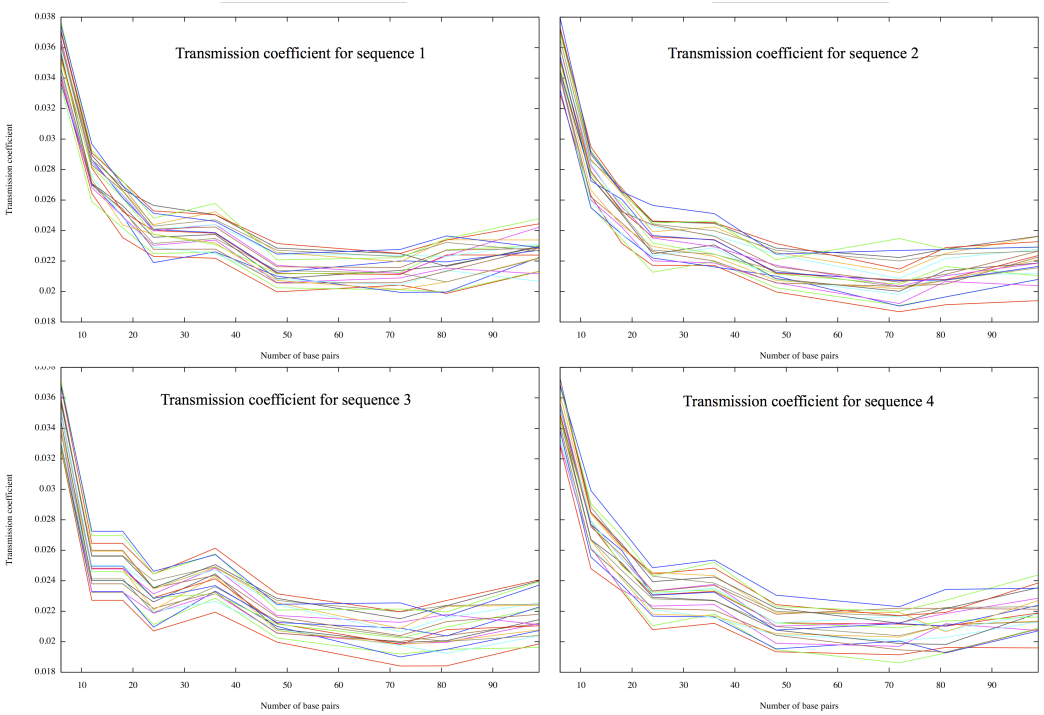


Figure 7: Plots of the transmission coefficient for sequence 1-4 as a function of the chain length, all chains containing 33 % AT and 66 % GC bps.

In figure 7 we see that for all four sequences, initially the transmission coefficient has a steep decrease with increasing chain length. At a chain length of 36 bps it has a small upturn before it starts to decrease again. It seems that it starts to have a slight increase again at a chain length of 80 bps. In figure 8, we have averaged the transmission coefficient for sequence 1-4, where we naturally see the same trend as in figure 7.

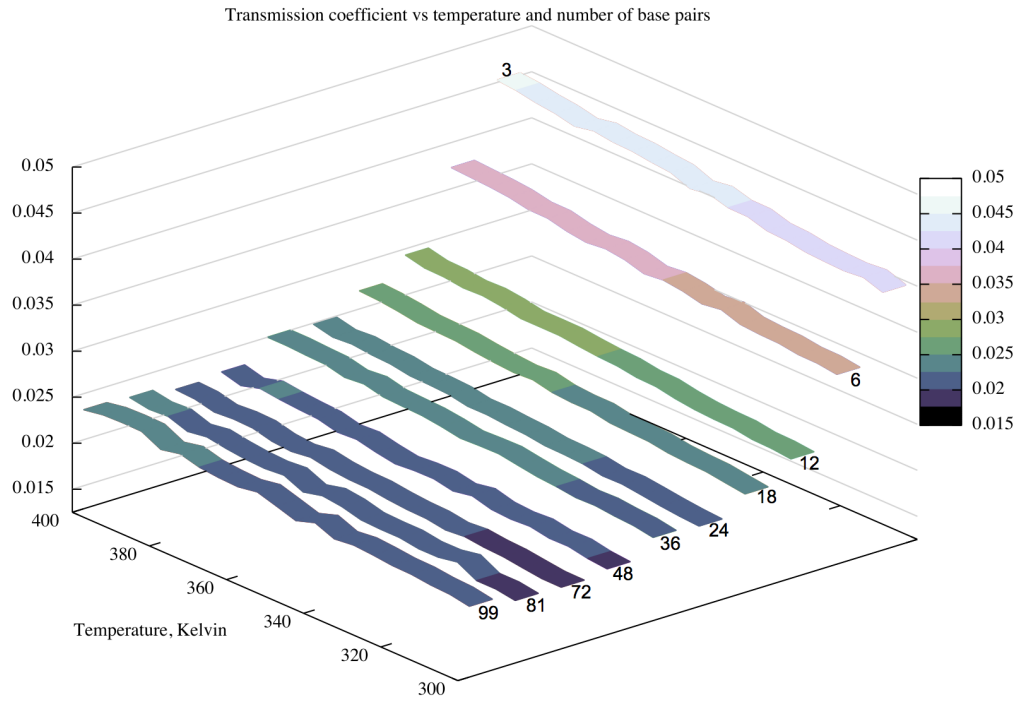


Figure 8: The normalized transmission coefficient as a function of chain length at temperatures 300-400 K. For each chain length at a certain temperature, we averaged the transmission coefficient for sequence 1-4.

In the paper of van Erp and Peyrard[22], they found that the transmission coefficent can be approximated to $\frac{1}{50} = 0.02$ for chains having a content of 50 % AT and 50 % GC bps. In this paper, they concluded that the transmission coefficient stabilizes for chain lengths of 100-120 bps at a value just slightly above 0.02. In our work, the largest chain investigated had a length of 99 basepairs, but we can assume that the transmission coefficent stabilizes at the value it has for the largest sequences. Thus, we conclude that 0.02 is a reasonable transmission coefficient value for sequences above 20 bps for 33 % AT and 66 % GC chains. The rate constant is given by $k = P(\lambda^*) \cdot R$, where R is the unnormalized kappa, given by $R = \frac{\kappa}{\sqrt{2\pi\beta m}}$. Thus, by assuming that $\kappa = 0.02$, one can avoid the calculation of the transmission coefficient, and thus, calculate the rate constant with a severe reduction of the computational cost.

4.1.2 Relative rate constants of sequence 1-4, chains consisting of 33 % AT and 66 % GC bps

In reference [22], van Erp and Peyrard obtained PBD model results for DNA chains consisting of 50 % AT bps and 50 % GC bps. In this paper, they found that the alternating sequence AGAGAGAG.., which corresponds to sequence 3 for 33 % AT 66 % GC chains, has a significantly larger rate constant than sequence 1 (strong block in the middle), sequence 4 (weak block in the middle) and sequence 2 (all AT bps at one end, all GC bps at the other end). Therefore, we investigated chains of sequence 1-4, all having 33 % AT and 66 % GC bps, too see if our results corresponds with the result obtained in reference [22]. We divided the rate constant of two sequences with each other, respectively, and obtained the relative rate constant of sequence X to sequence Y.

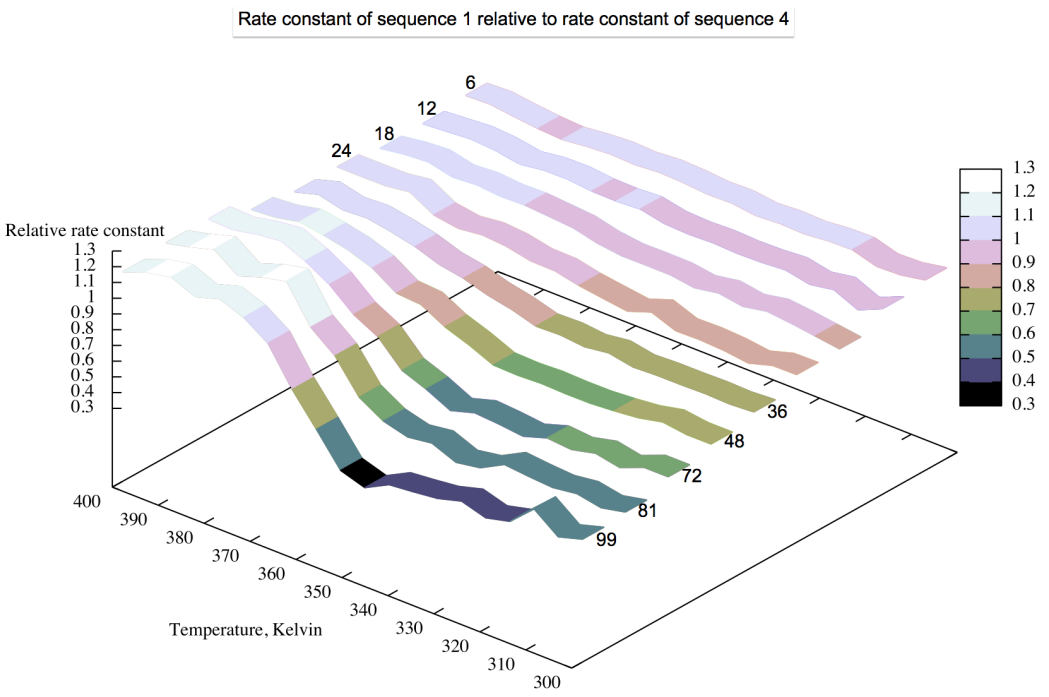


Figure 9: Rate constant of sequence 1 divided by rate constant of sequence 4, plotted as a function of the temperature and chain length.

Figure 9 shows that for short chains, the rate constant is more or less the same for the two sequences, which is expected, as short chains opens up quite easy regardless of the placement of the base pairs. However, the longer the chain, the larger the difference of the rate constants for the two sequences. For long chains at temperatures below 360 K, the alternating sequence 3 has a significantly larger rate constant than sequence 1, which is in agreement with reference[22], and the largest difference, seen as a valley in the plot, occurs at temperatures between 335-355 K. However, as the temperature exceeds 370 K, the rate constant of sequence 1 surpasses the rate constant of sequence 3, an unexpected behaviour which does not correspond with the results in reference[22]. One should notice that the relative rate constant has a steep increase with increasing temperature in the interval 355-370 K, around the assumed melting temperature of 33 % AT 66 % GC chains.

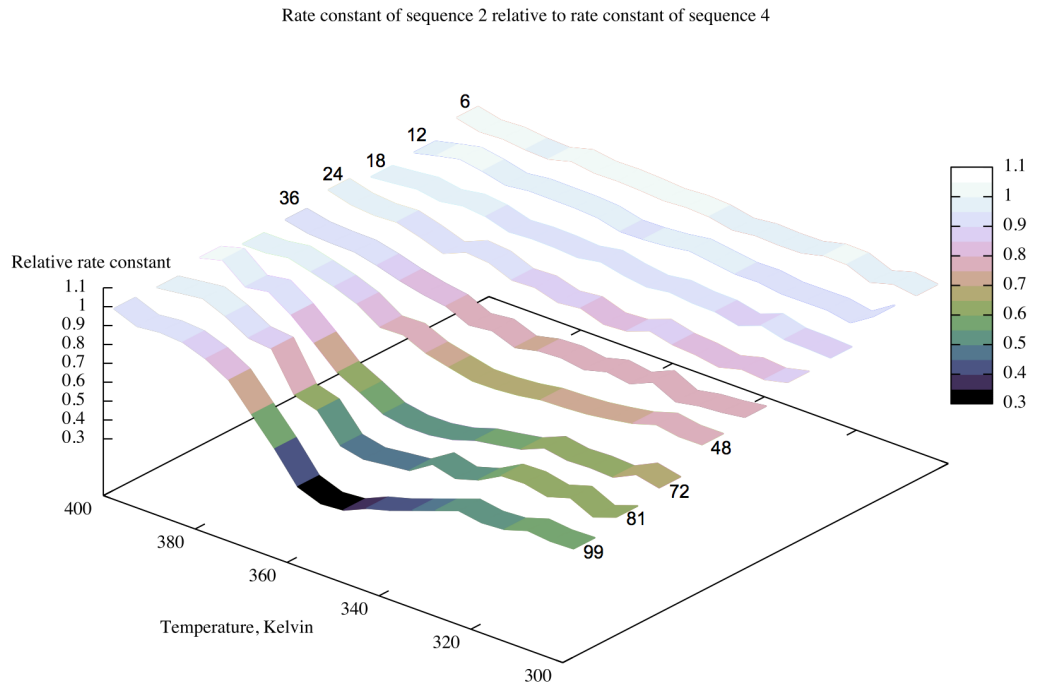


Figure 10: Rate constant of sequence 2 divided by rate constant of sequence 4 plotted as a function of temperature and chain length.

We can see from figure 10 that for short chains, the rate constants of the sequences is more or less the same. Also, for longer chains at temperatures above 380 K, the relative rate constant is approximately 1. However, for longer chains below 380 K, the rate constant for sequence 3 suddenly becomes much larger than the rate constant of sequence 2, and the largest difference is shown in a valley occuring at 350-360 Kelvin. Thus, there is a steep increase in the relative rate constant occuring between 360-380 K.

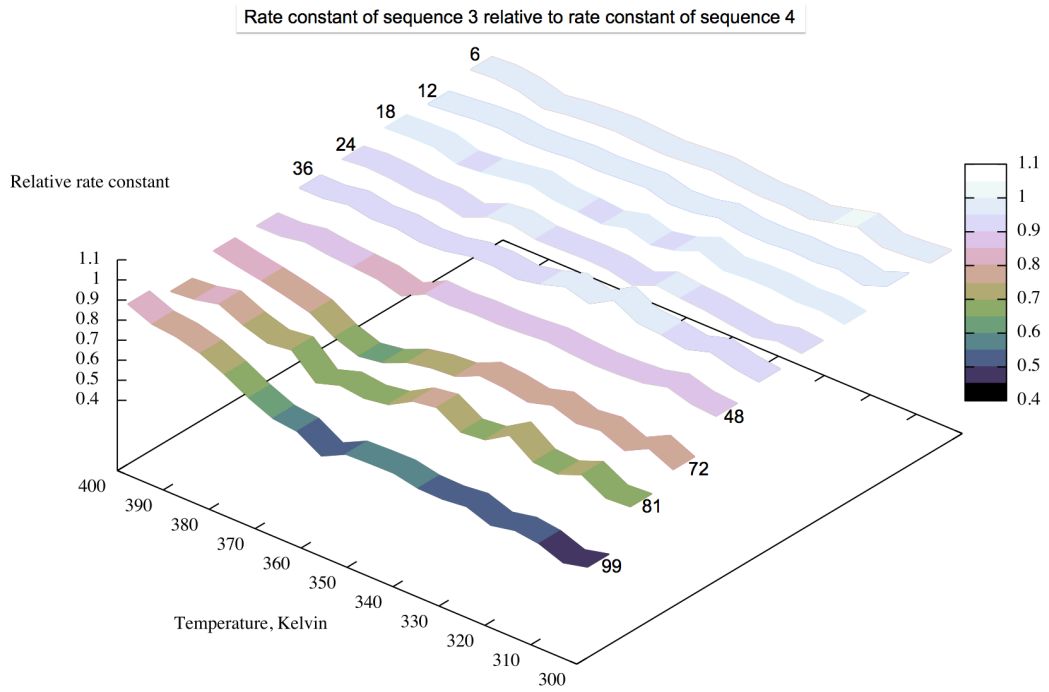


Figure 11: Rate constant of sequence 3 divided by rate constant of sequence 4 plotted as a function of temperature and chain length.

We can see from figure 11 that the rate constant of sequence 3 is lower than the rate constant of sequence 4 for all chain lengths and temperatures. However, at small chain lengths, the relative rate constant is only slightly below 1. For the 99 bps chain, we see that below 370 K, the rate constant of sequence 4 i much larger than the rate constant of sequence 3. However, this difference becomes smaller with increasing temperature between 360-400 K. The same trend is seen for chains of 81,72 and 48 bps, however, for these chains there is a valley occuring at 360-380 K, and the relative rate constant increases also towards lower temperatures. This increase in the rate constant with increasing temperature corresponds with the results found in figure 9 and 10, however, we see a less steep curve in figure 11, and the rate constant of sequence 3 never surpasses the rate constant of sequence 4.

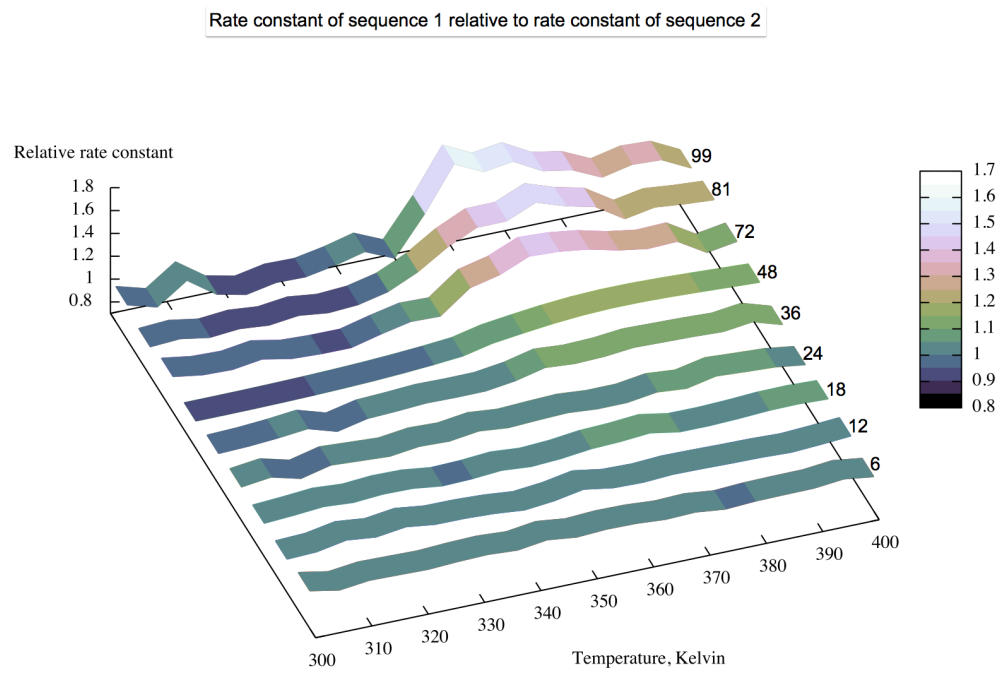


Figure 12: Rate constant of sequence 1 divided by rate constant of sequence 2 plotted as a function of temperature and chain length.

From figure 12 it can be seen that for short chains, the rate constant for sequence 1 is only slightly larger than the rate constant for sequence 2, and the relative rate constant is the same for all temperatures. For longer chains, at temperatures below 350 K the sequence 2 rate constant is actually slightly larger than the sequence 1 rate constant. However, for large chains above 355 K, the sequence 1 rate constant becomes much larger than the sequence 2 rate constant. The relative rate constant increases with increasing temperature between 350-360 K and peakes at about 360-370 K.

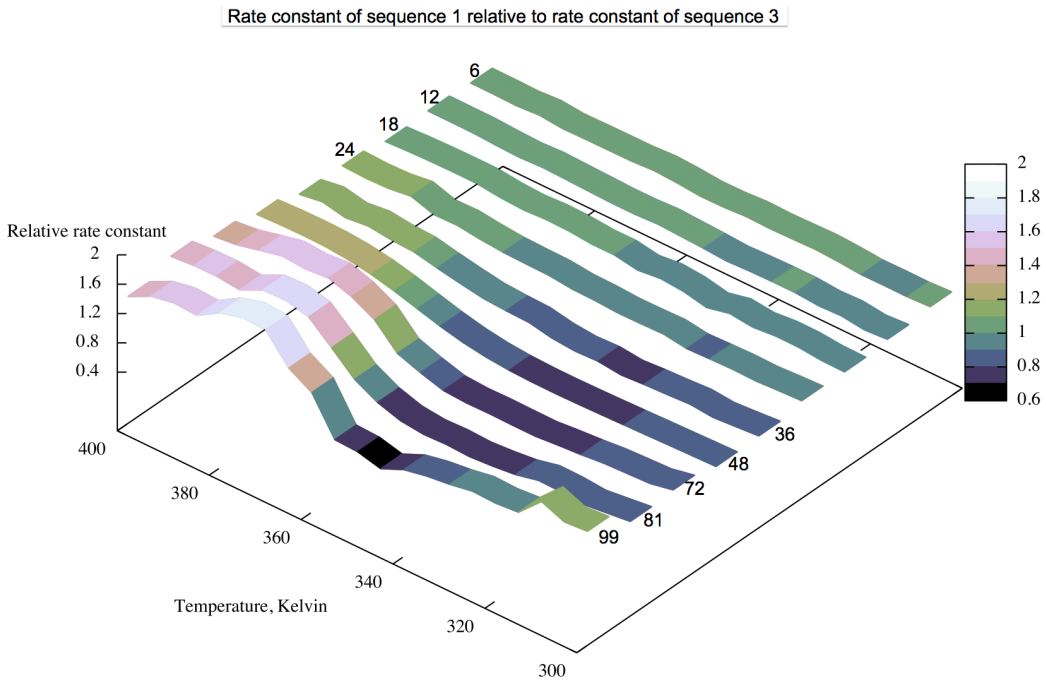


Figure 13: Rate constant of sequence 1 divided by rate constant of sequence 3 plotted as a function of temperature and chain length.

We see from figure 13, that for short chains, the rate constant of sequence 1 is more or less the same as the rate constant of sequence 3 for all temperatures. However, at longer chains, we see that for temperatures below 360 K, sequence 3 has a larger rate constant than sequence 1, with a valley occuring at 340-360 K, except from an upturn at 300-310 K for the 99 bps chain. In the temperature interval between 355-370 K, for large chains, the rate constant for sequence 1 grows larger than the rate constant for sequence 3, with a peak at 370-380 K.

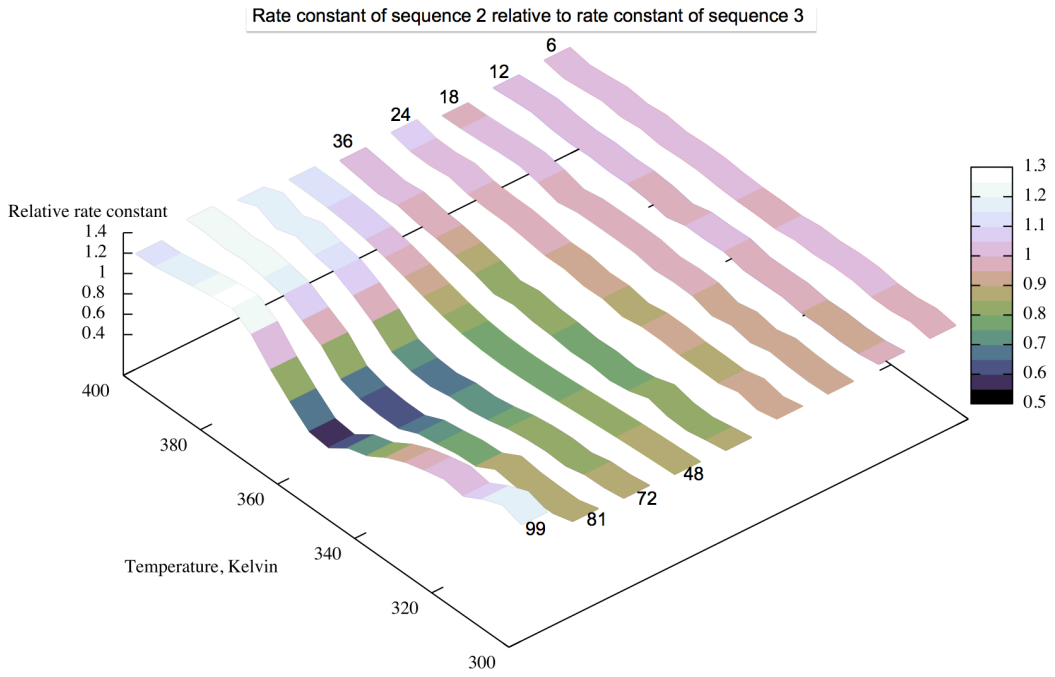


Figure 14: Rate constant of sequence 2 divided by rate constant of sequence 3 plotted as a function of temperature and chain length.

In figure 14, we see that for short chains, the rate constant of the two sequences is more or less the same. At longer chains, the rate constant of sequence 3 is larger at temperatures beelow 370 K, with a valley occuring at 350 K, except from an upturn at 305 K for the 99 bps chain. For long chains, there is a increase in the rate constant in the temperature interval at 360-370 K.

For all figures 9-14, we see that for long chains, all relative rate constants has a distinct increase with increasing temperature, between 355-375 K, which we assume is around the melting temperature of 33 % AT 66 % GC chains. There is also a increase in the relative rate constant with decreasing temperatures occuring between 350-310 K, however this increase is much less steep.
We also found the surprising result that sequence 4, the alternating sequence, does not consistently have a larger rate constant than the other sequences, which is contradictory to reference[22], where they found that for a 50 % AT 50 % GC chain, the alternating sequence without exception has a larger rate constant than the other sequences.

We further explored the rate constant of another sequences, which is a slight alteration of sequence 1. We denote the new sequence as sequence 5, and it consists of 11 AT bps at each end, and 11 AT bps in the middle of the chain, which divides the strong block in the middle. Thus, sequence 5 is A11G33A11G33A11. We investigated a chain of 99 bps having sequence 5 and compared it with a chain of 99 bps having sequence 1. In figure 15 the rate constant of sequence 1 and sequence 5 is shown. Inset shows the

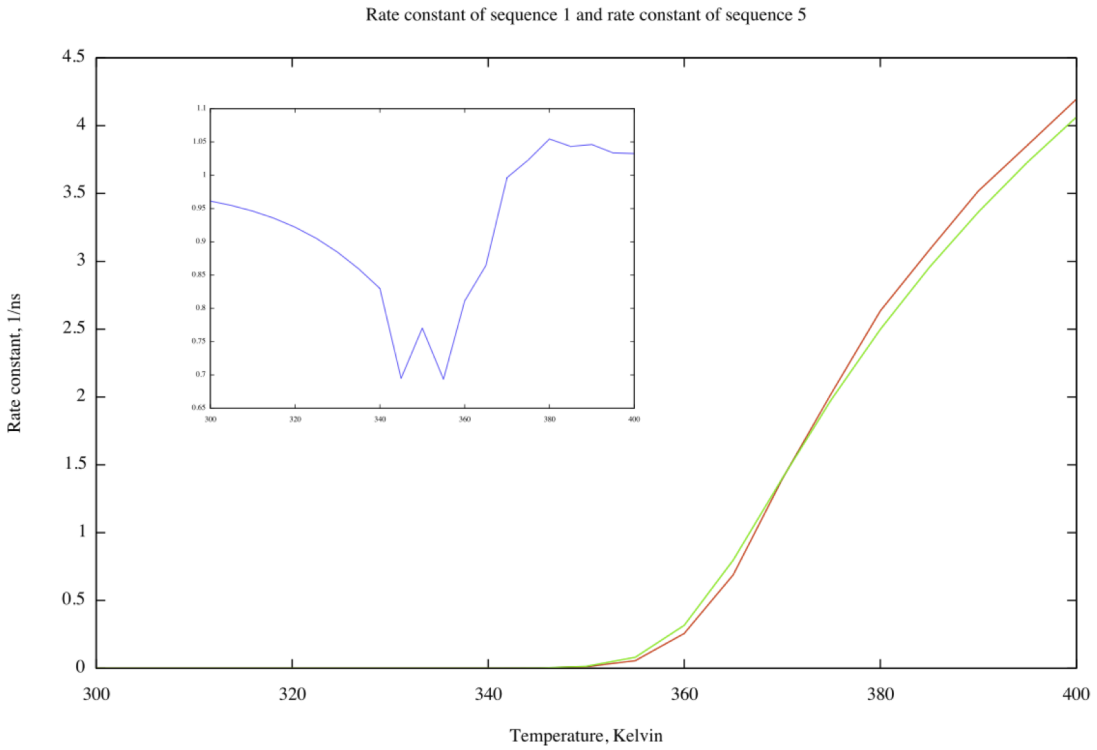


Figure 15: Rate constant of sequence 1 relative to sequence 5 plotted as a function of temperature, DNA chains of 99 bps. The curves are plotted with the same transmission coefficient=0.02. Inset shows the rate constant of sequence 1 relative to sequence 5.

relative rate constant, the rate constant of the sequence 1 divided by the rate constant of sequence 5. We see that below approximately 370 K, sequence 5 has a larger rate constant, with the largest difference at 345-355 K. Above 370 K, sequence 1 has a larger rate constant. Dividing up the strong GC-block should in principle ease the denaturation, but it seems that having AT bps at each end of the chain becomes more important than dividing up the strong block at temperatures above 375 K.

4.2 Investigation of DNA hairpins

We looked at three published articles where the rate constant of hairpins has been experimentally obtained, and we tried to reproduce these results by simulations applying the PBD model. Then, we compared our results with the experimental results to see how the PBD model results corresponds with the experimental results.

Cutoff value for both parameter sets To decide which cutoff values we should apply for the hairpin potential, we investigated how sensitive the PBD model is to the cutoff value for the two different parameters sets, respectively. We divided the rate constant for each cutoff-value with the rate constant for the ds DNA. The ds DNA is a chain having the same length and sequence as the hairpin stem. We found approximately the same results for all hairpins investigated, all included in appendix A. We present here an example graph to see the general trend. From figure 16, we can see two important aspects:

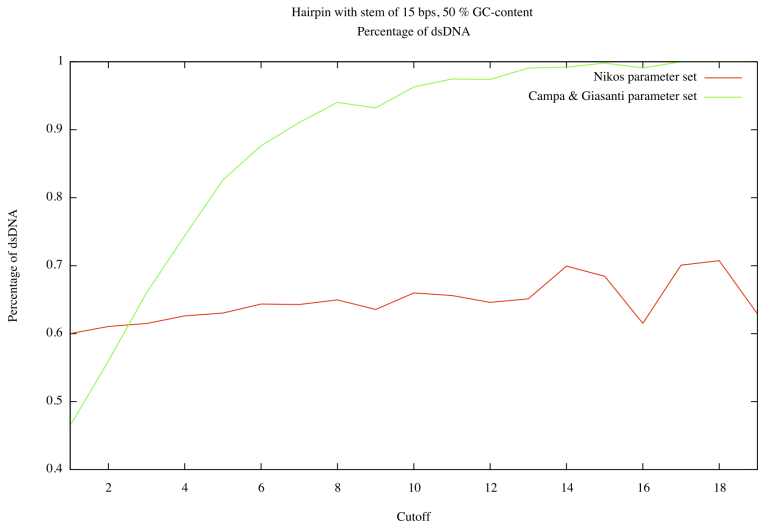


Figure 16: Percentage of dsDNA, 15 bps stem, 50 % GC content, loop of T4

First of all, that the Campa and Giansanti rates increase much more rapidly than the Theodorokopolous rates in the range of cutoff values below 10 Å. The cutoff value has an larger impact on the rate constant, seen from the steep line showing that a small change in cutoff results in a large change in rate constant, and at cutoff 9 Å , the rate constant for the PBD Campa and Giansanti results has almost reached the rate constant for the dsDNA. All simulations with cutoffs above 9 would then give more or less the same rate constant. From the PBD Theodorakopoulos results, we see that the rate constant only increases slightly with increasing cutoff value. For the Theodorakopoulos parameter set, the rate constant reaches about 80 % of the dsDNA value at cutoff 40 Å , and for the cutoffs we operated with usually(1-20 Å), it was about 50-70 % of the dsDNA value.
From this point, for all hairpins investigated, we applied cutoffvalues ranging from 1-9

Å for the Campa and Giansanti parameter set, and 1-19 Å for the Theodorakopoulos parameter set, except from in section 4.2.4, where we explored cutoffvvalues up till 90 Å for the Theodorakopoulos parameter set.

Hairpins with altering size of the stem and the loop, and varying GC-content of the stem We examined the hairpins from reference [31]. Here, the authors performed mechanical experiments in an optical trap using a force clamp arrangement on 20 different hairpins with various stem lengths, loop lengths and GC versus AT content of the stem. A summary of all details of the hairpins are given in appendix B. All experiments were performed at 23 degrees Celsius (296 Kelvin), and rate constants were obtained with a certain error (see table in appendix B). We applied the PBD model to obtain rate constants for these hairpins, applying cutoff values from 1 to 9 Å for simulations using the Campa and Giansanti parameter set, and 1 to 19 Å for the Theodorakopoulos parameter set We compare the PBD model with the experiments by dividing the rate constant obtained from the PBD model with the experimentally obtained rate constant. For each hairpin, we present two errorbars, one for cutoff min(1) and one for a cutoff max (9 Å for Campa and Giansanti and 19 Å for Theodorakopoulos). The experimental values are given with a certain error interval, and the errorbars thus shows how much larger the PBD model results are than the experimental mid value, max value and min value. When there are very large differences in the results in the same plot, we present the results with a logarithmic scale on the y-axis. The results are presented in three blocks: First, we present the results of hairpins with a stem of 15 bps with a 60 % GC-content with an increasing loop size consisting solely of T's. Next, we look at hairpins with a stem of 20 bps, all having a loop of T4 (Four T-bases), with increasing GC-content of the hairpin stem. Following, we look at hairpins of with a stem with 50 % GC-content and loop of T4 with increasing stem length.

4.2.1 Hairpins with a 15 bps long stem of 60 % GC-content with increasing loop size

Here, the authours of reference [31] performed experiments on hairpins having a 15 bps long stem of 60 % GC-content, with 7 different sizes of the loop. As mentioned in section 2.12, in our model we apply a potential to create a hairpin effect, and we can not say anything specific about the characteristics of the loop. Thus, we simulated a hairpin having this particular stem, and compared the rate constant obtained for this single hairpin with the experimentally obtained rate constant of the seven hairpins, respectively.

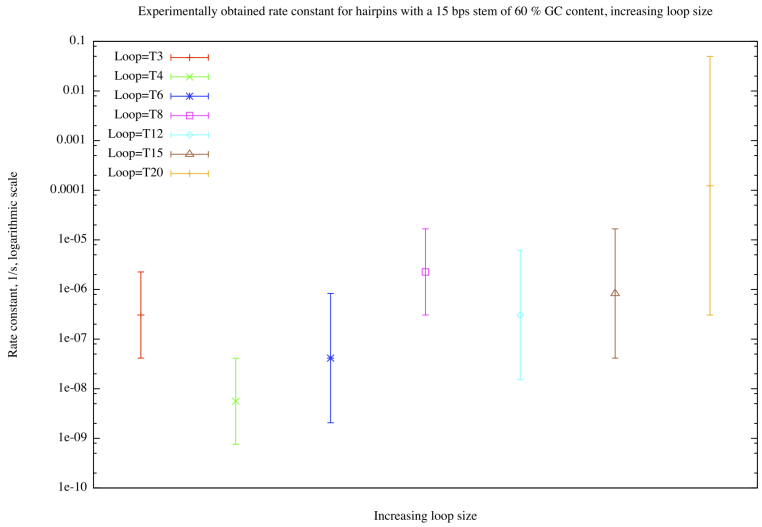


Figure 17: Plot of experimentally obtained rate constant for hairpins with a 15 bps stem of 60 % GC content, with increasing loop size.

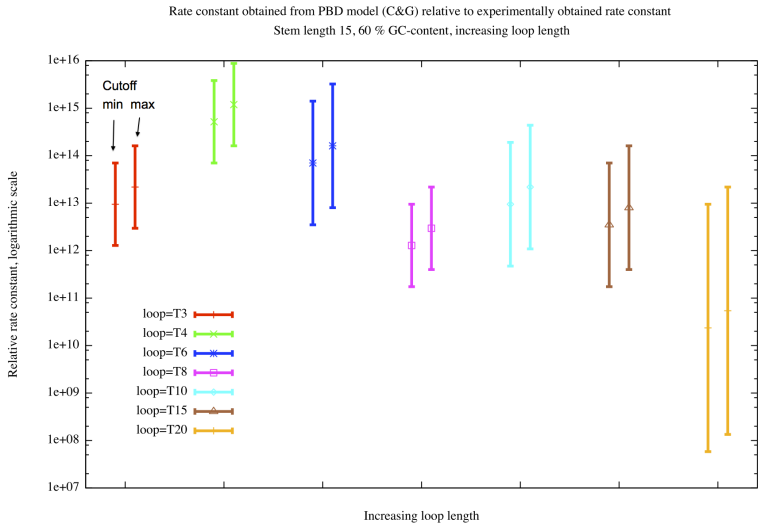


Figure 18: Plot of relative rate constant, the PBD rate constant relative to experimental rate constant. Hairpins with a 15 bps stem of 60 % GC content, with increasing loop size, Campa and Giansanti parameter set.

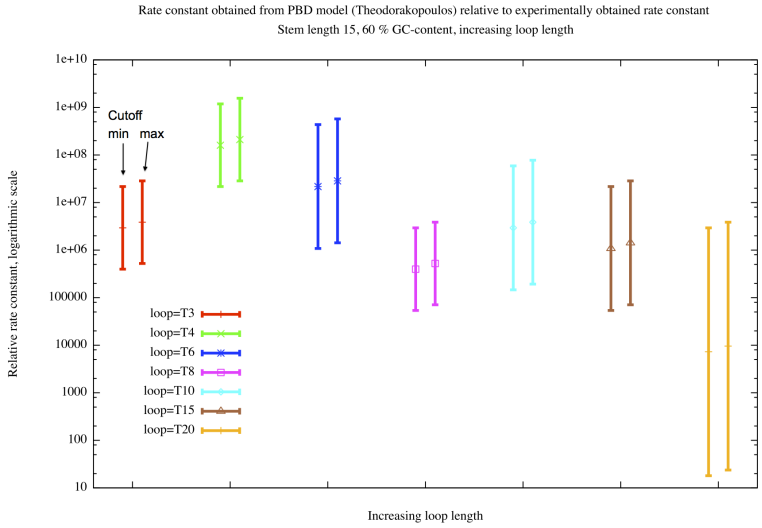


Figure 19: Plot of relative rate constant, the PBD rate constant relative to experimental rate constant. Hairpins with a 15 bps stem of 60 % GC content, with increasing loop size, Theodorakopoulos parameter set

From figures 18-20 we can see that the PBD model gives denaturation rates that are orders of magnitude larger than the experimental ones that, on the other hand, bear huge error bars. There seems no trend for the results with increasing loop size, so we

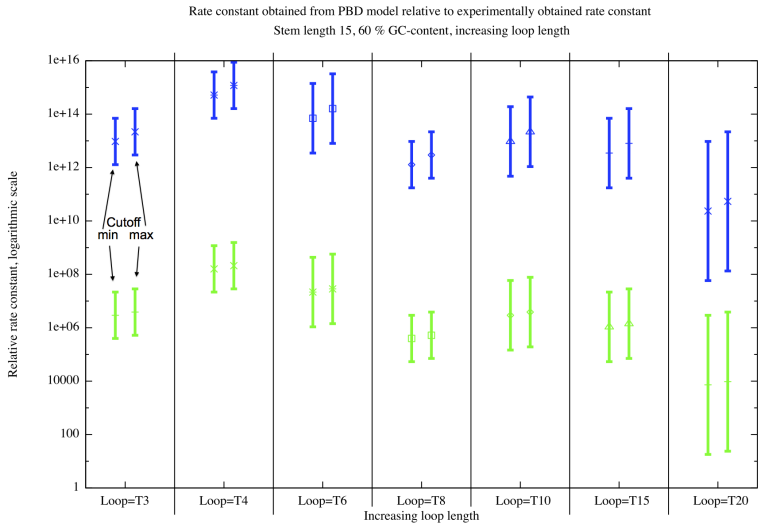


Figure 20: Here we see the results in figure18 and **??** together. Plot of relative rate constant, the PBD rate constant relative to experimental rate constant. Hairpins with a 15 bps stem of 60 % GC content, with increasing loop size. Here, the variation in the relative rates is only due to the experimental values seen in figure 17 since the theoretical rates are independent to the loop.

can conclude that the PBD model results does not fit better or worse with smaller or larger loop size. The average experimental opening rates differ for different loop sizes by factors upto 10^4. Though, a statement on the true dependence of opening rates and loop size is difficult to make since the experimental error bars are almost of the same order. It is a fact, though, that the closing rate is much more dependent to the loop size. Reference [30] therefore argue that the characteristics of the hairpin loop affects the opening rate constant only modestly. For all the subsequent experimental data, they were all performed with a loop of T4, and we can assume that these results might be representative for larger loops as well.

4.2.2 Hairpins with a 20 bps long stem of various GC content, all with loop T4

Here, the authors of reference[22] investigated six hairpins with a stem of 20 bps having a GC-content of 0, 25, 50, 55, 75 and 100 %, respectively. All hairpins had a loop of T4. From figure 21, we see that the PBD model results are order of magnitudes larger than

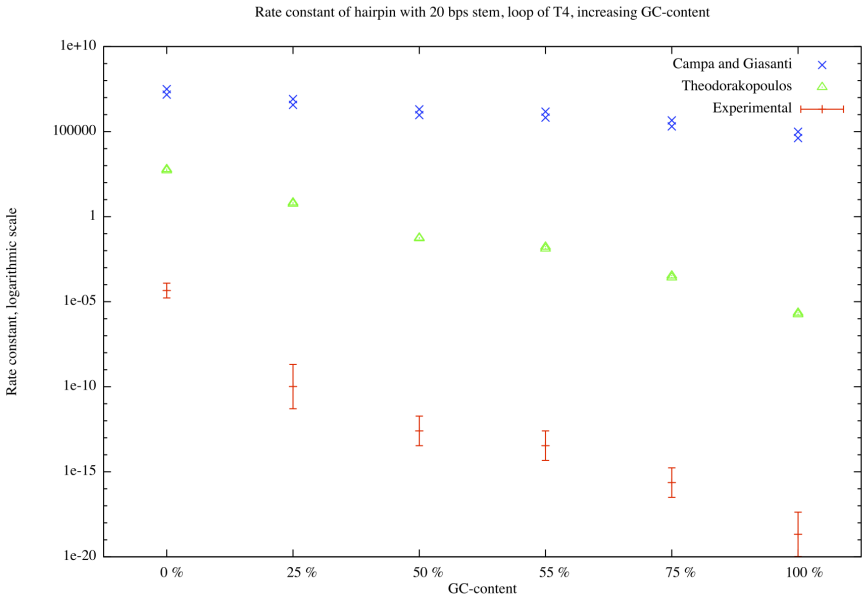


Figure 21: Plot of rate constant in units $1/s$ as a function of increasing GC content of stem for of hairpins with 20 bps stem and a loop of T4. PBD model results with Campa and Giansanti parameters, Theodorakopoulos parameters and experimental values.

the experimental results, for both parameter sets. Exactly how much larger the PBD results are than the experimental results is shown in figure 22-24. The PBD results from using the Campa and Giansanti parameter set is also much larger than the PBD results from using the Theodorakopoulos parameter set, with a factor of $3 \cdot 10^4$ for the hairpin with a pure AT stem, and $4 \cdot 10^{10}$ for the hairpin with a pure GC stem.

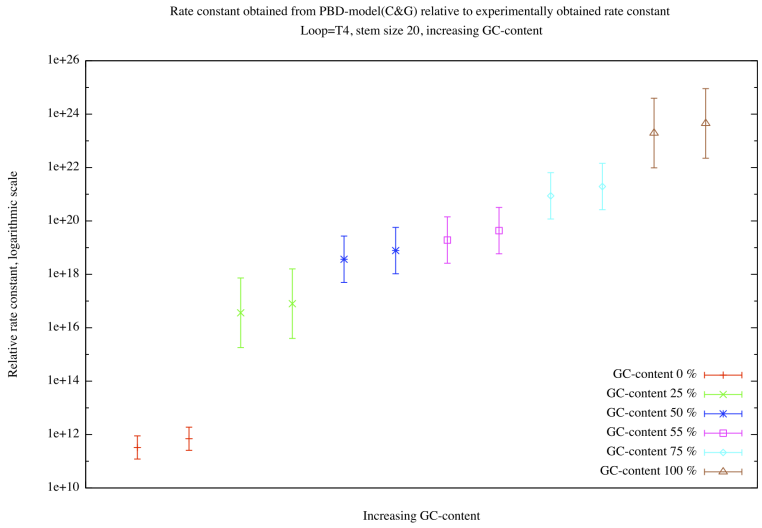


Figure 22: Plot of relative rate constant, the PBD rate constant relative to experimental rate constant as a function of increasing GC content of stem. Hairpins with a 20 bps stem and a loop of T4. Results from Campa and Giansanti parameter set

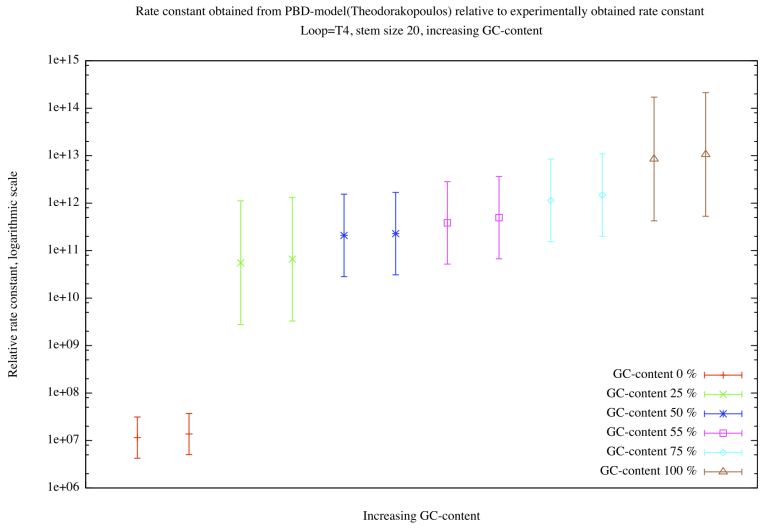


Figure 23: Plot of relative rate constant, the PBD rate constant relative to experimental rate constant as a function of increasing GC content of stem. Hairpins with a 20 bps stem and a loop of T4. Results from Theodorakopoulos parameter set

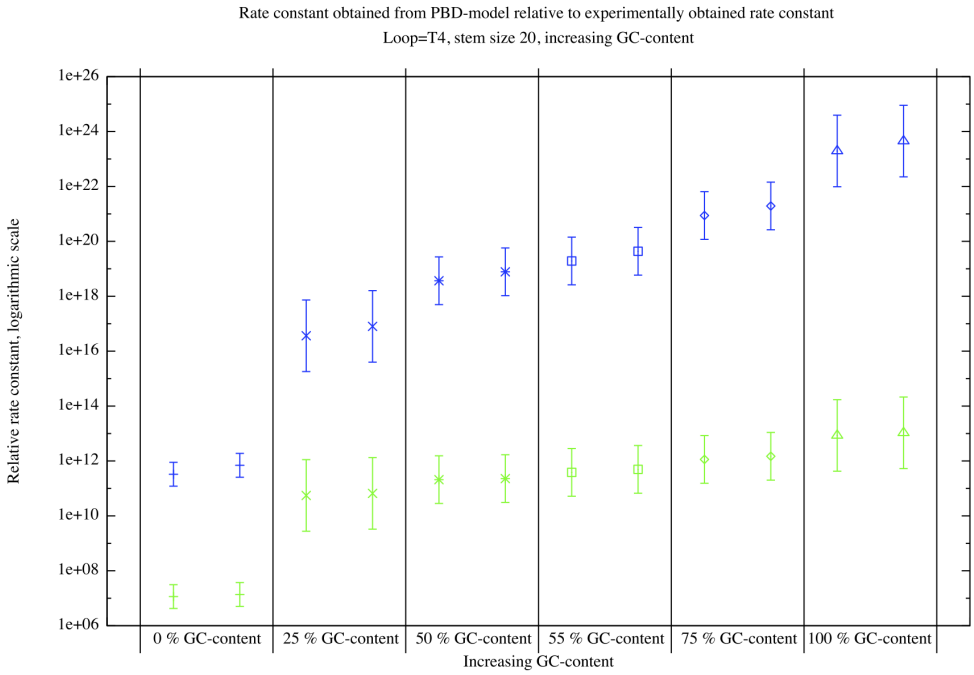


Figure 24: Here the results from figure22 and 23 are presented together. Plot of relative rate constant, the PBD rate constant relative to experimental rate constant as a function of increasing GC content of stem. Hairpins with a 20 bps stem and a loop of T4.

GC-content	Cutoff	PBD model results relative to Experimental results, Campa and Giansanti	PBD model results relative to experimental results, Theodorakopoulos
0 %	min	$1.2 \cdot 10^{11} - 8.9 \cdot 10^{11}$	$4.2 \cdot 10^{6} - 3.1 \cdot 10^{7}$
0 %	max	$2.6 \cdot 10^{11} - 1.9 \cdot 10^{12}$	$5.0 \cdot 10^{6} - 3.7 \cdot 10^{7}$
25 %	min	$1.8 \cdot 10^{15} - 7.3 \cdot 10^{17}$	$2.8 \cdot 10^{9} - 1.1 \cdot 10^{12}$
25 %	max	$4.0 \cdot 10^{15} - 1.6 \cdot 10^{18}$	$3.3 \cdot 10^{9} - 1.3 \cdot 10^{12}$
50 %	min	$5.0 \cdot 10^{17} - 2.7 \cdot 10^{19}$	$2.8 \cdot 10^{10} - 1.5 \cdot 10^{12}$
50 %	max	$1.1 \cdot 10^{18} - 5.8 \cdot 10^{19}$	$3.0 \cdot 10^{10} - 1.7 \cdot 10^{12}$
55 %	min	$2.6 \cdot 10^{18} - 1.4 \cdot 10^{20}$	$5.2 \cdot 10^{10} - 2.8 \cdot 10^{12}$
55 %	max	$5.9 \cdot 10^{18} - 3.2 \cdot 10^{20}$	$6.7 \cdot 10^{10} - 3.7 \cdot 10^{12}$
75 %	min	$1.2 \cdot 10^{20} - 6.4 \cdot 10^{21}$	$1.5 \cdot 10^{11} - 8.4 \cdot 10^{12}$
75 %	max	$2.6 \cdot 10^{20} - 1.4 \cdot 10^{22}$	$2.0 \cdot 10^{11} - 1.1 \cdot 10^{13}$
100 %	min	$9.8 \cdot 10^{21} - 3.9 \cdot 10^{24}$	$4.2 \cdot 10^{11} - 1.7 \cdot 10^{14}$
100 %	max	$2.2 \cdot 10^{22} - 9.0 \cdot 10^{24}$	$5.3 \cdot 10^{11} - 2.1 \cdot 10^{14}$

Table 1: Summary of figure 24.

From figures 22-24 and summarized in table 1 we see how much larger the PBD results are then the experimental results. The obvious trend is that the relative rates increases with increasing GC-content of the stem. The PBD results with the Campa and Giansanti parameter set is $10^{11} - 10^{12}$ times larger for a pure AT-stem, and $10^{21} - 10^{24}$ times larger for a pure GC-stem. For the and Theodorakopoulos parameter set, the PBD results are $10^{6} - 10^{7}$ times larger for a pure AT-stem, and $10^{11} - 10^{14}$ times larger for a pure GC-stem.

Further on, we calculated the decrease in rate constant with increasing GC content, that is, how much the rate constant decrease going from the hairpin with a stem of 0 % GC content to the hairpin with a stem of 25 % GC content, from 25 % to 50 % , from 50 % to 55 %, from 50 % to 75 % and from 75 % to 100 %. We chose to show the decrease from 50 % to 75 % rather than 55 % to 75 % for the reason that we could see the decrease for each increase of 25 % GC content of the stem. We present this decrease for the experimental results and the PBD results for both parameter sets. As the experimental results were given with a certain error, the decrease in rate constant is also given with a certain error, and these values are represented with an error bar.

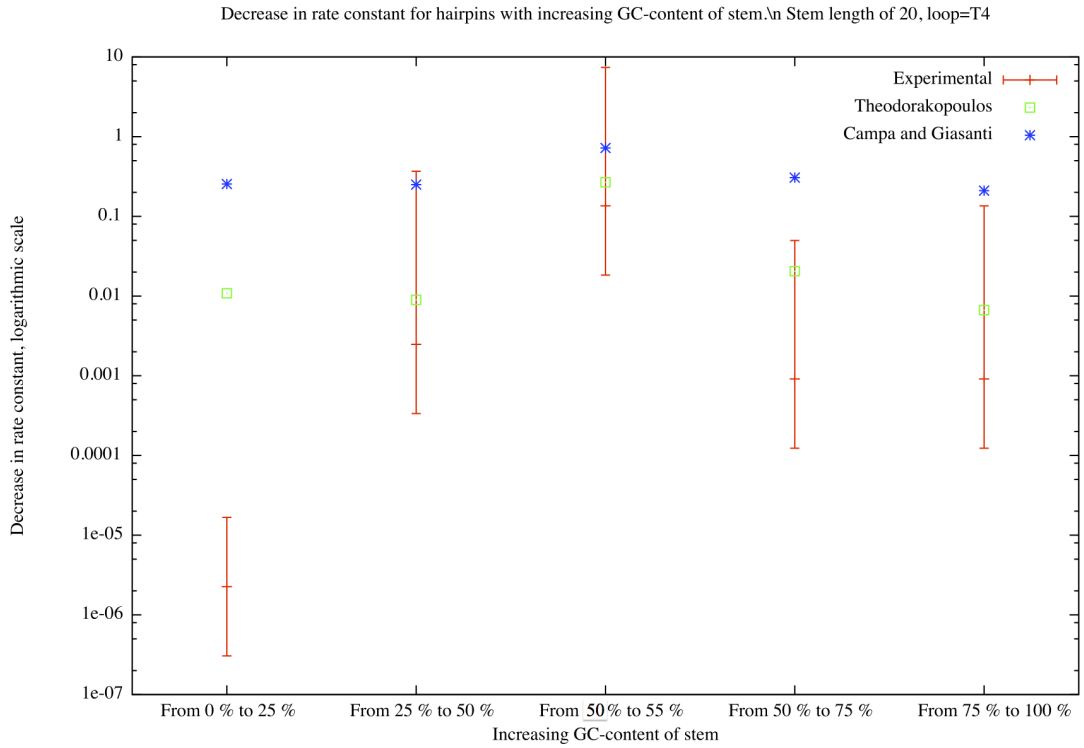


Figure 25: Decrease in rate constant plotted as a function of increase in GC content of stem. PBD results with Campa and Giansanti parameter set shown in blue, PBD results with the Theodorakopoulos parameter set in green and experimental results in red. Hairpins with a 20 bps stem and loop of T4.

We see from figure 25, that when the GC-content increases, the experimental results decreases faster than our results does, and the PBD Theodorakopoulos results decreases faster than the PBD Campa and Giansanti results. We see that for the decrease from 0 % to 25 % GC content, the decrease in experimental result is magnitudes of order larger for than for the decrease in the PBD results. For the subsequent calculations of decrease, the PBD results with Theodorakopoulos parameters lies within the decrease of the experimental results. PBD results with Campa and Giansanti parameters lies within the experimental decrease for 25-50 % and 50-55 %, and slightly above the experimental decrease for 50-75 % and 75-100 %.

4.2.3 Hairpins with 50 % GC-content of stem, increasing stem length, all with loop T4

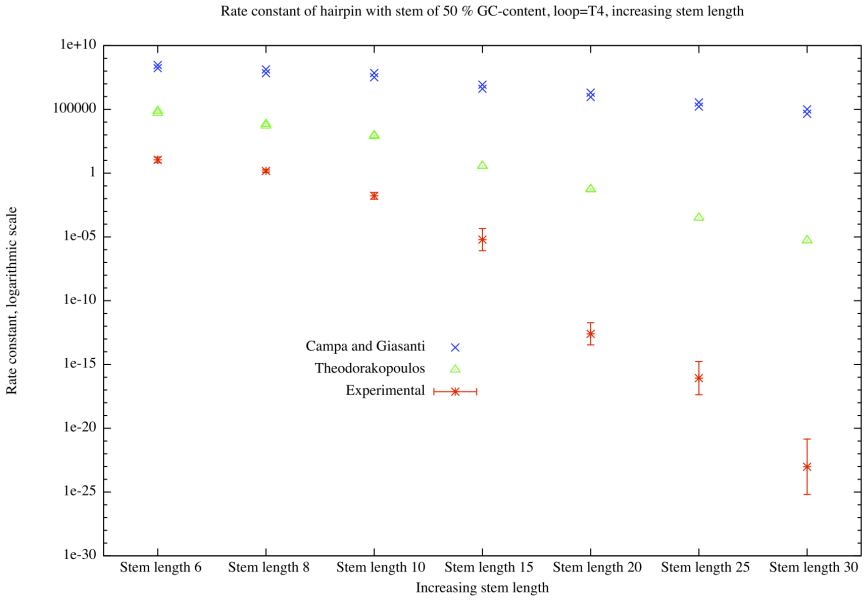


Figure 26: Plot of rate constant in units $1/s$ as a function of increasing stem length for of hairpins with stem of 50 % GC-content, and loop of T4. Experimental results seen in red, PBD model results with Campa and Giansanti parameter seen in blue and PBD model results with Theodorakopoulos parameters seen in green.

From figure 26, it can be seen that for both parameter sets, the results from the PBD model are significantly larger than the experimental results. The relative rates will be presented in figure 27-29. The PBD results from using the Campa and Giansanti parameter set is magnitude of orders larger than the PBD results from using the Theodorakopoulos parameter set, with a factor $3 \cdot 10^3$ for the hairpin with a 6 bps stem, and a factor of $2 \cdot 10^{10}$ for the hairpin with a 30 bps stem.

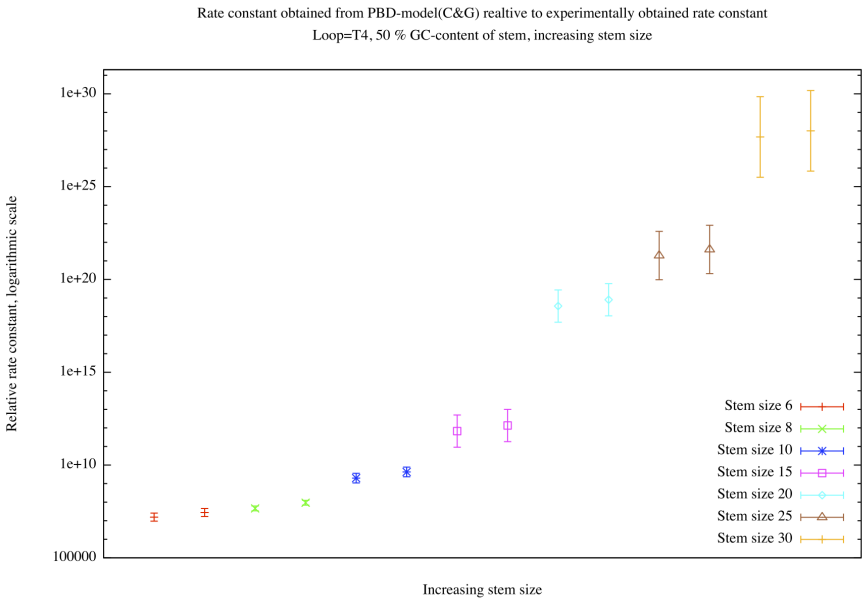


Figure 27: Plot of relative rate constant, the PBD rate constant relative to experimental rate constant as a function of increasing stem size. Hairpins with a stem of 50 % GC content, loop of T4. Campa parameter set

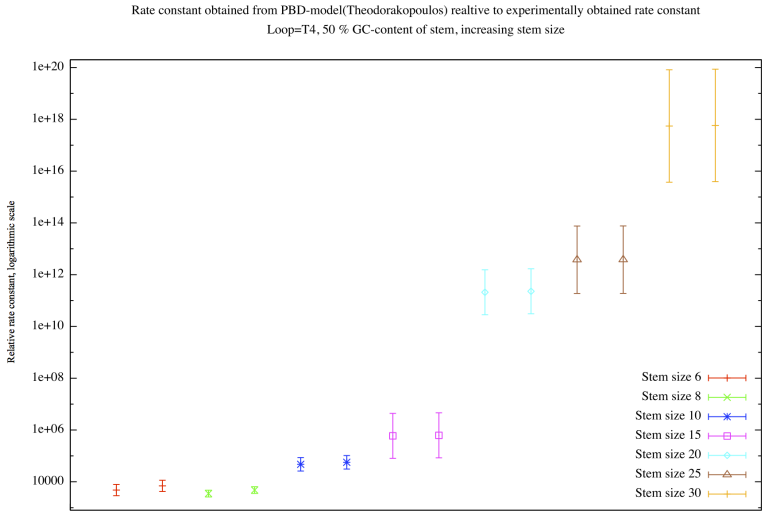


Figure 28: Plot of relative rate constant, the PBD rate constant relative to experimental rate constant as a function of increasing stem size. Hairpins with a stem of 50 % GC content, loop of T4, Theodorakopoulos parameter set

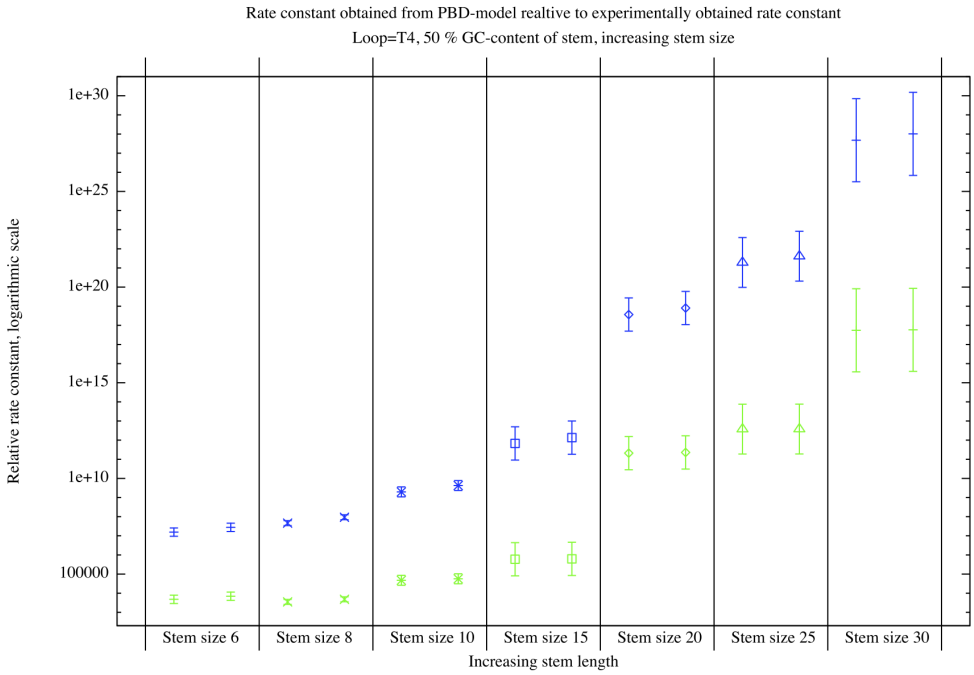


Figure 29: Here, the results from figure 27 and 28 are shown together. Plot of relative rate constant, the PBD rate constant relative to experimental rate constant as a function of increasing stem size. Hairpins with a stem of 50 % GC-content, loop of T4.

Stem length	Cutoff	PBD model results relative to Experimental results, Campa and Giansanti	PBD model results relative to experimental results, Theodorakopoulos
6	min	$9.5 \cdot 10^6 - 2.6 \cdot 10^7$	$2.9 \cdot 10^3 - 7.9 \cdot 10^3$
6	max	$1.7 \cdot 10^7 - 4.6 \cdot 10^7$	$4.2 \cdot 10^3 - 1.1 \cdot 10^4$
8	min	$3.4 \cdot 10^7 - 6.2 \cdot 10^7$	$2.6 \cdot 10^3 - 4.7 \cdot 10^3$
8	max	$6.9 \cdot 10^7 - 1.3 \cdot 10^8$	$3.6 \cdot 10^3 - 6.5 \cdot 10^3$
10	min	$1.1 \cdot 10^9 - 3.6 \cdot 10^9$	$2.6 \cdot 10^4 - 8.6 \cdot 10^4$
10	max	$2.3 \cdot 10^9 - 7.7 \cdot 10^9$	$3.1 \cdot 10^4 - 1.0 \cdot 10^5$
15	min	$9.1 \cdot 10^{10} - 5.0 \cdot 10^{12}$	$8.0 \cdot 10^4 - 4.4 \cdot 10^5$
15	max	$1.8 \cdot 10^{11} - 1.0 \cdot 10^{13}$	$8.4 \cdot 10^4 - 4.6 \cdot 10^5$
20	min	$5.0 \cdot 10^{17} - 2.7 \cdot 10^{19}$	$2.8 \cdot 10^{10} - 1.5 \cdot 10^{12}$
20	max	$1.1 \cdot 10^{18} - 5.9 \cdot 10^{19}$	$3.1 \cdot 10^{10} - 1.7 \cdot 10^{12}$
25	min	$9.7 \cdot 10^{19} - 3.9 \cdot 10^{22}$	$1.9 \cdot 10^{11} - 7.6 \cdot 10^{13}$
25	max	$2.1 \cdot 10^{20} - 8.3 \cdot 10^{22}$	$1.9 \cdot 10^{11} - 7.6 \cdot 10^{13}$
30	min	$3.2 \cdot 10^{25} - 7.0 \cdot 10^{29}$	$3.7 \cdot 10^{15} - 8.2 \cdot 10^{19}$
30	max	$6.8 \cdot 10^{25} - 1.5 \cdot 10^{30}$	$3.9 \cdot 10^{15} - 8.6 \cdot 10^{19}$

Table 2: Summary of figure 29.

From figures 27-29 and summarized in table 2, it can be seen we that the PBD results realtive to the experimental results increases with increasing stem length. The PBD results with the Campa and Giansanti parameter set is $10^6 - 10^7$ times larger for a stem of 6 bps, and $10^{25} - 10^{30}$ times larger for a stem of 30 bps. For the Theodorakopoulos parameter set, the PBD results are $10^3 - 10^4$ times larger for a stem of 6 bps, and $10^{15} - 10^{19}$ times larger for a stem of 30 bps.

Further on, we calculated the decrease in rate constant with increasing stem length, similar to figure 25, that is, how much does the rate constant decrease when going from the hairpin with a stem of 6 bps to the hairpin with a stem of 8 bps, from 8 bps to 10 bps and so on.

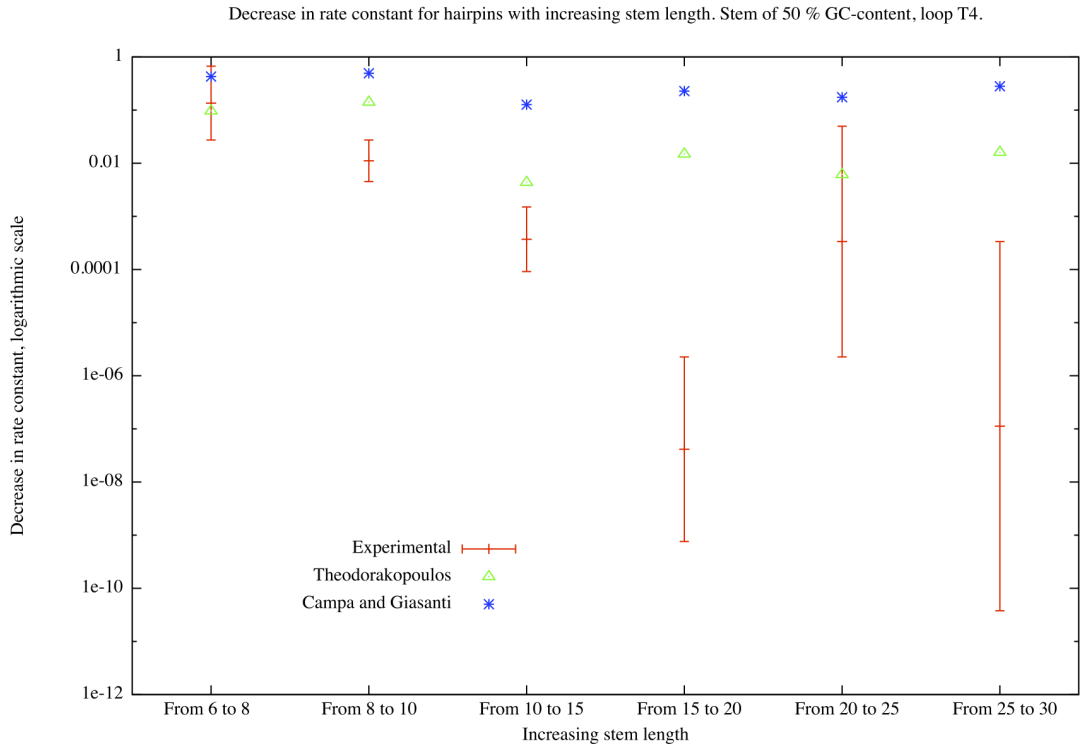


Figure 30: Decrease in rate constant as a function of increasing stem length. Campa and Giansanti, Theodorakopoulos and experimental results. Stems of 50 % GC-content, loop of T4.

From figure 30 we see the same trend as in figure 25, that the PBD Theodorakopoulos results decreases faster than the PBD Campa and Giansanti results, and the experimental results in general decreases faster than both parameter sets. We clearly see that the experimental values decreases irregulary with the PBD model results, and there can not be extracted any clear trend from figure 30.

4.2.4 Hairpins with a 10 bps stem of 60 % and 70 % GC content, 20 bps loop of 55 % GC content

In reference [29], the authors performed mechanical tension experiments on two hairpins with a 10 bps stem length and obtained rate constants for these two hairpins. Hairpin 1 has a stem of 60 % GC-content, with the stem sequence GAAGAGGGAG. The second hairpin has a 70 % GC-content with the stem sequence GAAGAGGGGG. The loop of these hairpins was the same, 20 bps long, with the sequence GGGGAGAAAGAGAGAAAGAA. This loop has a GC content of 55 %.
The experiments were performed in room temperature.
We obtained rate constants for a range of cutoffs. For the Campa & Giansanti parameter set, we used cutoffs from 1 to 10 Å , while for the Theodorakopoulos parameter set, we used cutoff values from 1 to 90 Å . The reason we used such large cutoff value for the Theodorakopoulos parameter set, was because, for these two hairpins, there were only a few simulations that needed to be done, and thus, we had the time and computational space available to explore such large cutoff-values. By doing this, we found results simular to 16. For the Campa and Giansanti parameter set, the expontential potential has no effect above cutoff 10, as the rate constant is then equal to the dsDNA rate constant, and for the Theodorakopoulos parameter set, the expontential potential has an effect up till cutoff 90, as the rate constant reaches the dsDNA rate constant at cutoff 90.

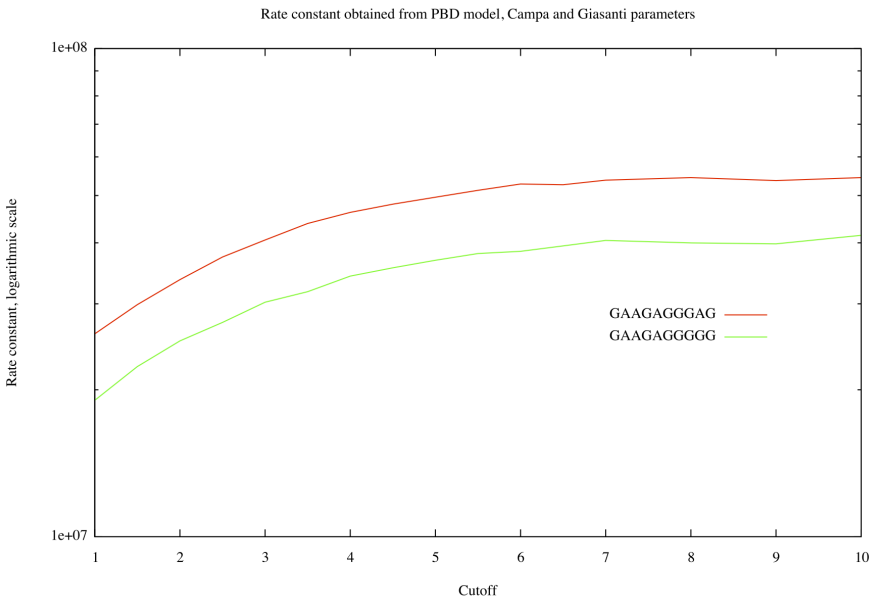


Figure 31: Plot of rate constant in units $1/s$ obtained from the PBD model as a function of the cutoff value, Campa and Giansanti parameters

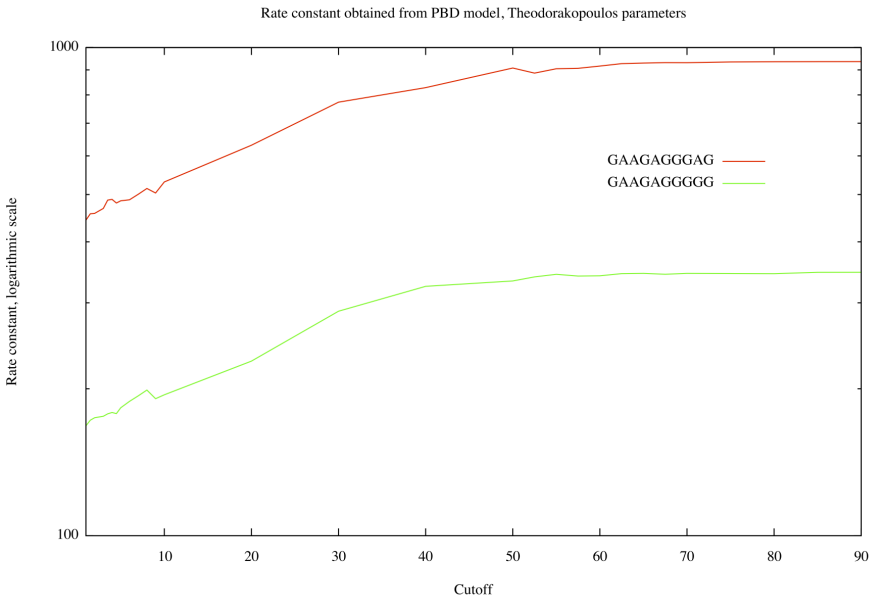
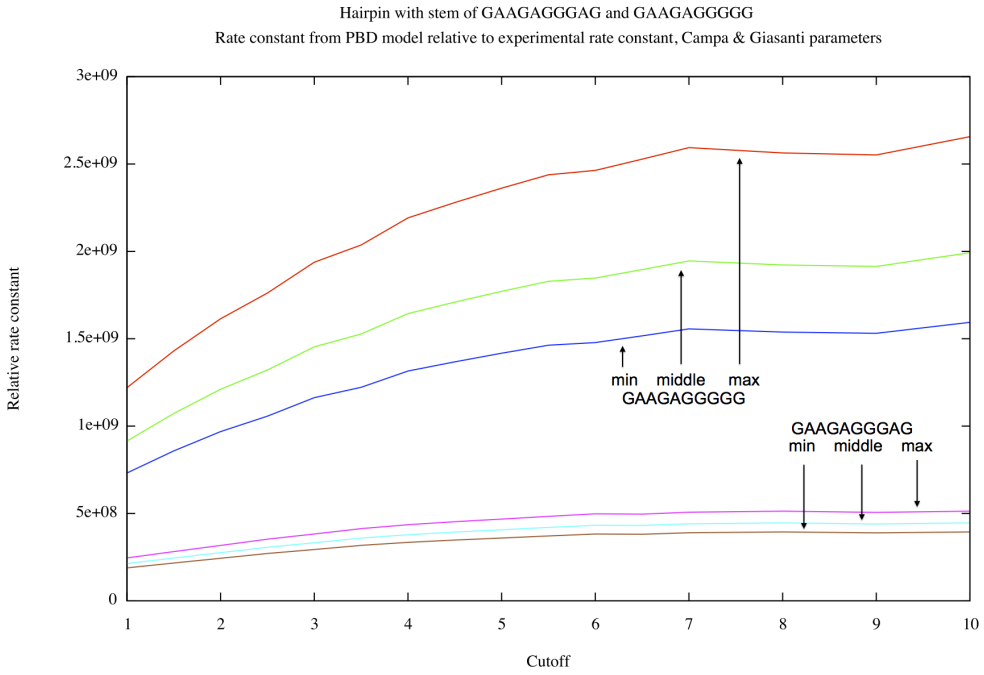
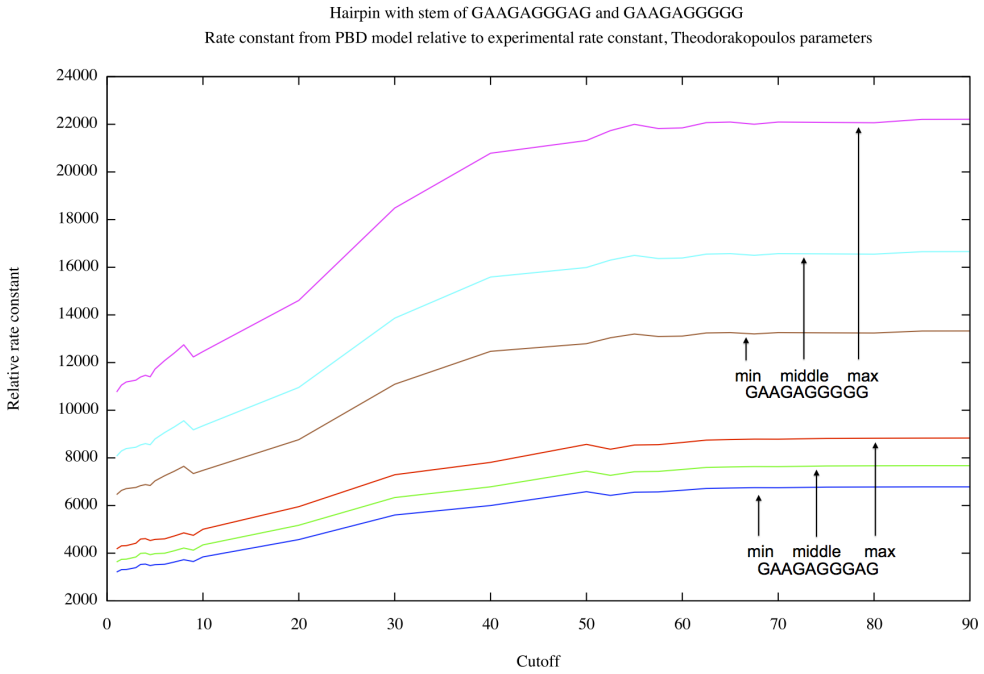


Figure 32: Plot of rate constant in units $1/s$ obtained from the PBD model as a function of the cutoff value, Theodorakopoulos parameters

For the two hairpins with a stem of GAAGAGGGAG and GAAGAGGGGG, the opening rates for zero applied force was experimentally found to be $r = 1.16 \pm 0.14 \cdot 10^{-1} s^{-1}$ and $r = 2.08 \pm 0.52 \cdot 10^{-2} s^{-1}$, respectively.
Thus, they found that the rate constant for the GAAGAGGGAG hairpin is $5.3 - 6.8$ times larger than the rate constant for GAAGAGGGGG.
From the two figures 31 and 32, we can see that as in the experimental results, the rate constant for GAAGAGGGAG is larger than the rate constant for the GAAGAGGGGG. However, the difference is smaller, approximately 1.35 for the Campa & Giansanti parameter set, and 2.7 for the Theodorakopoulos parameter set. The hairpins differ in the 9th base pair of the stem, so they differ by one base pair. As we know that AT bps have one less hydrogen bonds and thus opens up easier than GC bps, it is expected that the rate constant for GAAGAGGGAG is larger than the rate constant for GAAGAGGGGG. In section 2.3 we mentioned some of the limitations of the PBD model, among others that the stacking potential does not distinguish between AT and GC base pairs. Thus, it can be that the PBD-model does not differ adequately between these two hairpins. This might be a possible explanation for why the PBD model finds a smaller difference between these two hairpins than found in the experiments.



(a) Campa and Giansanti parameter set



(b) Theodorakopoulos parameter set

Figure 33: Rate constants from PBD model divided by the experimental rate constants, for hairpins GAAGAGGGAG and GAAGAGGGGG

Hairpin stem	Experimental value		Relative rate constant, Campa & Giansanti	Relative rate constant, Theodorakopoulos
GAAGAGGGAG	min	0.106	$2.5 \cdot 10^8 - 5.1 \cdot 10^8$	$4.2 \cdot 10^3 - 8.8 \cdot 10^3$
GAAGAGGGAG	middle	0.122	$2.1 \cdot 10^8 - 4.5 \cdot 10^8$	$3.6 \cdot 10^3 - 7.7 \cdot 10^3$
GAAGAGGGAG	max	0.138	$1.9 \cdot 10^8 - 3.9 \cdot 10^8$	$3.2 \cdot 10^3 - 6.8 \cdot 10^3$
GAAGAGGGGG	min	0.0156	$1.2 \cdot 10^9 - 2.7 \cdot 10^9$	$1.1 \cdot 10^4 - 2.2 \cdot 10^4$
GAAGAGGGGG	middle	0.0208	$9.1 \cdot 10^8 - 2 \cdot 10^9$	$8.1 \cdot 10^3 - 1.7 \cdot 10^4$
GAAGAGGGGG	max	0.026	$7.3 \cdot 10^8 - 1.6 \cdot 10^9$	$6.5 \cdot 10^3 - 1.3 \cdot 10^4$

Table 3: Table showing how much larger the rate constants obtained from PBD model results are than the rate constant obtained from experimental results, i.e. the relative rate constant, for both hairpins. Shows the relative rate constant for both experimental max, middle and min values. These numbers are a summary of figure 33a and figure 33b.

From figure 33a and 33b, and summarized in table 3 we see that for the Campa and Giansanti parameter set, the rate constant for the PBD model compared to the experimental rate constants is $1.9 \cdot 10^8 - 5.1 \cdot 10^8$ times larger for the GAAGAGGGAG hairpin, and $7.3 \cdot 10^8 - 2.7 \cdot 10^9$ for the GAAGAGGGGG hairpin. For the Theodorakopoulos parameter set, we found $3.2 \cdot 10^3 - 8.8 \cdot 10^3$ for the GAAGAGGGAG hairpin, and $6.5 \cdot 10^3 - 2.2 \cdot 10^4$ for the GAAGAGGGGG hairpin.

4.2.5 Hairpin with 5 bps stem of 60 % GC content with different bases in stem.

In reference [30], Bonnet et al. investigated two hairpins, both with a five bps stem, GGGAA, one having a 21 bps loop of pure T bases, the other having a 21 bps loop of pure A bases. Bonnet et al. performed experiments using a combination of fluorescence energy transfer and fluorescence correlation spectroscopy. The experiments were performed at temperatures ranging from 285 to 310 Kelvin, and both the opening rate constant and the closing rate constant were obtained. The authours report on a slightly increase in the opening rate when the loop is altered from T21 to A21. However, the authors of reference [30] emphasize that the rate constant for the closing rate is much more affected by alteration of the loop than is the opening rate constant. They found the opening rate constant to be 1.4-1.7 times larger for the A21 loop relative to the T21 loop, and the closing rate constant to be 2-12 times larger for the T21 loop relative to the A21 loop. As earlier mentioned, we can not say anything specific about the loop, and thus, we obtained PBD model results for the opening rate constant for a hairpin with a stem of GGGAA, which we compared with the experimental results for the opening rate constant for loop T21 and loop A21. We present the PBD model results for maximal and minimal cutoff, respectively, and how they vary with the temperature.

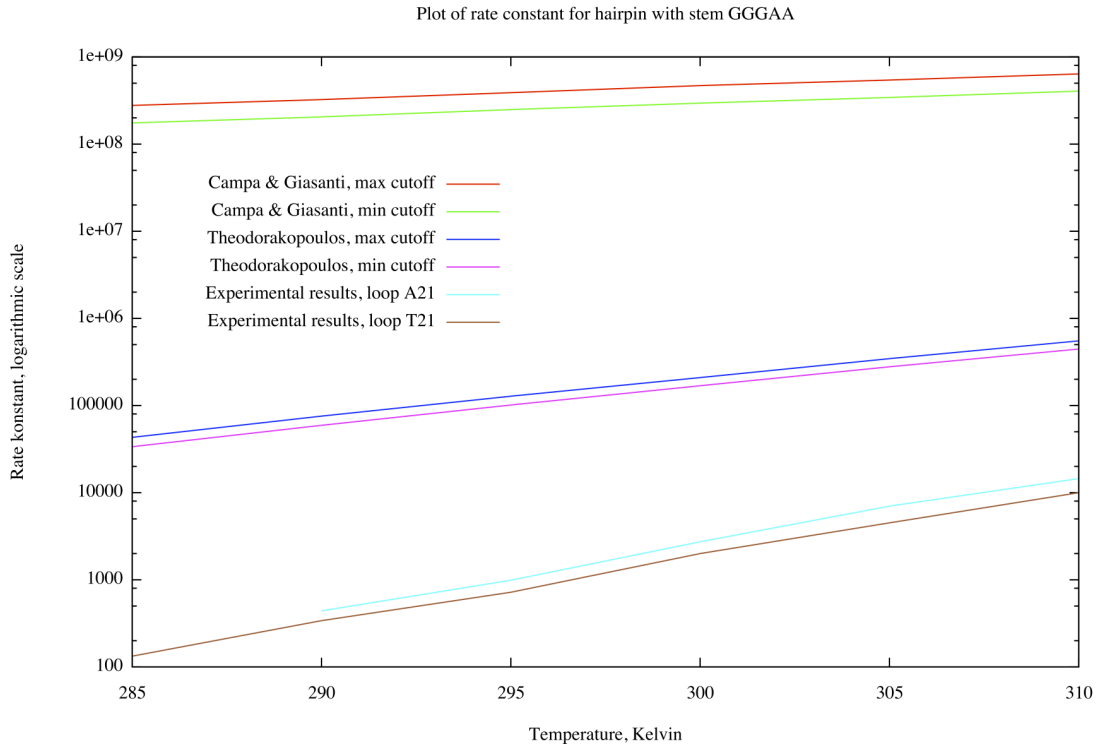
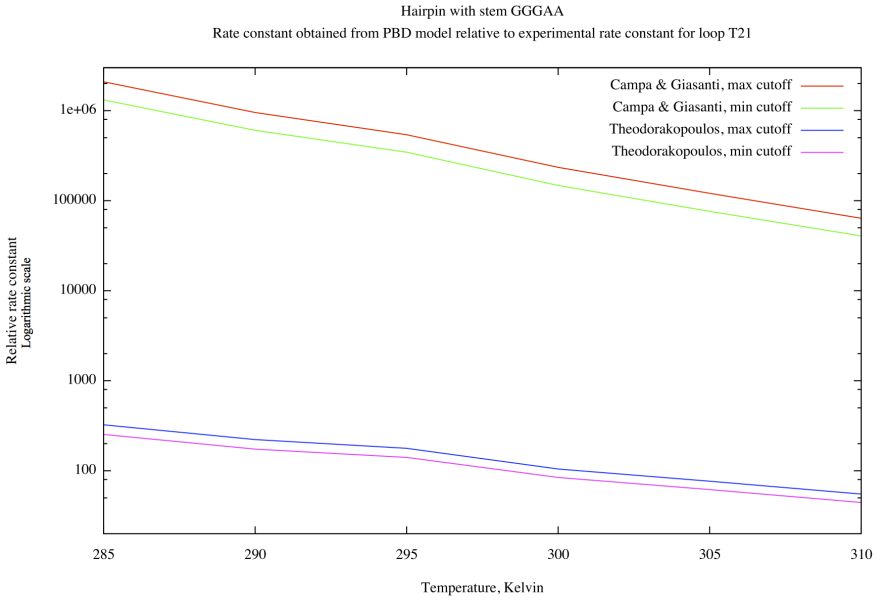
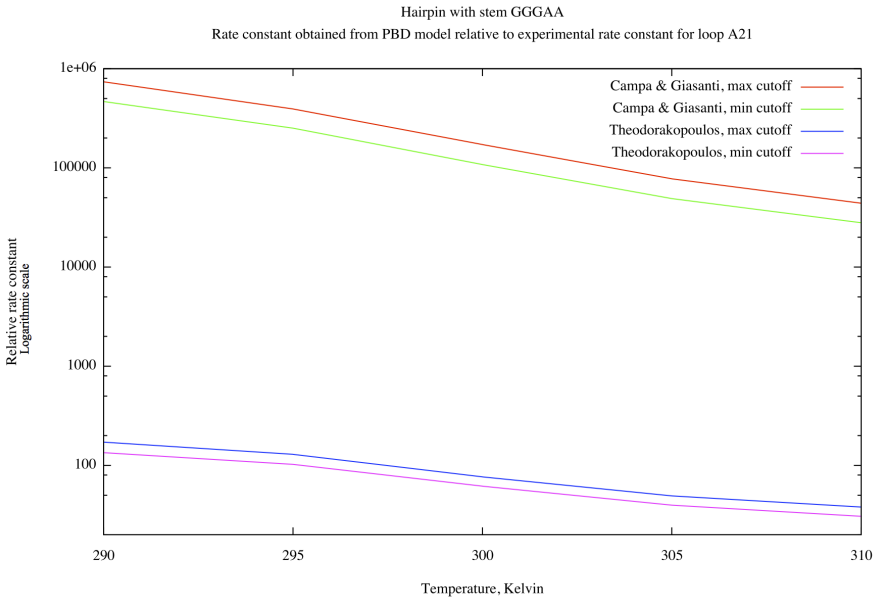


Figure 34: Plot of rate constant in units $1/s$ from experimental results for loop A21 and T21, PBD model results for Campa and Giansanti parameter set and Theodorakopoulos parameter set

Looking at figure 34, it is clear that for both the PBD model results and the experimental results, the rate constant increases with increasing temperature, which is expected. However, for the experimental results, the rate constant increases with a factor of 2.1-2.8 for each temperature interval of 5 Kelvin. For the PBD model results, the rate constant increases for each temperature interval of 5 Kelvin with a factor of 1.2 for the Campa and Giansanti parameter set and a factor of 1.6-1.7 for the Theodorakopoulos parameter set. The fact that the rate constant for the experimental results increases faster with temperature than the PBD model results, explains why the PBD results are relatively much larger than the experimental results for low temperatures compared to for high temperatures.
As for all the hairpins we have investigated until now, the Campa and Giansanti parameter set gives a result some magnitudes of order larger than the Theodorakopoulos parameter set, and the difference is relatively constant. For the GGGAA hairpin, the Campa and Giansanti parameter set gives results $9 \cdot 10^2 - 6.4 \cdot 10^3$ larger than the Theodorakopoulos parameter set.



(a) Loop T21



(b) Loop A21

Figure 35: Plot of relative rate constant, rate constant obtained from PBD model results for Campa and Giansanti parameter set and Theodorakopoulos parameter set relative to rate constant obtained from experimental results, as a function of temperature.

Loop	Cutoff	Relative rate constant, Campa & Giansanti	Relative rate constant, Theodorakopoulos
T21	Cutoff min	$4 \cdot 10^4 - 1.3 \cdot 10^6$	$44 - 254$
T21	Cutoff max	$6.4 \cdot 10^4 - 2 \cdot 10^6$	$55 - 325$
A21	Cutoff min	$2.8 \cdot 10^4 - 4.7 \cdot 10^5$	$31 - 135$
A21	Cutoff max	$4.4 \cdot 10^4 - 7.4 \cdot 10^5$	$38 - 172$

Table 4: Table showing how much larger the rate constants obtained from PBD model results are than the rate constant obtained from experimental results, for both T21 and A21 loop. These numbers are a summary of figure 35a and figure 35b.

From figure 35a and 35b, and summarized in table 4, we see that for the Campa and Giansanti parameter set, the rate constant for the PBD model divided by the experimental rate constant, i.e. the relative rate constant, is $4 \cdot 10^4 - 2 \cdot 10^6$ for the GGGAA hairpin with the T21 loop, and $2.8 \cdot 10^4 - 7.4 \cdot 10^5$ for the hairpin having the A21 loop. For the Theodorakopoulos parameter set, we found the relative rate constant to be $44 - 325$ for the T21 loop, and $31 - 172$ for the A21 loop.

We have found for all the hairpins investigated, that the PBD model results are orders of magnitude larger than the experimental results. Applying the Campa and Giansanti parameter set, we found relative rate constant values ranging from $2.8 \cdot 10^4 - 1.5 \cdot 10^{30}$. For Theodorakopoulos parameters, we found relative rate constant values ranging from $3.1 \cdot 10^1 - 8.6 \cdot 10^{19}$.
In this work, the friction coefficient was set to be $\gamma = 50\frac{1}{ps}$, a typical value for the friction of water[22]. It is not straightforward to select a friction coefficient for our system. From Stokes law, $\gamma = 6\pi\eta a$, where a is the radius of the particles and η the viscosity of the fluid, we can see that the friction coefficient depends on the size of the particles in the system. For our model, we operate with a mesoscopic system, which means that the radius of our particles is larger than the water molecules. Thus, it can be argued that the friction coefficient we have applied might indeed be too small for our system. As mentioned in section 3, at high friction, the relationship between the friction coefficient, γ, and the transmission coefficient, κ can be described by Kramers behaviour $\kappa = \frac{1}{\gamma}$. We know that the rate constant increases linearly with the transmission coefficient according to $k = k^{TST} \cdot \kappa$, and thus, by increasing γ by a certain factor, the rate constant will decrease by the same factor. If we obtained PBD rate constants that are a factor 10^5 too high compared to experiments, we could simply scale the friction coefficient by a factor of 10^5 larger to match experimental data. This would result in $\gamma = 500\frac{1}{fs}$. It seems kind of advantageous to change the friction coefficient in order to map the theoretical rate constants to the experimental ones. Changing the friction parameter will affect the rates, but not the statistics of the model. This implies that theoretical denaturation curves will look the same for all friction coefficients, which is important since PBD parameters have been fitted on these data. As mentioned, from Stokes law we can suggest that the friction coefficient should be larger for mesoscopic systems than for atomistic systems, but we are not sure if a friction coefficient of $\gamma = 500\frac{1}{fs}$ or even larger can be viewed as physical or not, as, to our knowledge, this is much larger than the friction coefficient applied in molecular simulations.
We must notice that the minimum and maximum relative rates for both parameters set is significantly different, with a factor of 10^{26} for the Campa and Giansanti parameter set and 10^{18} for the Theodorakopoulos parameter set. However, we know two important facts, that the relative rates increases with increasing stem length and increasing GC content of stem. Our results suggests that in order for the PBD model to describe dynamics correctly, fundamental changes have to be made to the model. Just increasing the friction parameter is not sufficient. Further and more extensive comparison of PBD model results to experimental data is needed in order to create these improvements to the model.

5 Sources of error

One of the main sources of error in this work lies in the limitations of PBD-model, as already described in section 2.3.
The integration of the free energy has no statistical errors, but choosing a too large integration step dy represents a possible systematic error. In calculations for chains lengths above 50 bps, we used $dy = 0.5$, due to the high computational cost for smaller integration steps. To investigate how large an error having an integration step of $dy = 0.5$ represented, we performed control simulations, where we simulated the same sequence with both integration step $dy = 0.5$ and $dy = 0.1$. We did this for all chain lengths, at high and low temperatures. What can be concluded, is that the rate constants obtained with the larger simulation step lies between 90-110 % of the rate constant obtained with the smaller integration step. However, for sequences of the same length and at the same temperature, they all increased or decreased the same amount when changing the integration step. Thus, the relative rates between the sequences was the same independent of the integration step.
The calculation of the transmission coefficient has a certain statistical error, which scales as $1\sqrt{N}$, where N is the number of trajectories. We used $N = 10^6$, and obtained an statistical error between $5 - 10\%$, which is very good results for such low reaction rates. In addition, the effect of the transmission coefficient on the rate constant was relatively small and for chain lengths > 20, the transmission coefficient it can be viewed as a constant of 0.02. Experimentally, reaction rates are obtained within an accuracy of one order or a few orders magnitude. This makes comparison with experiments tricky since their values might in fact be quite far away from the true physical denaturation rates.

6 Conclusion

The PBD model have been extensively fitted and tested to experimental equilibrium data, such as denaturation curves. In this work, it is the first time that the dynamcial properties of the PBD model have been tested and compared to experiments.
We explored the nature of ds DNA with the PBD model as a function of chain length, sequence order and temperature. Contrary to earlier studies conducted by [22], that mainly investigated at pure AT and GC chains, as well as 50 % AT 50 % GC chains, we looked at chains consisting of 33 % AT and 66 % GC bps. In the earlier studies, it was found that for 50 % AT 50 % GC chains, the rate constant of a sequence where the GC bps were evenly spread in the chain was much higher than other sequences which contained large blocks of GC bps. To our surprise, this rule is not always obeyed for the 33 % AT and 66 % GC chains that we studied. Especially for large sequences and at high temperatures, chains having large blocks of AT bps at the end of the chain, showed favourable opening rates. From our results, we also concluded that chains of 33 % AT and 66 % GC bps have a melting temperature close to 360 Kelvin. Next, we investigated a secondary structure of DNA and RNA, the hairpin structure. Here, we found PBD model results several orders of magnitude higher than experimental results. We argued that the friction coefficient used in this work might be too small, as it is considered reasonable for solute molecules in water described at the atomistic scale, while the PBD model operateat a mesoscopic scale. Increasing the friction coefficient with a certain factor will lead to a decrease in the rate constant by the same factor. We also found that PBD results relative to the experimental results increased with increasing stem length and GC content of stem. Thus, for the PBD model to have a correct description of the dynamics, some essential alterations of the model must be made. Simply increasing the friction coefficient is not adequate. Even though the PBD model results with the Theodorakopoulos parameters is still some orders magnitudes too high with the friction coefficient that we used, the general behaviour seems to fit better with the experimental results than the Campa and Giansanti parameters. In addition, as Theodorakopoulos parameters were intended for very long chains(between 1695 and 159662 bps)[9], PBD result from the Theodorakopoulos parameter set might fit better with experimental results for such long chains.

If the suggesions for improvements of the PBD model are conducted this could lead to a much more reliable mesoscopic model for ds DNA and DNA hairpins. The studies carried out in this work significates the importance of investigating dynamical quantities like the rate constant instead of only using equilibrium data for fitting model parameters. We believe that this work can initatiate a new approach for developing more accurate mesoscopic models for DNA.

.

References

[1] Pre-melting dynamics of DNA and its relation to specific functions. Boian Alexandrov, Nikolaos K Voulgarakis, Kim Ø Rasmussen, Anny Usheva and Alan R Bishop. *J. Phys.: Condens. Matter 21 (2009)*

[2] Dynamics and thermodynamics of a nonlinear model for DNA denaturation, T. Dauxois, M. Peyrard, and A. R. Bishop. *Phys. Rev. E 47, 684–695 (1993)*

[3] International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome". *Nature 431 (7011): 931-945*

[4] Komiya K., Yamamura M. and Rose J. A., *Nucleic Acids Res., 38 (2010) 4539.*

[5] Liu,D. and Balasubramanian,S. (2003) A proton-fuelled DNA nanomachine. *Angew. Chem. Int. Ed., 42, 5734–5736.*

[6] Chen,Y., Lee,S.-H. and Mao,C. (2004) A DNA nanomachine base on a duplex-triplex transition. *Angew. Chem. Int. Ed., 43, 5335–5338.*

[7] Takinoue M. and Suyama A., *Chem-Bio Inform. J., 4 (2004) 93; Small, 2 (2006) 1244.*

[8] Campa A, Giansanti A. Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains. *Phys. Rev. E. 1998;58:3585*

[9] N. Theodorakopoulos, *Phys. Rev. E 82, 021905 (2010).*

[10] R.B. Inman, R.L. Baldwin, *J. Mol. Biol. 8, 452 (1964).*

[11] Bubbles and denaturation in DNA. van Erp TS, Cuesta-López S, Peyrard M., *Eur Phys J E Soft Matter. 2006 Aug;20(4):421-34. Epub 2006 Sep 7.*

[12] Statistical Mechanics of a Nonlinear Model for DNA Denaturation, M. Peyrard and A. R. Bishop, *Phys. Rev. Lett. 62, 2755 (1989).*

[13] Stacking Interaction in DNA Molecule, S. Zdravkovic and M. V. Satarié. *Journal of Computational and Theoretical Nanoscience Vol. 7, 1-5, 2010.*

[14] Entropy-driven DNA denaturation, T. Dauxois, M. Peyrard, and A. R. Bishop, *Phys. Rev. E 47, R44 (1993).*

[15] Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains, A. Campa and A. Giansanti, *Phys. Rev. E 58, 3585 (1998)*

[16] Melting of DNA Oligomers: Dynamical Models and Comparison with Experimental Results, A. Campa, A. Giansanti, *J. Biol. Phys. 24, 141 (1999).*

[17] Helicoidal model for DNA opening, Maria Barbi, Simona Cocco, Michel Peyrard. *Physics Letters A 253 (1999) 358-369*

[18] Comment on 'Can One Predict DNA Transcription Start Sites by Studying Bubbles?' van Erp et al. reply. Titus S. van Erp, Santiago Cuesta-Lopez, Johannes-Geert Hagmann, and Michel Peyrard. *Phys. Rev. Lett. 97, 059802 (2006)*

[19] Keck, Dissociation and recombination reactions, *1962.*

[20] Multidimensional Tunneling, Recrossing, and the Transmission Coefficient for Enzymatic Reactions Jingzhi Pu, Jiali Gao, and Donald G. Truhlar, *Chem. Rev., 2006, 106 (8), 3140-3169*

[21] Efficiency analysis of reaction rate calculation methods using analytical models I: The two-dimensional sharp barrier, Titus S. van Erp. *J Chem Phys. 2006 Nov 7;125(17):174106.*

[22] *The dynamics of the DNA denaturation transition*, Titus S. van Erp and Michel Peyrard. EPL, 98 (2012) 48004

[23] Proctor, D. J., Ma, H., Kierzek, E., Kierzek, R., Gruebele, M. & Bevilacqua, *P. C. (2004), Biochemistry 43, 14004-14014.*

[24] Menger, M., Eckstein, F.& Pørschke, D. *(2000) Biochemistry 39, 4500-4507.*

[25] Jung, J. & van Orden, A. (2005) *J. Phys. Chem B 109, 3648-3657.*

[26] Froelich-Ammon, S. J., Gale, K. C. & Osheroff, N. *(1994) J. Biol. Chem. 269, 7719 –7725.*

[27] Zazopoulos, E., Lalli, E., Stocco, D. M. & Sassone-Corsi, *P. (1997) Nature (London) 390, 311–315.*

[28] Modeling DNA beacons at the mesoscopic scale, Errami, J. Peyrard, M. Theodorakopoulos, N., *European Physical Journal E – Soft Matter;Aug2007, Vol. 23 Issue 4, p397.*

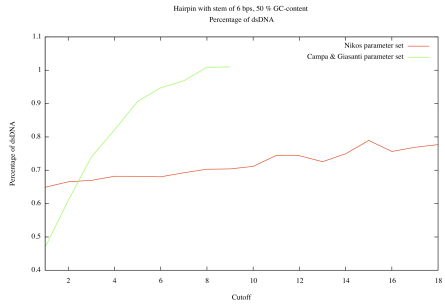
[29] Opening rates of DNA hairpins: Experiment and model, Jeungphill Hanne and Giovanni Zocchi, Nikolaos K. Voulgarakis, Alan R. Bishop, and Kim Rasmussen. *Phys. Rev. E 76, 011909 (2007)*

[30] Kinetics of conformational fluctuations in DNA hairpin-loops. Bonnet, Krichevsky and Linchaber. *Proc. Natl. Acad. Sci. USA, Vol. 95, pp. 8602–8606, July 1998.*

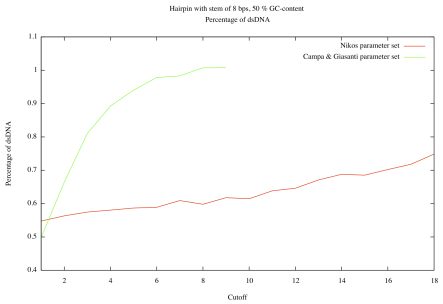
[31] Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins." M.T. Woodside, W.M. Behnke-Parks, K. Larizadeh, K. Travers, D. Herschlag, S.M. Block, *Proc. Natl. Acad. Sci. USA 103, 6190-6195 (2006)*

7 Appendix A

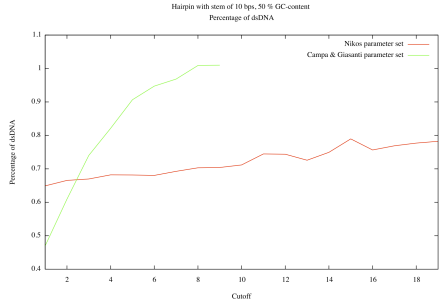
Percentage of dsDNA, all measurements.



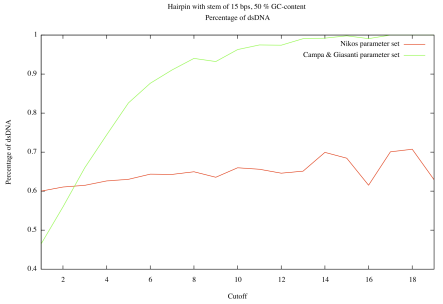
(a) 6 bps stem, 50 % GC-content



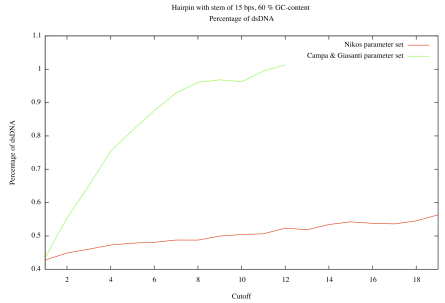
(b) 8 bps stem, 50 % GC-content



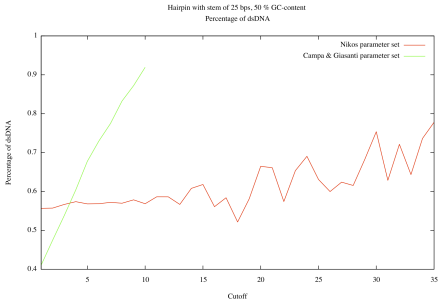
(c) 10 bps stem, 50 % GC-content



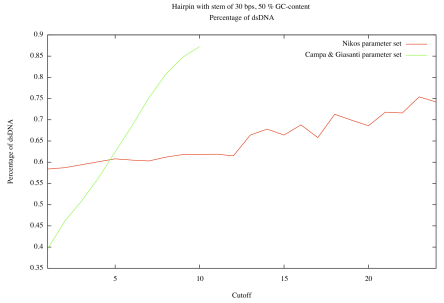
(d) 15 bps stem, 50 % GC-content



(e) 15 bps stem, 60 % GC-content

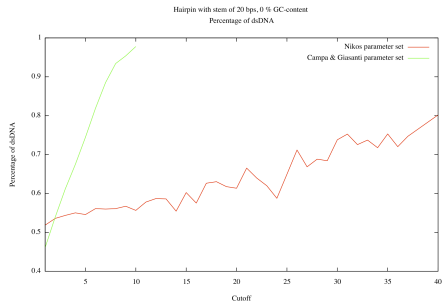


(f) 25 bps stem, 50 % GC-content

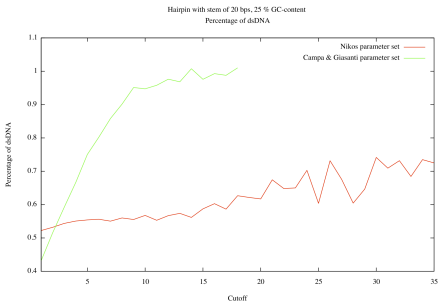


(g) 30 bps stem, 50 % GC-content

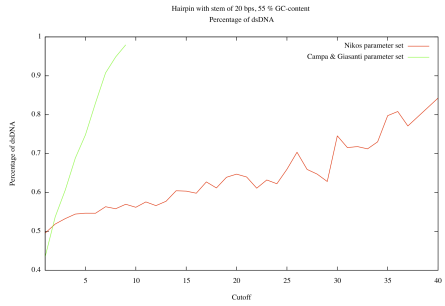
Figure 36: Percentage of dsDNA



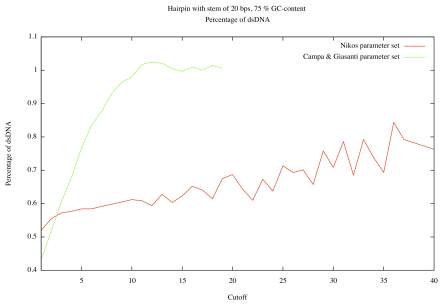
(a) 20 bps stem, 0 % GC-content



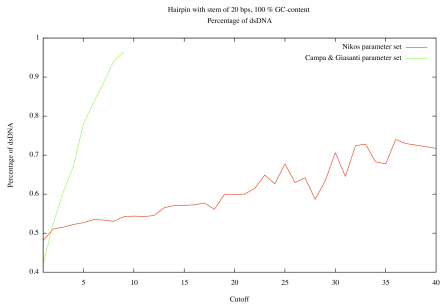
(b) 20 bps stem, 25 % GC-content



(c) 20 bps stem, 55 % GC-content



(d) 20 bps stem, 75 % GC-content



(e) 20 bps stem, 100 % GC-content

Figure 37: Percentage of dsDNA

8 Appendix B

Details of hairpins investigated in section 4.2.

Sequence	Loop	Experimental rate constant, $\ln k$, $\frac{1}{s}$
GAGAGGAGGAAGGAG	T3	-15 ± 2
GAGAGGAGGAAGGAG	T4	-19 ± 2
GAGAGGAGGAAGGAG	T6	-17 ± 3
GAGAGGAGGAAGGAG	T8	-13 ± 2
GAGAGGAGGAAGGAG	T10	-15 ± 3
GAGAGGAGGAAGGAG	T15	-14 ± 3
GAGAGGAGGAAGGAG	T20	-9 ± 6

Table 5: Table summarizing experimental values for hairpin with 15 bps stem of 60 % GC content, investigated in section 4.2.1

% GC content of stem	Sequence	Experimental rate constant, $\ln k$, $\frac{1}{s}$
0 %	AAAAAAAAAAAAAAAAAAAA	-10 ± 1
25 %	AAGAAAAGAAGAAGAAAGAA	-23 ± 3
50 %	GAGAGAAGGAGAGGAAGGAA	-29 ± 2
55 %	GAGAGAAGGAGAGGAAGGAG	-31 ± 2
75 %	GGAGGAGGGAGGGGAGGGAG	-36 ± 2
100 %	GGGGGGGGGGGGGGGGGGGG	-43 ± 3

Table 6: Table summarizing sequence and experimental results of hairpins with stem of 20 bps and loop of T4, investigated in section 4.2.2

Length of stem	Sequence	Experimental rate constant, $\ln k$, $\frac{1}{s}$
6	GAGGAA	2.4 ± 0.5
8	GAGAGGAA	0.4 ± 0.3
10	GAGAGAGGAA	-4.1 ± 0.6
15	GAGAGGAGGAAGGAA	-12 ± 2
20	GAGAGAAGGAGAGGAAGGAA	-29 ± 2
25	GAGAGAAGGAAGAGGGAGGAAGGAA	-37 ± 3
30	GAGAGAAGGAAGAGAAGAGGGAGGAAGGAA	-53 ± 5

Table 7: Table summarizing the length, sequence and experimental results for all hairpins with a 50 % GC content of stem and loop of T4 investigated in section 4.2.3