GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Bachelor's Thesis

Dye'ing for FRET: The HTH motif in smFRET-MD Simulation

Das HTH motif in smFRET-MD Simulation

prepared at the Max Planck Institute for Biophysical Chemistry, Göttingen
The Grubmüller Group
by Timo Graen from Hannover

Thesis period:	7th April 2009 until 20th July 2009
Supervisor:	Prof. Dr. Helmut Grubmüller
First referee:	Prof. Dr. Helmut Grubmüller
Second referee:	Prof. Dr. Jörk Enderlein

Abstract

The simulation of protein folding kinetics is a challenging task and modern computers are just reaching the performance levels required to push multiple folding trajectories into the microsecond range. Even with enough computing power at hand, the verification of the simulated protein kinetics usually relies on static parameters like average folding rates from folding experiments[34]. In recent years, single molecule FRET (smFRET) experiments have produced interesting data for the validation of folding kinetics simulations. This interest arises from the fact that single recorded photons, the raw experimental data, are directly connected to the kinetic processes during folding. Therefore, the prediction of experimental photon distributions can provide direct information on the physical correctness of the simulated folding trajectories. On this basis further understanding of the folding kinetics is desired.

The focus of this thesis is the setup of a smFRET MD simulation for the 44 amino acid Helix-Turn-Helix motif of the engrailed homeodomain protein [28, 29]. This work includes the parametrisation of the Alexa 488 and Alexa 647 dyes into the Generalized Amber Force Field. A suitable smFRET-MD simulation program was developed to simulate FRET kinetics using data from MD trajectories. Additionally, several methods for FRET efficiency calculations were compared using the smFRET simulation program.

Abstract

Die Simulation von Proteinfaltungen ist eine komplexe Aufgabe, welche auch morderne Computer schnell an ihre Grenzen bringt. Seit einigen Jahren ist es möglich mehrere durchgängige Trajektorien von wenigen Mikrosekunden zu berechnen. Interessant ist besonders die Untersuchung von dynamischen Vorgängen während der Faltung. Diese werden meistens mittels statischer Parameter wie zum Beispiel der mittleren Faltungsrate überprüft [34]. In den letzten Jahren wurden jedoch verstärkt Einzelmolekühl FRET Messungen (smFRET) durchgeführt, welche interessante Messdaten für den Vergleich mit Simulationen liefern. Dieses Interesse lässt sich haupsächlich durch die Verbindung der experimentellen Rohdaten mit den dynamischen Prozessen in dem untersuchten System erkären. Auf dieser Grundlage können Simulation und Experiment miteinander verglichen werden. So kann die physikalische Korrektheit der Simulation geprüft werden.

Die vorliegende Bachelor Arbeit befasst sich mit der Vorbereitung einer smFRET Simulation des Helix-Turn-Helix Motifs [28, 29]. Hierfür wurden die beiden FRET Farbstoffe Alexa 488 und Alexa 647 parametrisiert und für die Simulation mit dem HTH Motif vorbereitet. Zusätzlich wurde ein Programm entwickelt, mit welchen die FRET Kinetik des Systems simuliert werden kann. Anschließend wurde das Programm benutzt um einige bestehende FRET Analyse Methoden zu vergleichen und die Qualität des Programms zu überprüfen.

Keywords: HTH-motif, smFRET simulation, Alexa 647, Alexa 488, Molecular Dynamics

Contents

Nomenclature

Abbreviations

Abbreviation	
MD	Molecular Dynamics
smFRET	single molecule Fluorescence Resonance Energy Transfer
HTH-motif	Helix-Turn-Helix motif, see [37, 39]
smFRET-MD	Simulation program to simulate smFRET experiments ontop of MD trajectories
GAFF	Generalized Amber Force Field
ESP	Electrostatic Potential

1. Introduction

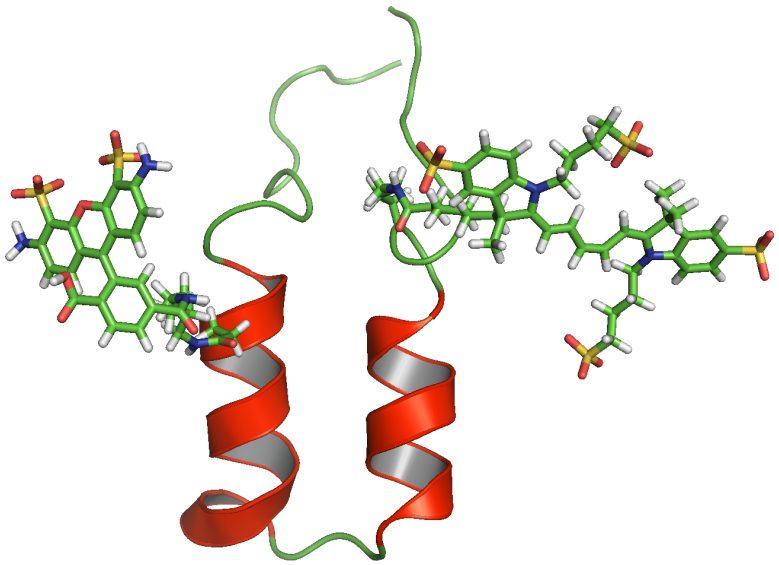


Figure 1.1.: The HTH-motif with Alexa 488 (left) and Alexa 647 (right) attached

It is the age of man-made machines. Similar to life, machines are being developed in an evolutionary process. New generations outperform their predecessors and explore new fields that were previously inaccessible. However, mankind is still very new to this game. In nature, highly sophisticated machines have evolved over billions of years. Nature provides us with the assembled machines but not with instruction manuals or assembly plans. These have to be reconstructed by science.

Proteins are part of nature's molecular machines. They consist of a polymer chain of about 20 different amino acids connected by peptide bonds. This chain has a specific three dimensional structure with the ability to self assemble. In recent years, large-scale sequencing projects have produced large amounts of protein sequences to study. It is desired to understand what purpose these proteins serve and what functionality they provide. The idea is to use the evolutionary advantage of nature to find protein mechanisms which solve existing challenges in many fields of science.

The functionality of Proteins highly depends on their three dimensional structure and can currently not be predicted from the amino acids sequence alone. This is due to the large number of possible arrangements for a given sequence. A small protein, like the 44 amino acid HTH-motif [37], would result in 8^{44} different conformations, assuming an average of eight conformations for each amino acid [17]. However, the HTH-motif folds in microseconds [39]. This time frame is not sufficient to sample all possible conformations. This is known as Levinthal's paradox.

Additionally, the folded structure is thermodynamically the most stable under a given set of specific conditions [12]. In other words, the folded structure can be obtained from the global free energy minimum. This is often referred to as Anfinsen's dogma. The global minimum is difficult to obtain from the sequence since the free energy landscape is usually unknown. In experiments, this free energy minimum corresponds to the native state of the protein.

Current structure prediction algorithms often rely on known similar structures, homology models, to predict the free energy minimum of unknown structures. Without homology to known structures the prediction of native structures remains difficult[38]. A more recent hypothesis explains the fast folding of proteins based on an optimization scheme. In this scheme, proteins solve the optimization problem by separating it into smaller pieces which each assemble separately. This drastically reduces the degrees of freedom resulting in faster folding rates [9]. Even though huge process has been made in the field of structure prediction, one drawback of these methods is their focus on the outcome and not the folding process. Therefore, only little information can be obtained about the protein folding kinetics.

The kinetics of protein folding can be studied at high resolution using MD simulations [11, 33, 34, 40]. The simulation trajectory, or many trajectories, are used to calculate a projected free energy landscape which is used to search the global free energy minimum. First attempts lagged proper sampling due to the stochastic nature of protein folding kinetics[35]. Given sufficient sampling, the MD method provides folded structures, folding transition states and the folding kinetics. The required computing power to sample microsecond folding kinetics on single trajectories has been available for over ten years now [11]. Since then, several studies have tested the predictions quality of MD simulations against experimental data. On the stochastic level, very large simulation setups with up to 0.5 ms sampling time predicted protein folding rates in good agreement with experimental values [30, 35]. Other approaches compared the minimum of the projected free energy landscape

from MD simulations to the native state in NMR measurements [40]. Replica exchange methods were used to enhance the sampling for shorter simulation times [33]. However, even with agreements in folding rates and native state prediction, the simulation trajectory still requires further validation of the underlying kinetic processes[35].

Molecular Dynamics approaches largely benefit from the emerging field of single molecule experiments. Predictions from MD simulations can be tested against data from single molecule experiments instead of many molecule averages. Within the last years, single molecule FRET experiments have developed from being pure proof of concept studies towards being a useful tool for answering questions on the single molecule level [4, 25, 32, 36]. The classical smFRET setup measures photon distributions of fluorescence dyes pairs. The photon distributions is used to calculate FRET efficiencies or FRET energy transfer rates [32]. Afterwards, Förster Theory is used to calculate molecular distances from the FRET efficiencies. Besides this attempt to measure molecular distances the photon distribution itself is valuable information as it is directly connected to the kinetic processes of the system.

Here, we present an attempt to compare simulated folding kinetics to experimental data, the smFRET-MD simulation. In this attempt, the kinetics of FRET energy transfer and photon emission in single molecule FRET experiments are modelled. The result is a simulated photon distribution which can be compared to the recorded photon distribution from smFRET experiments. As the photon emission of FRET systems highly depends on the system kinetics, we test the prediction quality of kinetic processes in MD protein folding simulations.

In this attempt, we used the previously mentioned Helix-Turn-Helix motif of the EnHD protein [37, 39] with Alexa 488 and Alexa 647 as the FRET donor and acceptor dyes. This setup is illustrated in figures 1.1 and 2.1. The HTH motif was chosen for its small size and fast folding rates[39]. This provides good accessibility of the folding kinetics using MD methods. This bachelor's thesis concentrates on the preparation of the MD simulation and the implementation of the smFRET-MD simulation program.

In preparation of the MD simulation, the two FRET dyes, Alexa 488 and Alexa 647, were parameterized using the Generalized Amber Force Field[21]. The HTH motif and the Alexa dyes were simulated in free MD simulation for 120 ns. The obtained trajectory was used to simulate the FRET kinetics and to calculate the time dependend photon distribution. Afterwards, the smFRET-MD program was used

to test the smFRET simulation prediction quality against the distance distribution from the MD trajectory. We also analyzed photon distributions according to the "Spectroscopic Ruler" experiments by Schuler et al. [32] and discussed approximations that entered the underlying data analysis.

2. Molecular Dynamics simulations

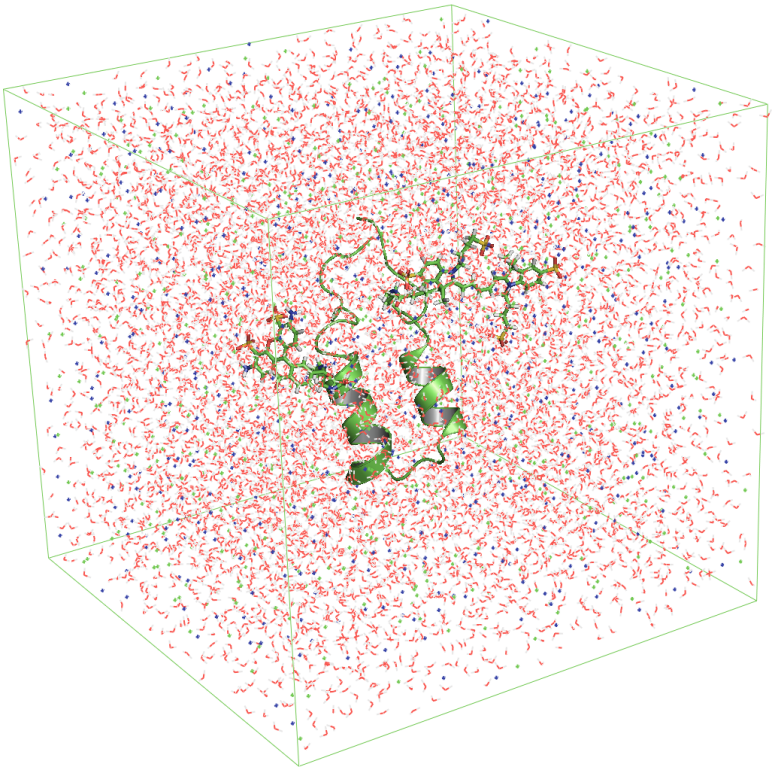


Figure 2.1.: The HTH-motif simulation box with dyes attached in 4M NaCl solution (HTH-motif shown in simplified cartoon representation of the all atom simulation)

Proteins are the workhorses of life. Proteins have different shapes and sizes forming a large ensemble of biological tools. The exploration of this biological set of tools is a very active field of research. Single molecule experiments are the focus of this thesis and will be monitored at atomic detail. We are modeling timescales that capture fast conformational changes of single amino acid residues in the protein. This level of detail can only be reached "in silico" and is not yet accessible in experimental setups. The underlying calculations are still very demanding and are just reaching the micro second range[22]. We have chosen the method of Molecular

Dynamics (MD) simulation for our simulations as it provides both computational efficiency as well as high resolution information on the dynamics. As always, this does not come free and therefore this chapter is dedicated to the approximations that enter our simulations.

2.1. Modeling Molecules

Molecular systems are determined by a number of interactions which determine their conformations and their dynamics. On the atomic level, molecular bonds between atoms restrain inter atomic distances. Pauli repulsions keep atoms from overlapping, and Coulomb interactions act on charged or partially charged atoms, forcing the whole system to move in the direction of the lowest free energy. For biological systems, the forming and breaking of hydrogen bonds is crucial for stabilizing secondary structures such as α-helices and β-sheets. A reasonable molecular model must therefore account for all these interactions.

The underlying equation for molecular dynamics is the time dependend Schrödinger Equation,

$$\imath\hbar\partial_t\Psi(\mathbf{R},\mathbf{r},t) \;=\; \hat{\mathbf{H}}\Psi(\mathbf{R},\mathbf{r},t). \tag{2.1}$$

In this equation the wavefunction $\Psi(\mathbf{R},\mathbf{r})$ depends on both the positions of the N nuclei $\mathbf{R} = (\mathbf{R_1},...,\mathbf{R_N})$ as well as the positions of the k electrons $\mathbf{r} = (\mathbf{r_1},...,\mathbf{r_k})$. The Hamiltonian is given by

$$\hat{\mathbf{H}} = \hat{\mathbf{T}}_{Nucl}(\mathbf{R}) + \hat{\mathbf{T}}_{Elec}(\mathbf{r}) + \hat{\mathbf{V}}_{Nucl/Nucl}(\mathbf{R}) + \hat{\mathbf{V}}_{Elec/Elec}(\mathbf{r}) + \hat{\mathbf{V}}_{Nucl/Elec}(\mathbf{R},\mathbf{r}). \tag{2.2}$$

Here, $\hat{\mathbf{T}}_x$ and $\hat{\mathbf{V}}_x$ denote the kinetic and the potential energy operators of the nuclei and electrons and their interactions. The subsequent goal is to obtain a solution for the time dependend Schrödinger Equation which is suitable for systems of several million atoms[16].

2.2. The Three Main Assumptions

The most important approximation in Molecular Dynamics is the Born-Oppenheimer approximation of the electron/nucleus interaction. We assume the electron motions to be orders of magnitude faster than the motions of the nuclei, due to their large

mass difference. As a consequence, the electrons follow the nucleic motions nearly instantaneously, thus decoupling the dependency of the wavefunction on the nuclei positions $\mathbf{R}$ and electron positions $\mathbf{r}$. This leads to the *ansatz*:

$$\Psi(\mathbf{R}, \mathbf{r}, t) \;=\; \Psi_{nuclei}(\mathbf{R}, t)\Psi_{electrons}(\mathbf{r}; \mathbf{R}). \tag{2.3}$$

With the Born-Oppenheimer approximation we can now solve the electronic motion by solving the time-independent Schroedinger-Equation. The nucleus position is now a parameter instead of a dynamic variable,

$$\hat{\mathbf{H}}_e\Psi(\mathbf{r}; \mathbf{R}) \;=\; E_e(\mathbf{R})\Psi(\mathbf{r}; \mathbf{R}). \tag{2.4}$$

Here, $\hat{\mathbf{H}}_e$ is the new electronic Hamiltonian,

$$\hat{\mathbf{H}}_e = \hat{\mathbf{T}}_{Electrons}(\mathbf{r}) + \hat{\mathbf{V}}_{Nucleus/Nucleus}(\mathbf{R}) + \hat{\mathbf{V}}_{Electron/Electron}(\mathbf{r}) + \hat{\mathbf{V}}_{Nucleus/Electron}(\mathbf{R}, \mathbf{r}). \tag{2.5}$$

Thus the eigenvalues of $\mathbf{H}_e$ depend on the static parameter $\mathbf{R}$, $E_e(\mathbf{R})$, and form an electronic potential which interacts with each nucleus upon motion. The dynamics of the nuclei can now be written as

$$\hat{\mathbf{H}}\Psi(\mathbf{R}, t) \;=\; \left[\hat{\mathbf{T}}_{nuclei} + E_e(\mathbf{R})\right]\Psi(\mathbf{R}, t) \;=\; \imath\hbar\partial_t\Psi(\mathbf{R}, t). \tag{2.6}$$

The second approximation describes the motion of the nuclei in classical approximation, i.e., using Newton's Equations of Motion.

$$F_i \;=\; m_i\partial_t^2\mathbf{R}_i \;=\; -\nabla E_e^0(\mathbf{R}) \tag{2.7}$$

Here the force on the i'th atom with mass m_i is calculated from the potential energy surface of the ground state E_e^0. The classical approximation has proven to be valid for modeling structural conformations of proteins and their dynamics. Some vibrational modes, i.e., of C-H, are approximated only roughly since their excitation energy exceeds the thermal energy $\approx kT$ of the system. Thus, the classical approach requires a low dependency of structural protein conformations on these fast vibrational modes. Empirically this assumption seems to hold as MD-simulations make experimentally valid predictions on many proteins [1, 10].

The third approximation of Molecular Dynamics is an approximation of the energy

function $E_e^0(\mathbf{R})$. The approximation consists of a computationally cheap ball and spring model coupled with terms for the Coulomb and Pauli interactions. We call this approximation a *force field*. The force field consists of a potential energy function and a set of constants to model the interaction strengths. Force fields are usually created from both *ab initio* quantum mechanics and thermodynamic experiments on model structures[45]. The backbone of all MD force fields is the potential energy function,

$$\begin{aligned} V(\mathbf{R}) &= V_{bonds}(\mathbf{R}) + V_{angles}(\mathbf{R}) + V_{dihedr.}(\mathbf{R}) + V_{LJ} + V_{coulomb}(\mathbf{R}) \\ &= \sum_{bonds} k_f(r - r_{eq})^2 + \sum_{angles} k_\theta(\theta - \theta_{eq})^2 + \\ &\quad \sum_{dihedr.} \frac{\nu_n}{2} \times [1 + cos(n\phi - \gamma)] + \\ &\quad \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right] + \sum_{i<j} \left[\frac{q_i q_j}{\epsilon R_{ij}} \right]. \end{aligned} \tag{2.8}$$

This approximation of the potential energy function uses a set of force constants k_f, k_θ and ν_n, multiplicities n, phase angles ϕ for torsional interactions and parameters A, B, q to model the Lennard-Jones and Coulomb potentials. Equation 2.8 only describes a simple "Class I" model. However, it can easily be modified to describe special conditions more accurate, i.e. bonds, by adding Morse potentials instead of harmonic potentials. Examples for common protein force fields are AMBER, CHARMM, GROMOS, and OPLS-AA[7, 20, 27, 42]. A practical approximation in these force fields is the definition of general atom types. The atom types are used to transform molecular structures into custom force field parameter sets.

In the following chapters we will make use of the AMBER99sb force field and use the AMBER based Generalized Amber Force Field (GAFF)[45] in extension. The GAFF defines 35 basic atom types with an additional set of 22 special atom types for common atomic conformations found in proteins and many organic compounds.

2.3. Simulation Theory

The level of approximation reached so far allows for solutions of the time dependend Schrödinger's Equation with reasonable accuracy. However, the practical imple-

mentation involves *quadratic scaling* system behavior like electrostatic interactions, which can only be handled poorly even for very small systems with $n \approx 10^4$ atoms. The focus of this section will therefore be the extension of the MD method to systems with several millions of atoms and timescales up to microseconds. This is only possible by reducing the scaling of the algorithms involved in the force calculations to $nlog(n)$ or less, yet retaining the physical correctness of the calculations. For sake of performance, all MD-simulations for this thesis were performed using the GRoningen MAchine for Chemical Simulations (GROMACS) in version 4.0[19].

2.3.1. The Integrator

The MD approach assumes classical movement of the nuclei. Therefore, we can obtain the forces F_i acting on each nucleus i from the potential energy $V(x_i)$ as follows:

$$F_i = -\nabla_{x_i} V(\mathbf{x}_i(t)) \tag{2.9}$$

MD-simulations often use a third order Leapfrog integrator[41] to solve Newton's equations of motion for the system. The algorithm is an equivalent formulation of the *Verlet* algorithm:

$$x(t_n + \Delta t) = 2x(t_n) - x(t_n - \Delta t) + \frac{F(t_n)(\Delta t)^2}{m} + O[(\Delta t)^4] \tag{2.10}$$

The Leapfrog algorithm was designed to explicitly contain the velocities in the integration step. Thus, the system can directly be coupled to a thermal bath. It's name originates from the leaping of velocities $v(t - \Delta t/2)$ and positions $x(t)$ in alternating time points.

$$\begin{aligned} v(t_n + \Delta t/2) &= v(t_n - \Delta t/2) + \frac{F(t_n)\Delta t}{m} + O[(\Delta t)^3] \\ x(t_n + \Delta t) &= x(t_n) + v(t_n + \Delta t/2) * \Delta t + O[(\Delta t)^3] \end{aligned} \tag{2.11}$$

The timestep Δt has to be reasonable small in order to minimize integration errors. A common integrationstep is 1 fs. The integration time step can be increased by neglecting high frequency / low amplitude bond vibrations, which are replaced by a set of distance constraints. The use of a LINear Constraints Solver (LINCS)[18, 26] allows an integration step sizes of 2 fs thereby increasing the effective length of the simulation.

2.3.2. Coupling of the Simulation Box

Systems for MD-simulations are usually, for practical reasons, placed inside a simulation box, see figure 2.1. The simulation box can then be equipped with periodic boundary conditions (pbc) where particles leaving one side of the simulation box reenter the box on the opposite end. Due to the small system size, temperature and especially the pressure are not represented correctly. In order to keep the temperature of the simulation box at a predefined value, the systems needs to be coupled to a thermal bath. This can be done using the Berendsen temperature[5] coupling algorithm. The velocities of the atoms in the simulation box are rescaled according to the temperature difference from the reference value T_0:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_0 - T(t)) \tag{2.12}$$

Here, τ defines the coupling strength of the system to the heat bath and $T(t)$ the current temperature. The pressure can be coupled in a very similar fashion by rescaling the coordinates, and thus the volume of the simulation box V.

$$\frac{dp(t)}{dt} = \frac{1}{\tau_p}(p_0 - p(t)) \tag{2.13}$$

Special care has to be taken whenever the system has to correctly represent a thermodynamic ensemble, which the Berendsen thermostat [5] does not. Other algorithms like the slightly more expensive velocity rescale algorithm can be used instead. However, the error of the Berendsen thermostat is only significant for calculations that rely on the accurate velocity distribution, which is not the case for this thesis. Therefore, the fast Berendsen thermostat and the Berendsen pressure coupling algorithms were used.

2.3.3. Electrostatics

The force field described in equation 2.8 has two terms contributing to the non-bounded interactions, the Lennard-Jones potential and the Coulomb potential. It can easily be seen how the Lennard-Jones potential decays like r^{-6}. This justifies a simple cut-off approach. For MD simulations, this is often chosen to be between 1.0 and 1.4 nanometers.

The second term, the coulomb potential, cannot be handled using a cut-off ap-

proach as it decays like r^{-1}. MD simulations therefore use Ewald sum methods to calculate coulomb electrostatics. A very powerful algorithm for calculating long range interactions is the Particle Mesh Ewald method [13] which splits the coulomb interactions into two separate sums. Short range coulomb interactions are calculated in direct space, while the long range interactions are shifted to Fourier space, both sums show quick convergence. The Fourier transformations are then calculated using three dimensional fast fourier transformations. This approach allows for longrange interactions to be calculated at the cost of only $nlog(n)$ compared to n^2, with n being the number of particles[13].

2.4. MD Simulation Setup

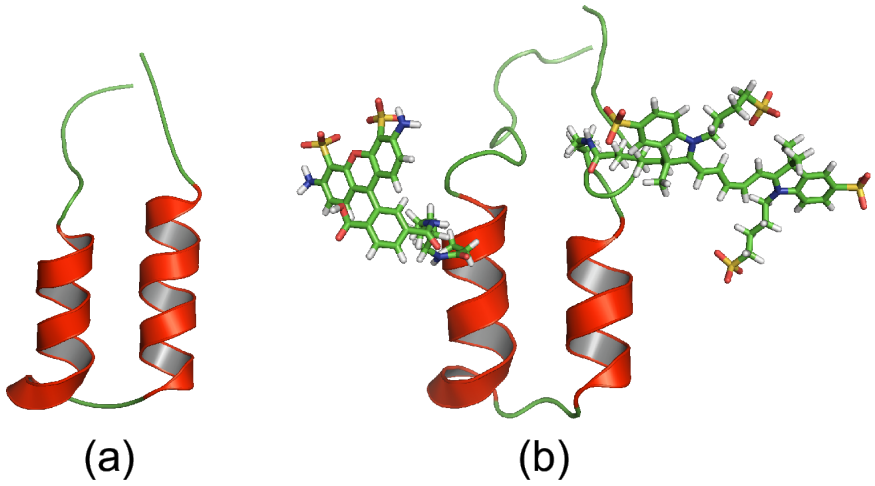


Figure 2.2.: (a) folded HTH motif after 100 ns of free MD simulation (b) Alexa 488/Alexa 647 mutant of the HTH motif

The L16A mutant of the 44 amino acid HTH motif [39] was used for the MD simulation as found in the protein data bank (entry 2p81). The partially folded structure was simulated in free MD simulation for 100 ns at 300 K until a folded conformation was observed, shown in figure 2.2 (a). The folded conformation was mutated at residues E28 and K52 to contain the FRET dye pair Alexa 488/Alexa 647, shown in figure 2.2 (b). The protein was parameterized using the Amber99sb force field and both of the Alexa dyes were parameterized using the generalized amber force field (GAFF) [21]. The system was placed in a 6.5x6.6x6.8 nm simulation box and solvated with tip3p water at 4 M NaCl concentration. The high ion concentration was chosen in analogy to the experimental setup. This is believed to

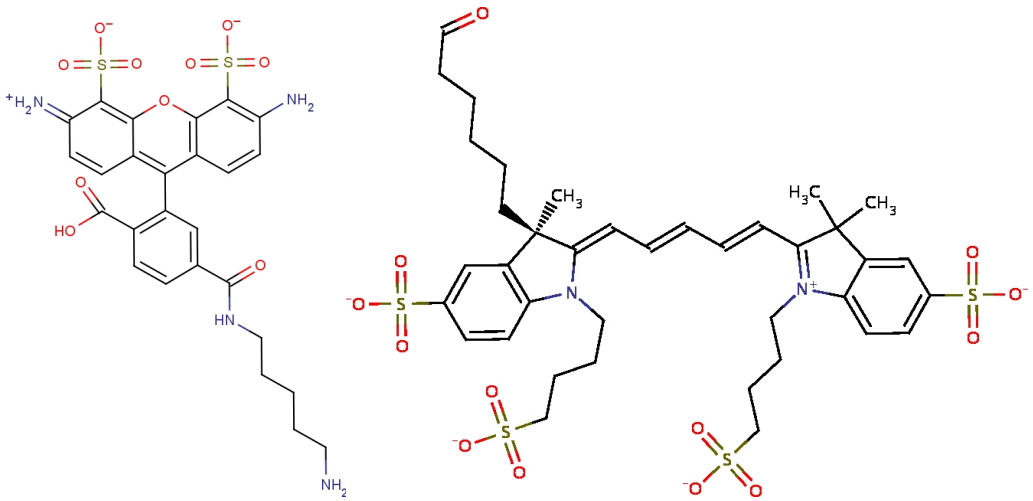
stabilize the folded conformation as suggested by Fersht et al. [39]. The system was energy minimized using the steepest descent algorithm and equilibrated for 5 ns in free MD simulation. The equilibration part of the trajectory was not used in the smFRET simulation.

All coulomb interactions were calculated using the Particle Mesh Ewald (PME) [13] method with a fourierspacing of 0.13 nm and a 1.4 nm cutoff was used for the van der Waals potential. Additionally, all bonds were restrained using the LINCS algorithm [18, 26]. Therefore, the integration step size could be increased to 2 fs. The system was simulated for 100 ns at 300 K using a 16 core beowulf cluster with four 2.66 Ghz Xeon processors.

2.5. MD Simulation Results

During the simulation, the protein was monitored by calculating the root mean square deviation (RMSD) of the protein backbone, residues T27-A54. The protein remained folded with only some movement in the flexible loop region as can be seen in figure A.1. The total energy of the system as well as temperature, pressure and the potential, Lennard Jones and Coulomb energies were monitored during the simulation. All parameters converged during equilibration and remained converged throughout the trajectory.

3. Parametrisation of the Dyes



(a) (b)

Figure 3.1.: a) Alexa 488 dye with linker b) Alexa 647 dye with linker

A particularly important task of this thesis was the parametrisation of the Alexa 488 and Alexa 647 dyes into the AMBER force field, see equation 2.8. As mentioned in section 2.2, MD-simulations rely on force field parameter sets. Numerous force fields exist for simulating protein dynamics, examples are GROMOS, OPLS-AA, AMBER and CHARMM[7, 20, 27, 42]. However, the two fluorescence dyes used for this thesis show little similarity to existing amino acid residues. Therefore, the mentioned force fields do not provide sufficient generalization for the required dye molecule parametrisation. This issue has previously been handled for the OPLS-AA force field by replacing unknown force field parameters with parameters from similar chemical configurations. Partial charges were calculated on OPLS-AA amino acid residues using quantum chemical methods. Subsequently, partial charges were compared to existing force field charges. The result was a scaling factor for new partial charges [23, 31]. In conclusion, dyes were parameterized using approximations from similar protein atom configurations and rescaling of partial charges from quantum

chemical calculations.

This thesis is based on a similar approach. The Generalized Amber Force Field (GAFF)[45] in extension to the AMBER99sb[27] force field was used for the dye parametrisation. The GAFF is compatible to the original AMBER force field, yet introduces parameters for many organic molecules[45]. This set of new atom types covers large parts of the Alexa 488 and Alexa 647 dyes. We consider generalized atom types to be the better approximation compared to using modified protein force field parameters. However, the usage of GAFF parameters for the dyes does not automatically guarantee to correctly reproduce the physical behavior of the system. The process of parametrisation and the measures taken to ensure their physical correctness will be described in this chapter.

3.1. Charge Calculations

We paid special attention to the partial charge calculations as they are an essential part of molecule parametrisation. Electrostatic interactions like the Coulomb potential, strongly rely on the correct partial charge distribution of parameterized molecules. Partial point charges were calculated for each atom using quantum chemical methods, see section 2.2. The process of partial charge calculation for the Generalized Amber Force Field is described in the following paragraph [8].

First, a structure of the dye molecule was created using the MARVIN [3] modeling software. The result was a file containing atomtypes and their coordinates. Then the molecular structure was optimized using the Hartree Fock method with a 6-31G* basis set followed by the generation of the ElectroStatic Potential (ESP) charges. Secondly, the optimized molecular coordinates and the charge grid (ESP) were used as input for a Restrained ElectroStatic Potential (RESP) fit. The RESP fit assigned partial point charges to the optimized atom coordinates from the charge grid (ESP). Symmetry groups were used to restrain the fit, for example all three oxygens in a SO_3 group should carry the same partial charge. Subsequently, the results from the RESP fits were checked to ensure the conservation of molecular symmetry.

In this thesis, all quantum chemical calculations were performed using the GAUSSIAN03 [14] software. The Antechamber [44] package was used for the RESP fit. Based on the optimized structure and RESP point charges, the teLEAP program of the AMBER simulation software [27] was used to assign the proper GAFF atomtypes to the molecular structure. The result was a complete set of parameters, the

"topology", for performing MD-simulations on the new molecule. Following standard procedure, all coulomb interactions were rescaled by a factor 0.8333 and all Lennard Jones interactions were rescaled by 0.5[45].

3.1.1. Alexa 488 Dye Parametrisation

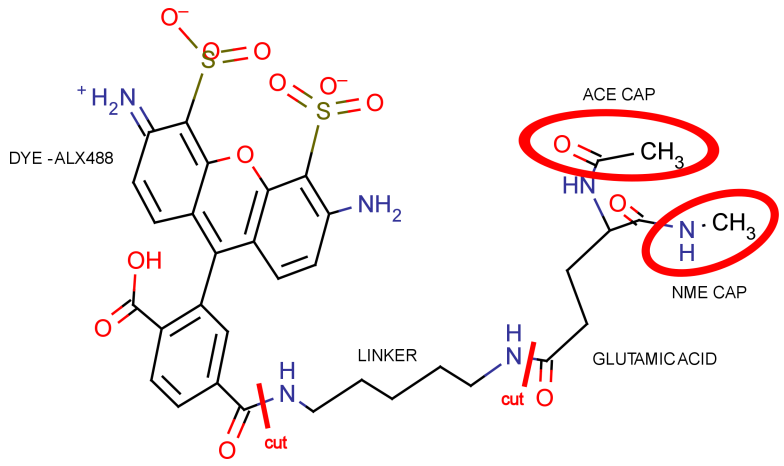


Figure 3.2.: Fragmentation of Alexa 488 prior to the partial charge calculations. (red cuts: ACE/NME caps were applied)

The Alexa 488 dye was attached to residue E28 of the HTH protein. Therefore, a modified glutamic acid residue (GLU488) had to be created. This required both a "coordinate" file as well as a matching force field parameter set, the "topology" file. The creation of the GLU488 structure was rather straight forward and done using the molecular modeling package MARVIN[3]. However, there exist many different conformations for the GLU / linker / Alexa 488 system, which are likely to show different partial charge distributions. For this thesis, we included only one conformation into the partial charge calculations.

The setup for the quantum chemical calculations reads as follows: The glutamic acid conformation was used as found in the HTH protein structure from the protein data base (id. 2P81). An extended conformation (planar) for the linker was chosen with maximal extension between dye and the glutamic acid. The Alexa 488 dye was modelled in its 6-isomere form. The van der Waals radii from the AMBER force field were used, and the atom type, angle, bond, pair and dihedral parameters were taken from the Generalized Amber Force Field (GAFF).

The calculations of the partial charges were performed as described in section 3.1. Prior to the calculations, the GLU488 molecule was cut into three smaller pieces

to avoid convergence problems in both the ESP calculations as well as the RESP fit[8]. However, cutting molecules prior to charge calculations causes problems as it leaves charged ends where the bond was located. To minimize this effect, the ends were capped with neutral molecular caps. The GLU488 pieces were chosen, so that the two cuts only left C- and N- termini like ends. The N-termini ends were then acetylated using ACE caps and the C-termini ends were capped with N-methyl groups (NME). We later restrained the RESP fit to assign zero total charge to these caps. This follows their original parametrisation in the AMBER force field.

Finally, the teLEAP software of the AMBER package was used for automatic force field parameter generation on the three parts. To model the resulting atom type, angle, pair and dihedral parameter changes in the two cut regions, the parameters were modified by hand.

3.1.2. Alexa 647 Dye Parametrisation

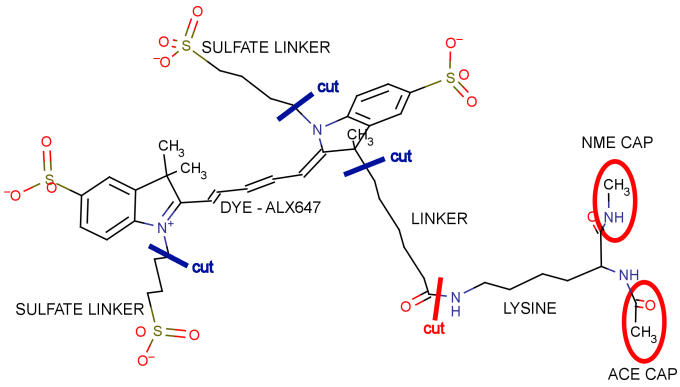


Figure 3.3.: Fragmentation of Alexa 647 prior to the partial charge calculations (green cuts: methyl caps were applied | red cuts: ACE/NME caps were applied)

The Alexa 647 dye was parameterized in analogy to the parametrisation of the Alexa 488 dye. For the Alexa 647 dye, we used residue K56 of the HTH protein for attachment. The result was a new residue, LYS647. The LYS647 residue was separated into five individual pieces prior to charge calculations. Two cuts were placed to separate the dye, the linker and the amino acid. Another two cuts were used to separate the dye from its two flexible sulfate linkers. The cut between lysine and the linker was capped using NME and ACE caps. The other cuts were capped using methyl groups, which were later restrained in the RESP fit to carry neutral overall charge. Force field parameters were generated as described for the GLU488

system. After parametrisation, the parametrisation quality of the Alexa 647 dyes inner aromatic carbon chain was monitored in free MD-simulation for 100 ns.

3.2. Results of Dye Parametrisation

3.2.1. Alexa 488

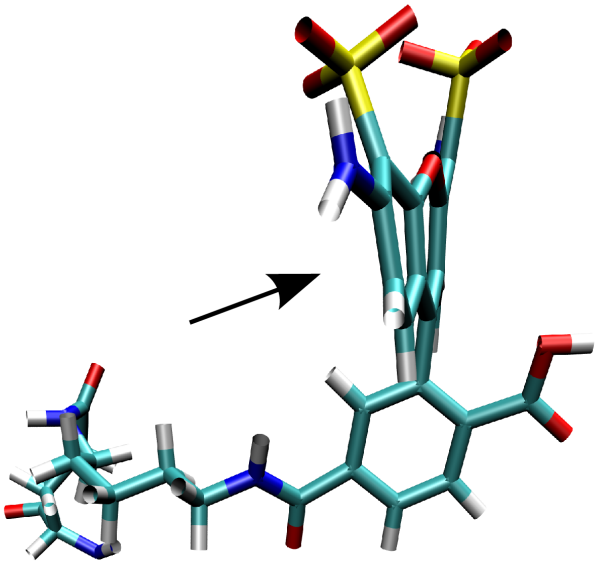


Figure 3.4.: Bending of the aromatic ring - Alexa 488 dye in free MD simulation

The structure optimization of the Alexa 488 dye in vacuum resulted in a slightly tilted conformation. We observed the extended π-electron system to be non non-planar. The tilting of the molecule increased the effective distance between the two charged sulfate groups of the system from 4.6 $\mathring{A}$ to 4.8 $\mathring{A}$. Applying implicit solvent boundary conditions to the structure optimization produced the same tilting with 4.8 $\mathring{A}$ sulfate-sulfate distance.

The same behavior was observed in free MD-simulations of the system. Frequent fluctuations out of the molecular plane were observed within the trajectory time resolution of 1 ps. However, the overall structure of the dye was conserved. See section 5.1 for further information.

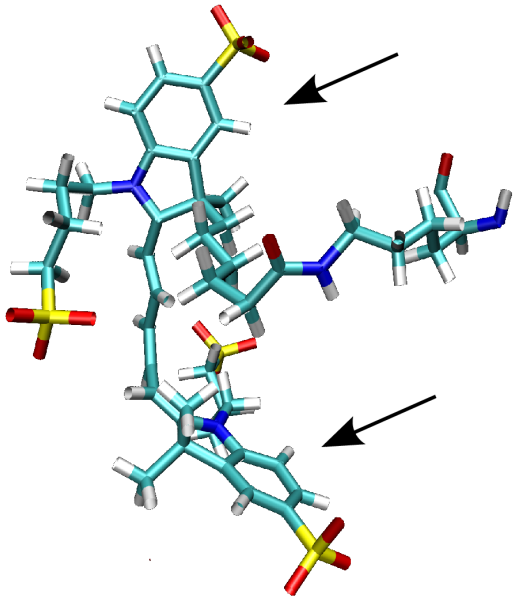


Figure 3.5.: Perpendicular orientation of the Alexa 647 dye in free MD simulation

3.2.2. Alexa 647

The structure optimization of the Alexa 647 dye body without the sulfate linkers resulted in a planar conformation of the dye. The separated sulfate linkers and the linker between dye and amino acid showed planar conformations after structure optimization and good flexibility in free MD-simulation. The MD-simulation of the dye with all linkers removed showed frequent bending and tilting within 100 ns of simulation time. This behavior was enhanced when simulating the whole LYS647 residue embedded into the HTH protein. The planar conformation of the LYS647 residue was not stable. The inner aromatic carbon chain was not conserved in the planer conformation, thus resulted in a tilting and bending of the extended π-electron system. Several events of dye conformations with perpendicular orientations of the two aromatic rings were observed and found stable for several nanoseconds. A more detailed analysis can be found in chapter 5.1.

4. smFRET-MD Simulation

Together with the experimental group, we monitor the dynamics of the HTH motif in both experiment as well as in computer simulation. These two efforts are linked using single molecule Fluorescence Resonance Energy Transfer (smFRET)[32]. To this aim, the experiment is conducted twice, "in vitro" by the experimentalists as well as "in silico" using MD-simulations. The smFRET experimental setup will be able to record data of single photon emission events from two dyes attached to the HTH motif[28, 29]. The FRET donor is Alexa 488 and the FRET acceptor is Alexa 647. The smFRET-MD simulation records all photon emission events in a similar way as the experiment. The smFRET-MD algorithm is based on and extended the previous work of Nicola Lima, Gunnar Schröder and Frank Beierlein et al.[4, 23, 31]

4.1. Fluorescence Resonance Energy Transfer

Fluorescence Resonance Energy Transfer (FRET) experiments measure the FRET rate k_{DA}. This observable is obtained from different experimental setups like fluorescence lifetime measurements, intensities or photon counting.[32, 43] These experiments all ask the same question:

How can molecular distances be obtained from FRET experiments?

The FRET rate enters the Förster theory[15], where it can be written in dependency of the distance R_{DA} as

$$k_{DA} = \frac{\kappa^2}{R_{DA}^6} \frac{9000\ ln(10) Q_D}{\tau_D 128 \pi^5 N n^4} \int_0^\infty F_D(\lambda) \epsilon(\lambda) \lambda^4 d\lambda. \tag{4.1}$$

Here, R_{DA} is the distance between the center of mass of donor and acceptor dye, τ_D is the fluorescence life time of donor dye in absence of the acceptor, N is Avogadro's number, n is the diffraction index of the medium and Q_D is the quantum yield of the donor in absence of the acceptor. The first of the two remaining terms is the orientation factor

$$\kappa^2 \;=\; \left[(\mathbf{D}\cdot\mathbf{A}) - 3(\mathbf{D}\cdot\mathbf{R}_{DA})(\mathbf{A}\cdot\mathbf{R}_{DA})\right]^2, \tag{4.2}$$

with $\mathbf{D}$, $\mathbf{A}$ being the normalized transition dipole vectors of the donor and acceptor dye and $\mathbf{R}_{DA}$ the normalized vector connecting the donor and acceptor centers of mass. This is often rewritten in terms of the angles between the two transition dipole vectors, θ_{DA}, the angle between $\mathbf{R}_{DA}$ and $\mathbf{D}$, θ_D, and the angle between $\mathbf{R}_{DA}$ and $\mathbf{A}$, θ_A,

$$\kappa^2 \;=\; \left[cos(\theta_{DA}) - 3cos(\theta_D)cos(\theta_A)\right]^2. \tag{4.3}$$

The overlap integral between the donor emission spectrum and the acceptor absorption spectrum can be written as

$$J_{DA} \;=\; \int_0^\infty F_D(\lambda)\epsilon(\lambda)\lambda^4 d\lambda. \tag{4.4}$$

For practical calculation of distances, k_{DA} can be written as

$$k_{DA} \;=\; C\frac{\kappa^2}{R_{DA}^6}, \tag{4.5}$$

summarizing all constant terms in C. In the experimental setup, $\langle\kappa^2\rangle$ is often assumed to be 2/3 which is the mean value of κ^2 for isotropic orientation of the dyes. This assumes the distances R_{DA} and the orientation factor κ^2 to be uncorrelated,

$$\langle\frac{\kappa^2}{R_{DA}^6}\rangle \;=\; \frac{\langle\kappa^2\rangle}{\langle R_{DA}^6\rangle}. \tag{4.6}$$

For proteins, the dye molecules often interact with the amino acid residues. Therefore, protein system are likely to have anisotropic orientation factors. A customized orientation factor $\langle\kappa^2\rangle$ is required. The measurement of κ^2 remains problematic, as it is not accessible at timescales of single FRET events. The smFRET simulation of a given system can provide a new orientation factor. However, before the customized orientation factor can be applied to the experimental data, the validity of equation 4.6 needs to be checked. In case of strong correlation between κ^2 and R_{DA} the smFRET simulation can also provide a rough error estimation on the customized $\langle\kappa^2\rangle$. The method of smFRET simulation will be discussed in the next section.

4.2. smFRET Simulation

The established approach for single molecule FRET distance measurements has been described in the previous section 4.1. At first sight, it seems natural to simulate the smFRET experiment "in silico" and compare the calculated distances from the experiment with the distances from the computer simulation. However, this approach adds additional approximations which can be avoided. The focus of our simulation is the simulation of the raw experimental data, the photon distribution of the donor and the acceptor dye. This allows us to directly compare the smFRET *simulation* to the smFRET *experiment* at the photon level. As the result, a microscopic description of the FRET system can be obtained.

The smFRET simulation can also provide further information like additional distance distributions, orientation factors or error estimations for the analysis of the experimental data.

4.2.1. Modeling of the FRET System

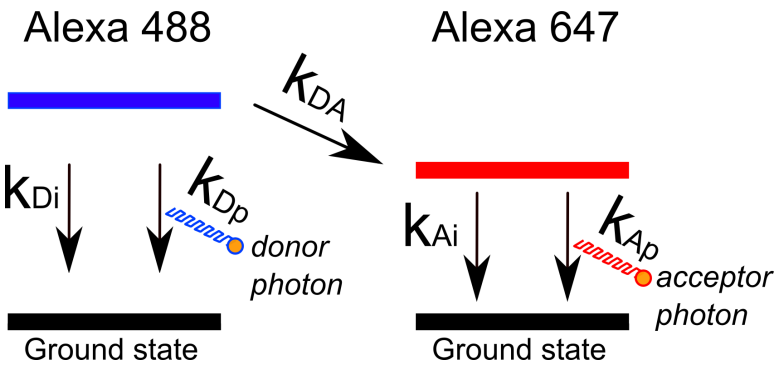


Figure 4.1.: Deactivation diagram of the FRET donor Alexa 488: The rates k_x were used for probability calculations of the deexcitation pathway

We created a program called "smFRET-MD" which takes an MD-simulation trajectory as input and produces photon distributions for donor and acceptor photons as output. The algorithm determines several starting points from the MD trajectory for donor excitation. From these starting points, the deexcitation pathway of the donor dye is simulated based on Förster Theory, see diagram **??**. Each donor excitation results in one out of four different outcomes: 1) a donor photon 2) donor deexcitation without photon emission 3) an acceptor photon 4) acceptor deexcitation without photon emission. Unlike the established algorithm [23], the smFRET-MD

algorithm calculates the four deexcitation probabilities p_x using one picosecond windows: $p_x = k_x \cdot 1ps$, instead of $p_x = k_x \cdot \Delta t$ with increasing Δt values. Here, k_x denotes the different deexcitation rates for event 1)-4). Therefore, the smFRET-MD algorithm calculates correct lifetimes and quantum yields for the FRET dyes which can be used for further analysis. The core donor deexcitation routine reads as follows:

$$
\begin{aligned}
&k_{dp} = 2.24\ 10^{-4}\ [ps^{-1}]; \quad p_{dp} = k_{dp} \cdot 1\ [ps] \\
&k_{di} = 1.95\ 10^{-5}\ [ps^{-1}]; \quad p_{di} = k_{di} \cdot 1\ [ps] \\
&time = rand(in\_traj) \\
&\Delta t = 0 \\
&loop: \\
&\qquad \Delta t = \Delta t + 1 \\
&\qquad k_{fret} = calc\_fret\_rate(time + \Delta t) \\
&\qquad p_{fret} = k_{fret} \cdot 1[ps] \\
&\qquad choice = rand(0, 1) \\
&\qquad if\ choice \leq p_{dp}: \\
&\qquad\qquad donor\_photon\_count + 1 \\
&\qquad\qquad break \\
&\qquad elif\ choice \leq p_{dp} + p_{di}: \\
&\qquad\qquad break \\
&\qquad elif\ choice \leq p_{dp} + p_{di} + p_{fret}: \\
&\qquad\qquad excite\_Acceptor(time + \Delta t) \\
&\qquad\qquad break
\end{aligned}
$$

In this routine, k_{dp} , k_{di}, k_{fret} represent the rates for donor photon emission, donor deexcitation without photon emission and donor fret deexcitation. p_{dp} , p_{di}, p_{fret} are the corresponding probabilities. All rates except for k_{fret} were calculated using matching quantum yields and lifetimes as provided in the Molecular Probes Handbook[2]. The FRET rates were calculated as described in equation 4.1 using an approximated value for the overlap integral. We used the overlap integral from the Alexa 647/Cy5 [24] system for its high similarity to the Alexa 488/Alexa 647 emission and absorption spectra [6].

Additionally, we recorded the deexcitation times τ for each deexcitation event. From this data, the donor lifetime τ_{DA} and the donor quantum yield Q_{DA} in presence of the acceptor can be calculated. The fluorescence lifetime τ is defined as the

average time between donor excitation t_0 and donor photon emission t_p,

$$\tau = \langle t_p - t_0 \rangle \tag{4.7}$$

which was recorded during the smFRET-MD simulation. The quantum yield Q is defined as

$$Q = \frac{\#\ photons\ emitted}{\#\ dye\ excitations} \tag{4.8}$$

and can also be obtained from the recorded deexcitation data.

The FRET system consists of two dyes. Both dyes were monitored during the smFRET-MD simulation while only the FRET donor was deexcitated according to Förster Theory. The FRET acceptor was deexcitated by either photon emission or deexcitation without photon emission. Therefore, the FRET acceptor served as a model system to test the functionality of the simulation program.

4.2.2. Prediction Quality of the smFRET-MD Simulation

In this section we ask the questions: "How good is the prediction quality of our smFRET-MD simulation?"

To answer this question, we compared the calculated distances from the photon trajectories d_{ph} to the "true" distances d_{MD} which were obtained directly from the MD trajectory. Since the MD trajectory showed a rather narrow distance distribution, we compared the mean distance from the MD trajectory $\overline{d}_{MD} = \langle d_{MD} \rangle$ to the calculated mean FRET distances $\overline{d}_{ph}$. To account for the $\langle d^6 \rangle$ averaging error in FRET distance calculations, we calculated a new average according to $\overline{d}_{MD,6} = (\langle d_{MD}^6 \rangle)^{1/6}$.

The prediction quality was tested using two mean distances, $\overline{d}_{yield}$ and $\overline{d}_{lifetime}$, which were calculated from the FRET efficiencies E using the relation:

$$E = \frac{1}{1 + \left(\frac{d}{R_0}\right)^6}. \tag{4.9}$$

The matching FRET efficiencies for the donor quantum yield $E_{D,yield}$ and donor lifetime $E_{D,lifetime}$ in presence of the acceptor read as follows:

$$E_{D,yield} = 1 - \frac{Q_{DA}}{Q_A}, \tag{4.10}$$

$$E_{D,lifetime} = 1 - \frac{\langle \tau_{DA,excitation} \rangle}{\tau_D}. \tag{4.11}$$

In these equations, R_0 is the Förster radius, $\tau_{DA,excitation}$ is the donor lifetime in presence of the acceptor calculated over all deexcitation events instead of just the donor photon emissions, τ_D is the donor lifetime in absence of the acceptor and Q_{DA} is the quantum yield of the donor in presence of the acceptor.

4.2.3. The "Spectroscopic Ruler" Revisited by smFRET-MD

Single molecule FRET experiments have been suggested as a "Spectroscopic Ruler" to measure distances on the molecular level. A study by Schuler et al. [32] used polyproline chains in smFRET experiments to calculate the chain length from measured FRET efficiencies. The measured length of the chain was then compared to the calculated length from the number of proline monomers in the chain. In this section we will use the folded HTH motif in analogy to the polyproline chain in the Schuler experiment. We will also use similar methods for photon distribution analysis and compare our results to the polyproline measurements.

First, we calculated the FRET efficiency E_{av} using the relation

$$E_{av} = \frac{\#Acceptor\ Excitations}{\#Donor\ Excitations} = \frac{n_{a,total}}{n_{d,total}} = \frac{n_{a,p} + n_{a,i}}{n_{d,p} + n_{d,i} + n_{d,fret}}. \tag{4.12}$$

Here, $n_{a,p}$ is the number of acceptor photon emission events and $n_{a,i}$ is the number of radiationless acceptor deexcitation events. $n_{d,p}$ and $n_{d,i}$ are the corresponding numbers for the donor. The donor fret events $n_{d,fret}$ can also be written as $n_{d,fret} = n_{a,p} + n_{a,i}$ since the acceptor can only be excited through the donor.

A limiting case of equation 4.12 was used in the Schuler et al. experiments. For $n_{a,i} \ll n_{a,total}$ and $n_{d,i} \ll n_{d,total}$ the FRET efficiency can be written as

$$E_s = \frac{n_{a,p}}{n_{d,p} + n_{a,p}}. \tag{4.13}$$

E_s was calculated in analogy to the experiment using photon bursts of $n = 100$ photons. Afterwards the mean value of the resulting efficiency distribution $\overline{E}_s$ was compared to the global efficiency $E_{s,av}$ which was obtained from the total donor and acceptor photon counts using equation 4.13. The photon counting method for FRET efficiency calculations has experimental advantages over donor based methods, see equations 4.11 and 4.10, as it increases the total number of photons available for

analysis. Thus, the time resolution can be increased.

In favor of the photon counting method, we suggest a different counting approach. This approach conserves the increased sampling and experimental accessibility yet increases the accuracy for distances $d < R_0$. The main idea is to rewrite equation 4.13 to contain the radiationless acceptor deexcitation events:

$$E_g = \frac{n_{a,p} + n_{a,i}}{n_{d,p} + n_{a,p} + n_{a,i}}. \tag{4.14}$$

The FRET efficiency can now be rewritten using the acceptor quantum yield Q_A to estimate $n_{a,p} + n_{a,i}$:

$$n_{a,p} + n_{a,i} = \frac{n_{a,p}}{Q_A}. \tag{4.15}$$

The result is a transfer efficiency that contains terms for the radiationless acceptor deexcitation

$$E_g = \frac{n_{a,p} + n_{a,i}}{n_{d,p} + n_{a,p} + n_{a,i}} = \frac{\frac{n_{a,p}}{Q_A}}{n_{d,p} + \frac{n_{a,p}}{Q_A}}. \tag{4.16}$$

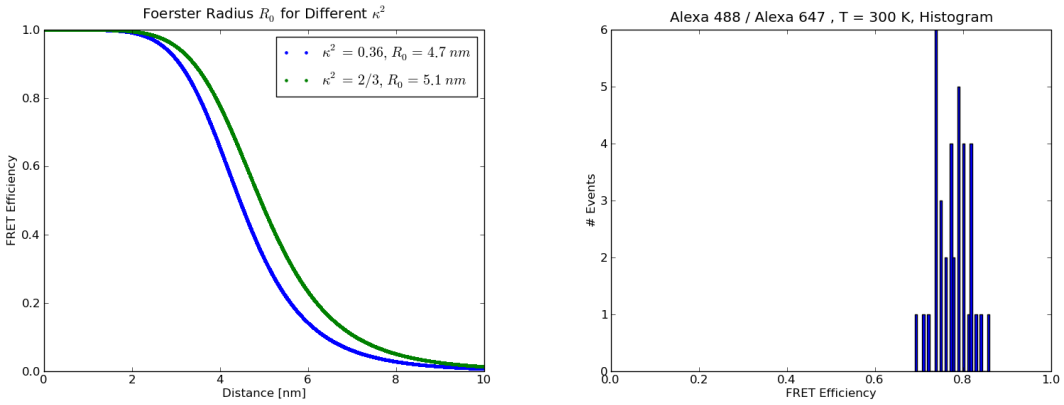
This still assumes $n_{d,i} \ll n_{d,total}$ for the donor dye. The accuracy of this approximation increases with increasing FRET rates and larger donor quantum yields. The donor used in this simulation is the Alexa 488 which has very high quantum yield of $Q_D = 0.92$.

4.3. smFRET-MD Results

4.3.1. The smFRET-MD Program

During the test run, the acceptor lifetime $\tau_A = 1\ ns$ as well as the acceptor quantum yield $Q_A =\ 0.33$ could correctly be reproduced using equations 4.7 and 4.8. The productive run was based on the sampling of the 100 ns MD trajectory and a total of 10000 donor excitations in the smFRET-MD simulation. The raw smFRET-MD simulation data is shown in figure B.1. The simulation also produced donor and acceptor photon distributions which were compared to the FRET rate in figures B.2 and B.3. The donor photon density distribution showed anticorrelation between the FRET rate and the donor photon emission.

From these photon distributions we predict a donor lifetime in presence of the acceptor of $\tau_{DA} = 380 \pm 10\ ps$ and a donor quantum yield of $Q_{DA} = 0.08 \pm 0.01$. Based on 100 ns of sampling, we also obtained an anisotropic orientation factor



(a) (b)

Figure 4.2.: (a) FRET efficiencies calculated for isotropic dye orientation $\kappa^2 = 2/3$ [green] and from the MD trajectory $\kappa^2 = 0.36$ [blue] (b) FRET efficiency histogram after 10000 donor excitation (photon bin size $n = 100$)

$\langle\kappa^2\rangle = 0.36$ from the smFRET-MD simulation.

4.3.2. Prediction Quality of the smFRET-MD Simulation

	Efficiency	Distance [nm]
$\overline{d}_{MD}$		2.6 ± 0.2
$\overline{d}_{MD,6}$		2.7 ± 0.2
$E_{D,yield}$	0.91	3.2
$E_{D,lifetime}$	0.91	3.2

Table 4.1.: smFRET-MD prediction quality measured using different FRET efficiencies: (1-2) Distances from the MD trajectory (3) Efficiency and mean distance calculated from the donor yield (4) Efficiency calculated from the donor lifetime, average over all deexcitation events

We studied the accuracy of the smFRET-MD simulation by calculating the FRET efficiencies $E_{D,yield}$ and $E_{D,lifetime}$. The calculation included radiationless donor deexcitation events which are usually inaccessible in smFRET experiments, yet result in higher accuracies of the calculated values. Both FRET efficiencies resulted in a distance of 3.2 nm which was compared to the d^6 averaged distance from the MD simulation $\overline{d}_{MD,6} = 2.7 \pm 0.2\ nm$.

4.3.3. The "Spectroscopic Ruler" Revisited by smFRET-MD

	Efficiency	Distance [nm]
$\overline{d}_{MD}$		2.6 ± 0.2
$\overline{d}_{MD,6}$		2.7 ± 0.2
$\overline{E}_s$	0.78±0.04	3.8± 0.1
E_{av}	0.78	3.8
$E_{av,Cy5}$	0.75	3.9
$E_{D,lifetime,phot}$	0.88	3.4
E_g	0.91	3.1

Table 4.2.: (1-2) Distances from MD trajectory (3-4) FRET efficiencies from photon counting (5) FRET efficiency with Cy5 as acceptor dye (6) FRET efficiency calculated from the donor lifetime considering only photon emission events (7) Corrected photon counting method, including radiationless acceptor deexcitation from Q_A

The orientation factor was used to calculate a distance-efficiency diagram, as shown in figure 4.2a. This diagram can be used to interpret the FRET efficiency distribution from the photon counting approach E_s, shown in figure 4.2b. The average efficiency for the photon counting approach was calculated $\overline{E}_s = 0.78 \pm 0.04$ which is considerably lower than the values for $E_{D,yield}$ and $E_{D,lifetime}$. Since radiationless deexcitations are difficult to access experimentally, we also calculated the FRET efficiency from the donor photon lifetimes without radiationless deexcitation events $E_{D,lifetime,phot} = 0.88$. This value is experimentally accessible, yet more accurate than the photon counting calculation.

The suggested correction for the radiationless acceptor deexcitation using the acceptor quantum yield resulted in a FRET efficiency of $E_g = 0.91$. This is in good agreement with the $E_{D,yield}$ and $E_{D,lifetime}$ efficiencies from the last section. E_g was calculates from $n_{d,i} = 73$ radiationless donor deexcitations, $n_{d,p} = 845$ donor photon emission events and $n_{d,fret} = 9073$ energy transfers events.

5. Discussion

5.1. Discussion of the Simulation Preparation

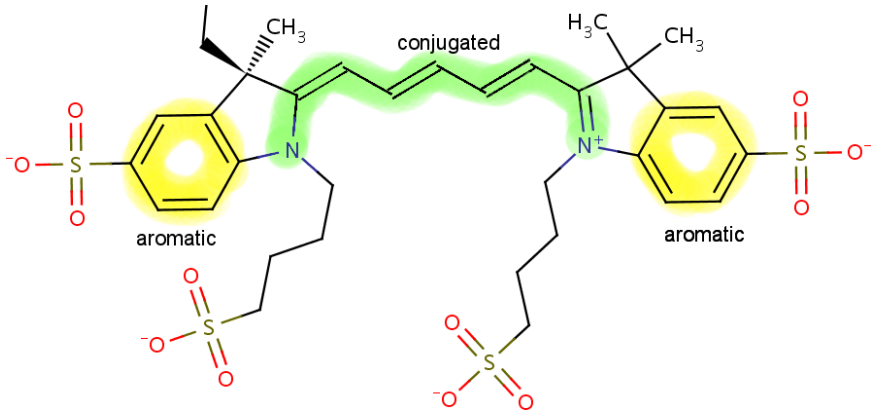


Figure 5.1.: Alexa 647 teLEAP aromaticity classification: (yellow: classified as aromatic atoms | green: classified as conjugated atoms)

The parametrisation of new molecules into MD force fields is a challenging task. Most parameter sets will produce a functional force field which simulates molecular motion in an MD simulation. However, it is difficult to determine whether the MD simulation describes a physical process or just large amounts of expensive random numbers. We parameterized the Alexa 488 and the Alexa 647 dye into the Generalized Amber Force Field. GAFF is an established force field for computer aided drug design, a field that strongly relies on the parametrisation of custom molecules on large scales. During our work, we realized that dye molecules with large extended π-electron systems and several free charges are more challenging to parameterize than small drug ligands.

The result of parameter generation for the Alexa 647 dye is shown in figure 5.1. We see a mixture of aromatic carbon rings and conjugated carbon chains. The aromatic rings were correctly describes using the same bonds and atom types for all carbon atoms, representing delocalized electrons. This was not the case for carbon chain

between the to aromatic rings. Here, two different atom types were assigned. The consequence is a localization of the π-electrons, where they should be delocalized. We corrected this missclassification by relabeling all atoms in the chain region to have delocalized π-electron systems. The modifications could easily be done for both the Alexa 488 and the Alexa 647 dye. However, the integration of the new parameters into the protein residues is a very demanding task. Therefore, all MD simulations for this thesis had to be performed using the old parameter set. The result is an overestimation of the dye flexibility, see section 3.2.2.

The MD simulation itself is very promising as we were able to simulate an average of 16 ns/day. This result was obtained without parameter optimization which is likely to improve this value to 20ns/day using a tuned system setup. This will allow us to increase the conformational sampling of the system to the microsecond range. For this thesis, only short 100 ns trajectories were calculated due to the short timeframe of the Bachelor thesis. We plan to simulate the system at different temperatures with trajectory lengths $> 1\ \mu s$ each.

A more sophisticated approach than the MD simulation would be a replica exchange MD simulation setup covering the entire temperature range from the experiment. Replica exchange simulations (REX) drastically increase the conformational sampling of the simulation system, yet the time information is lost. We tested the REX sampling at higher temperatures by simulating several short trajectories between 290 and 360 K. The simulations showed an increased sampling of the dyes which suggest that REX could improve the quality of the photon distribution. We tested the current smFRET-MD algorithm on shredded trajectories of these short runs. However, due to the strong time dependency of the algorithm we did not reach converge.

There is no X-ray structure of the folded HTH motif and only one partially folded set of NMR structures in the protein data bank. A folded structure was obtained from free MD simulation of the partially folded NMR structure for 100 ns using 4 M NaCl concentration. This folded structure has not been compared to any reference structure and might not represent the native state of the HTH motif under the given conditions. Further structure studies are desired.

5.2. Discussion of the smFRET-MD Simulation

We begin the discussion of the smFRET-MD simulation by reminding the reader that the sampling of the MD trajectory used for testing the smFRET-MD program is not sufficient to make concrete predictions about experimental values! Therefore, the predicted values for the donor lifetime $\tau_{DA} = 380 \pm 10\, ps$ and the donor quantum yield in presence of the acceptor $Q_{DA} = 0.08 \pm 0.01$ have to be taken with great caution. Instead, these values should be taken as a demonstration of the capabilities of the smFRET-MD algorithm. We were able to greatly improve the benefit of the existing smFRET simulation algorithm. Unlike the previous implementation which produced unphysical lifetimes, we are now able to calculate dye lifetimes and extract quantum yields for the donor and acceptor dyes.

Additionally, we also implemented a new routine which approximates the acceptor deexcitation without photon emission:

$$E_g = \frac{n_{a,p} + n_{a,i}}{n_{d,p} + n_{a,p} + n_{a,i}} = \frac{\frac{n_{a,p}}{Q_A}}{n_{d,p} + \frac{n_{a,p}}{Q_A}}. \tag{5.1}$$

Let us briefly revise this approach in the following paragraph. The radiationless energy transfer from the donor to the acceptor is described by Förster Theory. Once excited, the acceptor follows a deexcitation pathway independent of Förster Theory. Therefore, acceptor quantum yield and acceptor lifetime are not affected. This is not the case for the donor, here quantum yield and lifetime change due to FRET quenching. When calculating FRET efficiencies, it is desired to use photons from both the donor and the acceptor. The established approach in the Schuler et al. experiment [32] used

$$E_{fret} = \frac{n_{d,photons}}{n_{a,photons} + n_{d,photons}}. \tag{5.2}$$

While looking at the FRET system from Förster Theory, the efficiency should be defined as

$$E_{fret} = \frac{\#\ Acceptor\ Excitations}{\#\ Donor\ Excitations} = \frac{n_{a,p} + n_{a,i}}{n_{d,p} + n_{d,i} + n_{a,p} + n_{a,i}}, \tag{5.3}$$

including radiationless donor and acceptor deexcitations, see equation 4.12. Based

on this line of thought, a better approximation can be made

$$E_g = \frac{n_{a,p} + n_{a,i}}{n_{d,p} + n_{a,p} + n_{a,i}} = \frac{\frac{n_{a,p}}{Q_A}}{n_{d,p} + \frac{n_{a,p}}{Q_A}}. \tag{5.4}$$

This only neglects the radiationless deexcitation of the donor. Therefore, one limiting case of the E_g approximation are low FRET rates which will result in low $n_{a,p}$ and $n_{a,i}$ values but high $n_{d,i}$ values which are coupled to Q_D. This leads to an overestimation of low FRET efficiencies. Whereas the original approach, equation 5.2, will also suffer from underestimating high FRET efficiencies due to large values of $n_{a,i}$.

The modification seems rather simple and has not been discussed sufficiently with other group members as of writing this thesis. There might be experiments which have already used this correction or studies that proved this approach inaccurate. However, we find this modification rather appealing as it might explain the experimental observation from the original Schuler et al. study [32]. In this study, an overestimation of low efficiencies and an underestimation of high FRET efficiencies was reported. This is in agreement with our smFRET-MD simulations and theoretical considerations of limiting cases.

The results from the smFRET-MD simulation test run were in good agreement with the distances from the MD simulation trajectory. A dye distance of 3.2 nm was calculated from Förster Theory using donor lifetimes and donor quantum yields while the much more accurate measurement from the MD simulation suggested 2.7 ± 2 nm. Given the many approximations that enter our simulation this result represents the maximal accuracy of the smFRET-MD program. The overlap integral had to be approximated from the the Alexa 647/Cy5 system and the diffraction index of the medium was approximated to be 1.4. There are also several approximations that enter Förster Theory itself like the ideal dipole approximation. Due to the short length of the trajectory compared to the large number of photons, oversampling becomes an issue, too. However, given the simplicity of the algorithm and the quality of the distance prediction we consider the smFRET-MD simulation a working algorithm for smFRET simulations.

In the future, it will be highly interesting to compare photon distributions from better sampling MD simulations directly to measured photon distributions from smFRET experiments. The quality of the smFRET-MD simulation gives us reason to believe that, given proper sampling, it is possible to predict smFRET photon

distributions “in silico”.

A. Appendix I

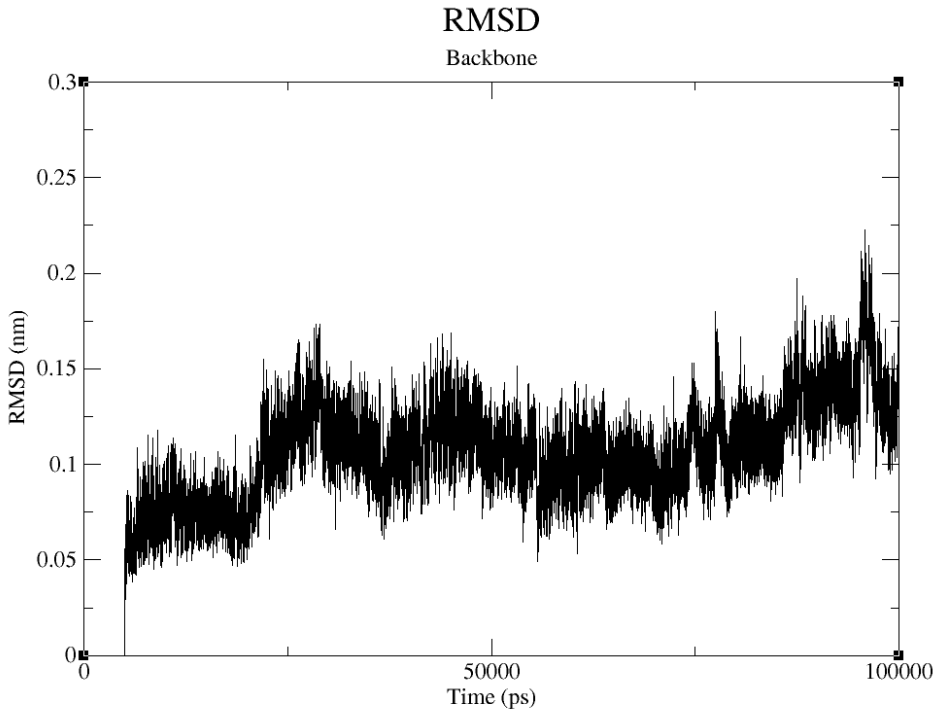


Figure A.1.: The backbone RMSD of residues T27-A54 (helix-turn-helix) is shown as a function of simulation time. The first frame after equilibration was taken as the reference structure.

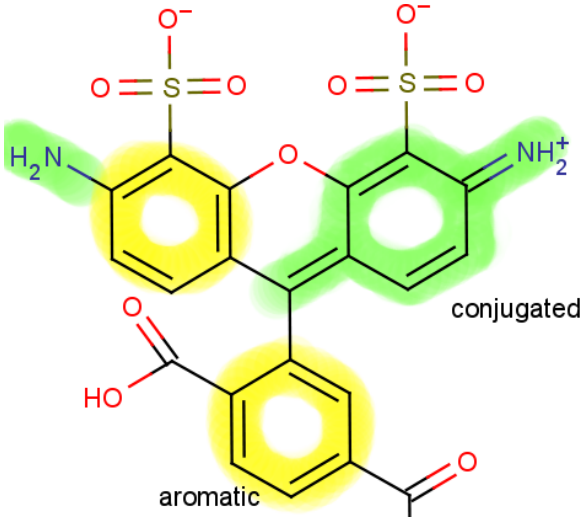


Figure A.2.: Alexa 488 teLEAP aromaticity missclassification: (yellow: classified as aromatic atoms | green: classified as conjugated atoms)

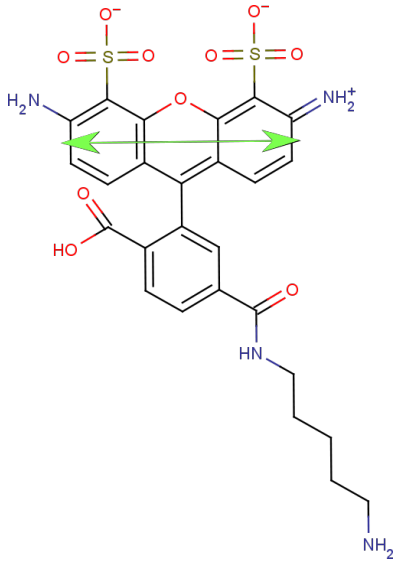


Figure A.3.: Calculated transition dipole vector for the Alexa 488 dye using a small basis set CIS/3-21G*

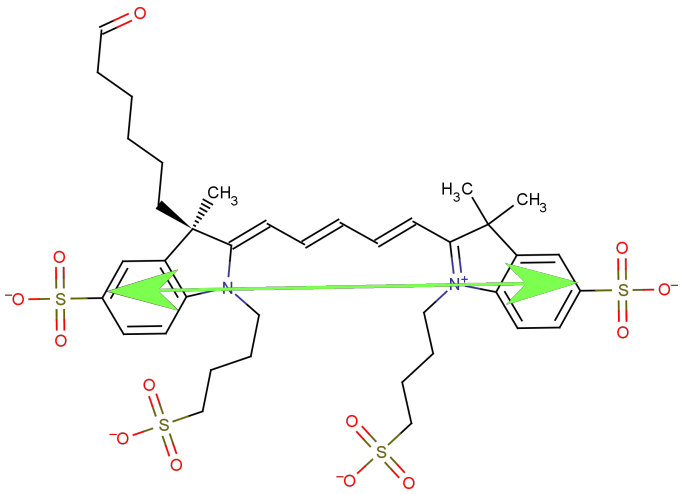


Figure A.4.: Calculated transition dipole vector for the Alexa 647 dye using a small basis set CIS/3-21G*

B. Appendix II

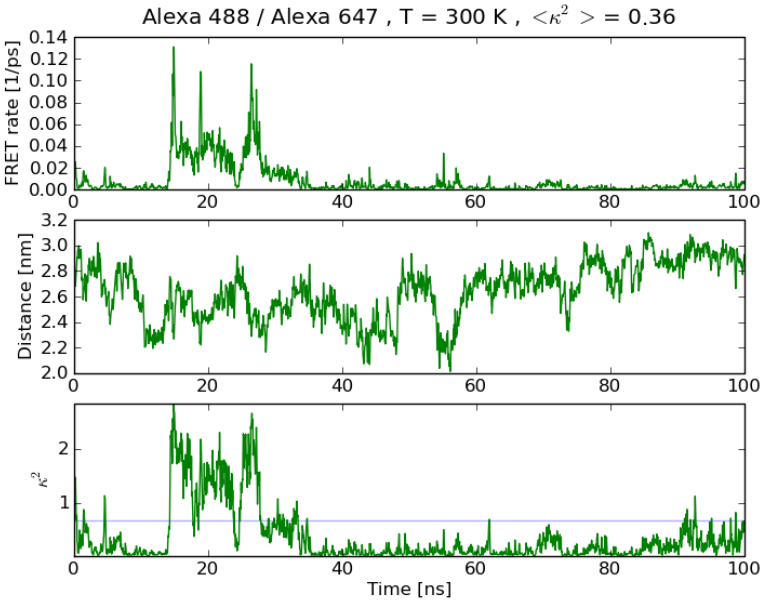


Figure B.1.: (top) FRET rates from smFRET-MD simulation (middle) Alexa 488/Alexa 647 distance from MD trajectory (bottom) Orientation factor κ^2

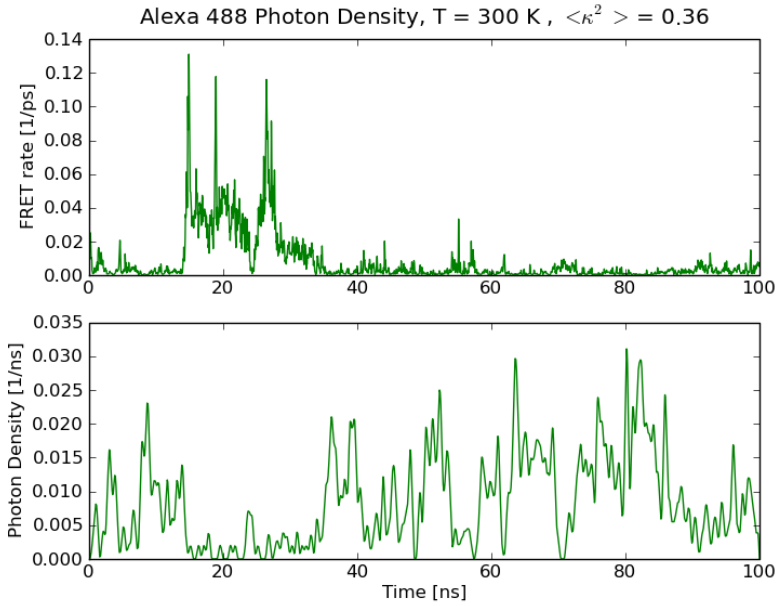


Figure B.2.: (bottom) Donor photon distribution (top) FRET rates from smFRET-MD simulation

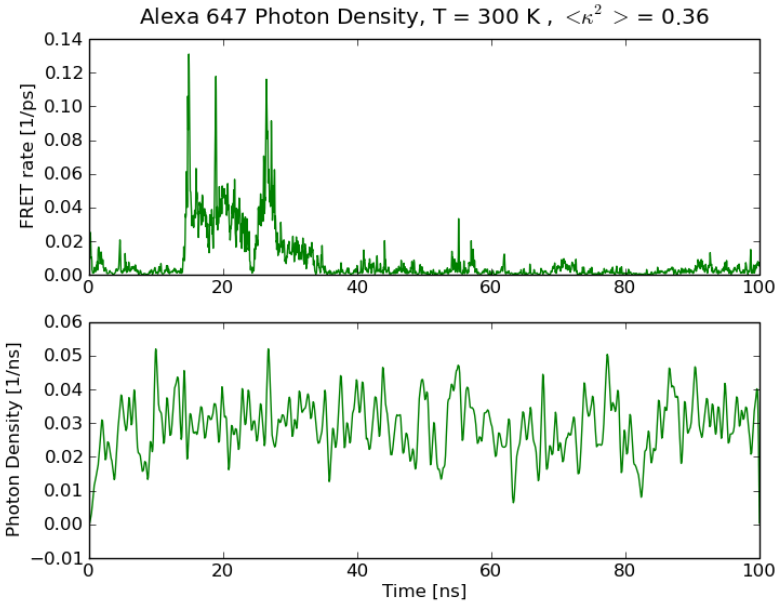


Figure B.3.: (bottom) Acceptor photon distribution (top) FRET rates from smFRET-MD simulation

Bibliography

[1] The nobel prize in chemistry 2003. URL http://nobelprize.org/nobel_prizes/chemistry/laureates/2003/public.html.

[2] Invitrogen 2009. URL http://www.invitrogen.com/site/us/en/home/References/Molecular-Probes-The-Handbook/tables/Fluorescence-quantum-yields-and-lifetimes-for-Alexa-Fluor-dyes.html.

[3] Marvin 5.2.11. Chemaxon, 2009. URL http://www.chemaxon.com.

[4] Frank R Beierlein, Olaf G Othersen, Harald Lanig, Siegfried Schneider, and Timothy Clark. Simulating fret from tryptophan: is the rotamer model correct? *J Am Chem Soc*, 128(15):5142–5152, Apr 2006. doi: 10.1021/ja058414l. URL http://dx.doi.org/10.1021/ja058414l.

[5] HJC Berendsen, JPM Postma, WF Van Gunsteren, A. DiNola, and JR Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684, 1984.

[6] J.E. Berlier, A. Rothe, G. Buller, J. Bradford, D.R. Gray, B.J. Filanoski, W.G. Telford, S. Yue, J. Liu, C.Y. Cheung, et al. Quantitative comparison of long-wavelength Alexa Fluor dyes to Cy dyes: fluorescence of the dyes and their bioconjugates. *Journal of Histochemistry and Cytochemistry*, 51(12):1699, 2003.

[7] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus. *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations Supported in part by grants from the National Science Foundation and the National Institutes of Health.*, volume 4. John Wiley & Sons, Inc. New York, 1983.

[8] Brent Kruegera Ross Walker Bryan Lelanda, David Paula. Amber advanced tutorials tutorial a1. Technical report, San Diego Supercomputer Center, University of California, San Diego Dept. of Chemistry, Hope College, 2008.

[9] H.S. Chan and K.A. Dill. The protein folding problem. *Physics today*, 46(2): 24–32, 1993.

[10] B. L. de Groot and H. Grubmüller. Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and glpf. *Science*, 294(5550):2353–2357, Dec 2001. doi: 10.1126/science.1062459. URL `http://dx.doi.org/10.1126/science.1062459`.

[11] Y. Duan and P.A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282(5389):740, 1998.

[12] C.J. Epstein, R.F. Goldberger, and C.B. Anfinsen. The genetic control of tertiary protein structure: studies with model systems. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 28, page 439. Cold Spring Harbor Laboratory Press, 1963.

[13] U. Essmann, L. Perera, M.L. Berkowitz, T. Darden, H. Lee, and L.G. Pedersen. A smooth particle mesh Ewald method. *Journal of Chemical Physics*, 103(19): 8577–8593, 1995.

[14] A.E. Frisch, M.J. Frisch, and G.W. Trucks. *Gaussian 03 User's Reference*. Gaussian Wallingford, CT, 2003.

[15] T. Förster. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik*, 437, 1948.

[16] Helmut Grubmüller. Theorie und Simulation induzierter Konformationsdynamik von Proteinen, 2002.

[17] NSF Science Harvey Rubin, David Harlan Wood, Technology Center in Discrete Mathematics, and DIMACS (Group) Theoretical Computer Science. *DNA based computers III: DIMACS Workshop*. AMS Bookstore, 1999.

[18] B. Hess. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput*, 4(1):116–122, 2008.

[19] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, 2008.

[20] W.L. Jorgensen, D.S. Maxwell, and J. Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Q. ReV. Biophys*, 26:49, 1993.

[21] JAMES W. CALDWELL PETER A. KOLLMAN DAVID A. CASE JUNMEI WANG, ROMAIN M. WOLF. Development and testing of a general amber force field. *Development and Testing of a General Amber Force Field*, 25:1157–1174, 2004.

[22] John L Klepeis, Kresten Lindorff-Larsen, Ron O Dror, and David E Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol*, 19(2):120–127, Apr 2009. doi: 10.1016/j.sbi.2009.03.004. URL `http://dx.doi.org/10.1016/j.sbi.2009.03.004`.

[23] Nicola Lima. Fret studies on polyproline. (unpublished).

[24] A.A. Martí, X. Li, S. Jockusch, N. Stevens, Z. Li, B. Raveendra, S. Kalachikov, I. Morozova, J.J. Russo, D.L. Akins, et al. Design and characterization of two-dye and three-dye binary fluorescent probes for mRNA detection. *Tetrahedron*, 63(17):3591–3600, 2007.

[25] Xavier Michalet, Shimon Weiss, and Marcus Jäger. Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem Rev*, 106(5):1785–1813, May 2006. doi: 10.1021/cr0404343. URL `http://dx.doi.org/10.1021/cr0404343`.

[26] S. Miyamoto and P.A. Kollman. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):3985, 1992.

[27] D.A. Pearlman, D.A. Case, J.W. Caldwell, W.S. Ross, T.E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, 1995.

[28] T. L. Religa, J. S. Markson, U. Mayor, S. M V Freund, and A. R. Fersht. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437(7061):1053–1056, Oct 2005. doi: 10.1038/nature04054. URL `http://dx.doi.org/10.1038/nature04054`.

[29] Tomasz L Religa, Christopher M Johnson, Dung M Vu, Scott H Brewer, R. Brian Dyer, and Alan R Fersht. The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of engrailed homeodomain. *Proc Natl Acad Sci U S A*, 104(22):9272–9277, May 2007. doi: 10.1073/pnas.0703434104. URL `http://dx.doi.org/10.1073/pnas.0703434104`.

[30] Y.M. Rhee, E.J. Sorin, G. Jayachandran, E. Lindahl, and V.S. Pande. Simulations of the role of water in the protein-folding mechanism. *Proceedings of the National Academy of Sciences*, 101(17):6456–6461, 2004.

[31] Gunnar Schröder. *Simulation of Fluorescence Spectroscopy Experiments*. PhD thesis, University of Göttingen, 2004.

[32] Benjamin Schuler, Everett A Lipman, Peter J Steinbach, Michael Kumke, and William A Eaton. Polyproline and the "spectroscopic ruler" revisited with single-molecule fluorescence. *Proc Natl Acad Sci U S A*, 102(8):2754–2759, Feb 2005. doi: 10.1073/pnas.0408164102. URL `http://dx.doi.org/10.1073/pnas.0408164102`.

[33] M.M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *Journal of molecular biology*, 354(1):173–183, 2005.

[34] C.D. Snow, H. Nguyen, V.S. Pande, and M. Gruebele. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420:102–106, 2002.

[35] C.D. Snow, E.J. Sorin, Y.M. Rhee, and V.S. Pande. How well can simulation predict protein folding kinetics and thermodynamics? 2005.

[36] D. F. OGLETREE D. S. CHEMLA P. R. SELVIN T. HA, TH. ENDERLE and S. WEISS. Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *PNAS*, 93:6264–6268, 1996.

[37] U. Mayor S. M. V. Freund & A. R. Fersht T. L. Religa, J. S. Markson. Solution structure of a protein denatured state and folding intermediate. *Nature*, 437: 1053, 2005.

[38] S.A. Teichmann and N.V. Grishin. Sequences and topology: from genome structure to protein structure. *Current Opinion in Structural Biology*, 18(3): 340–341, 2008.

[39] Dung M. Vu Scott H. Brewer R. Brian Dyer Tomasz L. Religa, Christopher M. Johnson and Alan R. Fersht. The helix–turn– helix motif as an ultrafast independently folding domain: The pathway of folding of engrailed homeodomain. *PNAS*, 104:9272–9277, 2007.

[40] D. van der Spoel and E. Lindahl. Brute-force molecular dynamics simulations of villin headpiece: Comparison with NMR parameters. *Journal of Physical Chemistry B-Condensed Phase*, 107(40):11178–11187, 2003.

[41] WF Van Gunsteren and HJC Berendsen. A leap-frog algorithm for stochastic dynamics. *Molecular Simulation*, 1(3):173–185, 1988.

[42] WF Van Gunsteren, SR Billeter, AA Eising, PH Hunenberger, P. Kruger, AE Mark, WR Scott, and IG Tironi. *In Biomolecular simulation: The GROMOS96 manual and user guide. 1996.*

[43] Darren B VanBeek, Matthew C Zwier, Justin M Shorb, and Brent P Krueger. Fretting about fret: correlation between kappa and r. *Biophys J*, 92(12):4168–4178, Jun 2007. doi: 10.1529/biophysj.106.092650. URL `http://dx.doi.org/10.1529/biophysj.106.092650`.

[44] J. Wang, W. Wang, P.A. Kollman, and D.A. Case. Antechamber: an accessory software package for molecular mechanical calculations. *J. Am. Chem. Soc*, 222:U403–U403, 2001.

[45] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *J Comput Chem*, 25(9):1157–1174, Jul 2004. doi: 10.1002/jcc.20035. URL `http://dx.doi.org/10.1002/jcc.20035`.

Danksagung

I would like to thank Helmut for providing such a great group environment and for providing me with full access to all of the group's resources including his own time and discussions throughout the last months. I would also like to thank Nicola for kindly introducing me to his work and the rich discussions we had. The same applies to Carsten and Sören who showed great patience especially during my the first weeks here. And Carsten, next time I promise I will read that genion manpage before I keep you in office on a Friday afternoon scrolling through the source code searching for that imaginary bug of mine. Greetings to the experimental group around Jörk Enderlein and Qui Van, who do all the hard experimental work and supplied me with the topic and ideas for this thesis. Special thanks to the system administrators who keep the clusters up and running and prevent this enormous computing power from becoming self aware. I would also like thank the rest of the group for the high level input I received in so many discussions, especially Jan who helped me boost the smFRET-MD algorithm last minute.

Last but not least I would like to thank my mother who supported me both mentally as well as finacially throughout my studies. Special thanks to Finja who did not get to see me very much during my thesis work. I hope I can make it up to you!

Erklärung nach §13(8)der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den February 11, 2011

(Timo Graen)