Simulation of Fluorescence Spectroscopy Experiments

Dissertation

Zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Georg-August-Universität zu Göttingen

Vorgelegt von

Gunnar Schröder

aus Bremen

Göttingen, 2004

D7
Referent: Prof. Dr. T. Salditt
Korreferent: PD Dr. H. Grubmüller
Tag der mündlichen Prüfung:

Contents

"I accept the universe!"
– Margaret Fuller

1

Introduction

Proteins are abundant in all organisms and are involved in almost all cellular processes. They have functions such diverse as building the cytoskeleton, catalyzing chemical reactions, controlling cell signaling, performing muscle contraction, and many others.[1,2] Structural biology aims at gaining insight into the function of proteins by determining their three-dimensional structures, which is typically done by x-ray crystallography[3,4] or nuclear magnetic resonance (NMR) spectroscopy.[5,6] Furthermore, protein motions, particularly their conformational dynamics, regulate and often constitute protein function. Therefore, a large variety of experimental and theoretical techniques aims at probing the internal dynamics of proteins, with a particular focus at the picosecond to microsecond timescale.[3]

NMR,[7] electron paramagnetic resonance (EPR),[8] neutron scattering,[9,10] as well as fluorescence spectroscopy[11–14] have indeed provided much insight in this respect. Fluorescence spectroscopy, in particular, in combination with site-directed fluorescent labeling became an established tool in molecular biology and biochemistry to investigate the dynamics and interactions of biomolecules.[15] The fluorescent labels, or dyes, are bound to proteins typically via cysteines. To attach a dye to a specific position in the protein, the present amino acid at this position is therefore mutated into a cysteine prior to labeling. Two of the most important optical techniques are fluorescence resonance energy transfer (FRET)[16] and fluorescence anisotropy (or depolarization),[17–19] which both will be studied in detail within this work and will be briefly be introduced in Chap. 2.

FRET is a distance-dependent interaction between two dye molecules in which electronic excitation is transferred from an excited donor molecule to

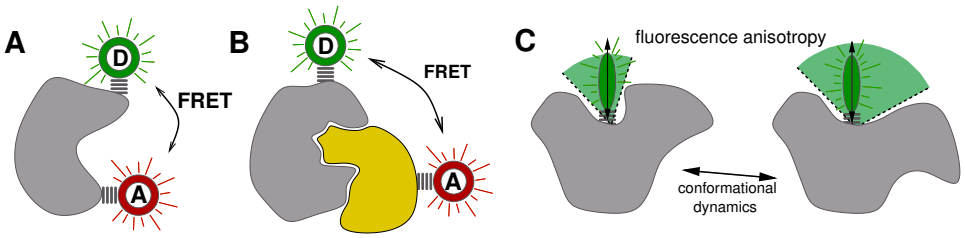

Figure 1.1: Fluorescence spectroscopy in combination with site-directed labeling can be used to study structure and dynamics of single biomolecules. (A) and (B) Fluorescence resonance energy transfer (FRET) experiments measure the distance between two dyes. Determination of *intra*molecular distances (A) reveals conformational changes and *inter*molecular distances give information on interactions between differently labeled biomolecules. (C) The motional freedom of a protein-attached dye, indicated by the green cones, depends on the conformational dynamics of the protein to which the dye is attached. Fluorescence anisotropy experiments therefore detect conformational changes.

an acceptor molecule (see Fig. 1.1 A and B). The FRET efficiency, which is obtained from the separately detected light intensities of the donor and the acceptor, depends on the distance (as well as on the relative orientation of the dyes) and is sensitive at a length-scale of about 10–75 Å.[20,21] Thus, FRET experiments allow to determine intra- and intermolecular distances within and between biomolecules.[11,14] Moreover, the labeling with FRET pairs of many different protein sites yields multiple intramolecular distances, which then can be used to build a three-dimensional model of the protein via triangulation.[22,23]

Single molecules

For a few years, it is possible to measure *individual* fluorescent photons from a *single* dye molecule in solution. Big advances have been achieved since the first successful detection of a molecule labeled with multiple fluorophores by Hirschfeld in 1976.[24] Several groups have contributed to the improvement of single-molecule detection techniques with only a single fluorophore.[25–34] Single-molecule methods became particularly popular in the last years, because they offer the chance to determine *distributions* of observables, like in this case the distance between two dyes, rather than just *ensemble averages* as obtained from conventional bulk measurements. Therefore, subpopulations of distances can be resolved, which reveals exciting insights into conformational substates and dynamics.[11,20,35]

Single molecule detection in fluorescence spectroscopy is achieved by using

a confocal microscope setup to focus a laser beam to a very tiny excitation volume of about one femtoliter. The dilution of the dye molecules is chosen to maximize the number of events, where only one single dye passes through the excitation volume. For a dye attached to a protein of e. g. 30 kD, the mean passage time is in the millisecond range and is longer for larger proteins, due to the larger diffusion coefficient. Typical dyes used in fluorescence spectroscopy have a fluorescence lifetime of a few nanoseconds. Therefore, during the passage of the dye-protein system through the laser focus, several hundred measurements can be perfomed, yielding a few hundred arrival times of the detected photons, all emitted by the same dye.

The small number of obtained photons is the main drawback from which single-molecule experiments suffer in general, which implies low statistical accuracy. The additional information on the distribution of the measured quantity has thus to be paid for by an increase of the statistical uncertainty. Therefore, methods have to be developed, which particularly account for limited or noisy data from single-molecule experiments. In this context, maximum likelihood approaches have been successfully applied to several related problems,[36–42] but not to the case at hand. In this work, we will therefore use a maximum likelihood approach to address the problem of distance determination from single-molecule FRET experiments.

Time-resolved distance

One example of particularly noisy data is the detection of photon arrival times from single molecule fluorescence experiments.[43–45] Fluorescence intensity *variations* are obtained from these photon arrival times,[23,46,47] which allow to track distance changes $R(t)$ between the two dyes, and hence to monitor conformational motions of the studied biomolecule.[46,47] If, however, one wants to achieve millisecond time-resolution,[23,48,49] only very few photons are available to determine the distance, therefore the statistical noise is considerable.

In the conventional analysis, the required FRET intensities are computed from photon counts in time windows.[23,50,51] For a typical window size of 1 ms, however, the small number of only $10 \ldots 50$ photons per window[23] implies considerable statistical uncertainty ('shot noise'[52]) and thus limits the time resolution for $R(t)$. Furthermore, the choice of the window size is somewhat arbitrary and only guided by the requirement to trade off shot noise and time resolution. Finally, the traditional method saliently assumes a uniform *a priori* probability for the FRET *intensities* (rather than for the distances). Therefore, and contrary to what one might intuitively assume at first sight, the traditional method cannot be considered a model-free ap-

proach, but introduces a non-physical bias to the distance measurement. In Chap. 3, we thus develop a maximum-likelihood theory to reconstruct $R(t)$ from the photons recorded in single molecule FRET measurements, which does not suffer from any unwanted bias and additionally yields rigorous error bounds.

Time-resolved fluorescence anisotropy

In contrast to FRET, which probes the distance between two dyes, fluorescence anisotropy experiments probe the *rotational motion* of one dye. This is achieved by exciting the dye, or an ensemble of dyes, using a short polarized laser pulse. If the dye has a certain mobility, it will emit the photon in a possibly changed orientation after typically a few nanoseconds. The changed polarization of the emitted light with respect to the exciting light is detected in the experiment. Time-resolved fluorescence anisotropy experiments therefore provide information on both, motional freedom and dynamics of a fluorophore.[53,54] This can be exploited to probe the local environment of the dye, e, g. , a protein, which affects the rotational motion of the dye.

Fluorescence anisotropy experiments are frequently used to study protein conformational dynamics taking advantage of naturally occurring fluorophores, like tryptophan residues [55–61] as well as artificially introduced fluorescent probes.[41,59,62,63] Other biomolecular systems like membranes[64–72] or DNA[73–76] were also studied by fluorescence anisotropy.

Figure 1.1 C illustrates how the mobility of a dye is then restricted by the presence of the protein. This restriction of the mobility will usually depend on the structure and electrostatics of the protein surface to which the dye is attached, as indicated by the two different cones. Unfortunately, analysis of the obtained anisotropy is not straightforward. Therefore, models have been proposed to facilitate the interpretation. The situation shown in Fig. 1.1 C usually is described by the *wobbling-in-a-cone* model, which assumes that the dye diffuses freely inside the depicted cone. Unfortunately, it is often impossible to show if this simple model is actually justified. Furthermore the dynamics of the dye, which is probed in the experiment, is affected by the motion of the protein fragment to which it is attached. Therefore, the anisotropy yields information on the local protein structure and conformational changes as well as on the local protein *flexibility*. The simple *wobbling-in-a-cone* model, thus, has to be extended to also include the flexibility of the protein, which is the subject of Chap. 2.

MD-simulation as a model-free approach

Many results obtained by fluorescence anisotropy experiments depend on the particular choice of the model used for their interpretation. A model-free approach, therefore, would enable one to drop these assumptions and thus to provide more accurate interpretations of the experiments at the molecular level. The subject of Chaps. 6 and 7 is to check to what extent molecular dynamics (MD) simulations of the full anisotropy experiment can serve that purpose. To this aim, we have carried out MD simulations of the complete experimental system including a protein, a protein-attached dye and an explicit solvent environment, which allow to extract the individual contributions to the depolarization and to analyze the dye-protein interactions in detail. From this simulation, the fluorescence anisotropy of the dye can be calculated and compared to the experiment, which also allows to directly validate the simulations.

In a similar spirit, but at faster time scales, MD simulations of tyrosine,[77] tryptophan[61,78–84] and phenylalanine[85] containing proteins have been used to predict the fluorescence anisotropy decay function. Simulations of free fluorophores in a solvent could reproduce temperature and solvent dependence of the experimental fluorescence anisotropy.[86,87] The rotational diffusion of tryptophan in water has been simulated and its dependency on different water models discussed.[88,89] The anisotropy of a fluorescein bound to the Fab fragment has been calculated from relatively short (174 ps) MD-simulations.[90]

Here, and in contrast to the previous studies, the focus is on the interaction and dynamic coupling between the protein fragment and the attached dye. In addition, determination of dye conformation often is a key to the interpretation of fluorescence spectroscopy experiments. We will also study to which extent dye conformations can be obtained by MD simulations, which would therefore provide valuable information complementary to the experiment.

Bacteriorhodopsin as a test system

The system studied here is the well-known membrane protein Bacteriorhodopsin (bR), which is considered a prototype for a large class of membrane proteins, the G-protein coupled receptors (GPCR), since they share the heptahelical transmembrane motif. The GPCRs are used by many cell signaling pathways to convert external and internal stimuli into intracellular responses. The dynamics and flexibility of surface exposed protein segments of these GPCRs have been shown to play a central role in molecular recognition and activation of the receptor.[91] Investigation of the local protein surface dynamics and flexibility can thus provide much insight into

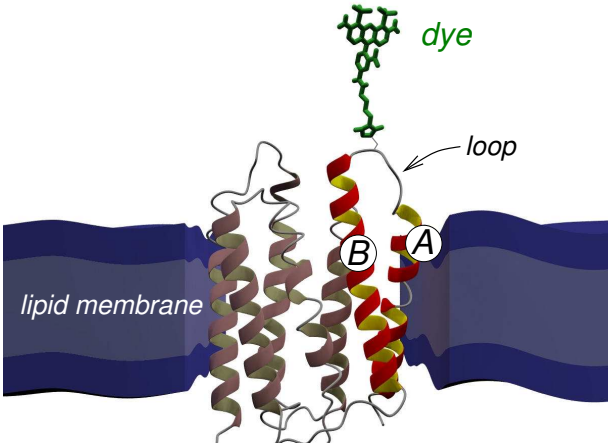

Figure 1.2: The membrane protein bacteriorhodopsin. The system studied in this work comprises the two highlighted helices A and B of bacteriorhodopsin and a fluorescent label, which is covalently attached to a cysteine in the loop connecting both helices.

the mechanism of the binding of ligands to the protein surface. The importance of the surface loops for recognition and binding of ligands has been suggested.[92] Therefore, fluorescence anisotropy experiments have been carried out by U. Alexiev (FU Berlin) and co-workers, to study the dynamics of the surface loops bR.[13] This study revealed temperature and pH dependent conformational changes of the loop. Unfortunately, only the *relative* mobility of the surface loops could be determined.

Our aim is to gain insight into these fluorescence anisotropy experiments by providing an interpretation of the experiment in atomic detail. We will particularly address the question, which processes influence the reorientational dynamics of the dye and thus contribute to the observed anisotropy decay, and how to extract information on the protein conformational dynamics from the anisotropy decay curves. Finally, we ask if and to what extent, vice versa, the attached dye affects the unperturbed loop dynamics. This effect is commonly — and necessarily — assumed to be negligible. The present study offers the chance to test this assumption.

Analysis of the correlation of the dye-protein dynamics is crucial

For studying the interactions between the dye and the protein, which lead to the characteristic anisotropy of the dye, careful analysis of the correlation of the dye-protein is crucial. The main focus is here on the question which parts or modes of the motion of the protein affect the dye motion, and are

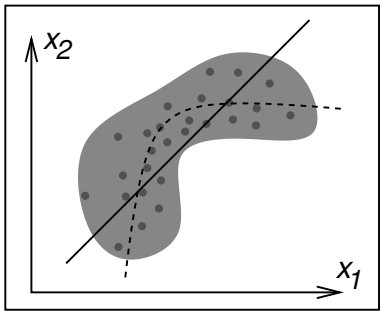

Figure 1.3: Sketch of an ensemble of protein structures (gray circles) in $3N$-dimensional configurational space, projected onto two dimensions, x_1 and x_2. Principal component analyses yield a linear principal coordinate (solid line). But often a curvilinear coordinate (dashed curve) should be more suitable to describe the shape of the ensemble.

therefore probed by the dye motion and, thus, by the experiment. The question is, of course, how this mode should be extracted from the ensemble of dye/protein structures obtained by the MD simulation.

To answer this question and following a statistical mechanics approach, a dye/protein structure containing N atoms is described by a single point in the $3N$-dimensional configurational space. An ensemble of these structures is then represented by a 'cloud' of points in the, usually high-dimensional, configurational space. Figure 1.3 illustrates a two-dimensional projection of such a 'cloud'. Large — and usually collective — conformational motions or structural rearrangements of the dye/protein system manifest themselves as large extensions of the ensemble 'cloud' in the configurational space. Those 'relevant' or 'essential' motions are to be distinguished from high-frequency but small-amplitude thermal fluctuations, the details of which are usually functionally irrelevant.

The widely used method to obtain a principal coordinate describing the largest conformational motion (solid line in Fig. 1.3), is the principal component analysis[93,94] (PCA). The obtained principal component (or coordinate) maximizes the variance of the ensemble projected onto this coordinate, i. e., it yields the direction of the largest extension of the 'cloud' in the configurational space. This principal coordinate has the convenient property, that it correlates best with the given dataset (the 'cloud'), i. e., the motion along this principal coordinate is optimally correlated with the largest motion of the dye/protein system. Thus, to find the best correlated mode of motion, the largest extension of the structural ensemble has to be determined.

However, the PCA is not optimal for this purpose in this case for two reasons. First, the PCA yields a linear coordinate, whereas typical motions, especially those of the dye, are clearly nonlinear, which requires the determination of principal *curvilinear* coordinates (dashed line in Fig. 1.3). Second, the

principal coordinate obtained by the PCA maximizes the correlation with the ensemble in the complete configurational space, whereas we are here interested in the correlation between *two* configurational subspaces (dye and protein). Therefore a generalization of the PCA is required.

In Chap. 5, we develop a method, which particularly accounts for the nonlinearity of the conformational motions and which enables one to calculate correlations between configurational subspaces.

This work is organized as follows: After an introduction to the theory of FRET and fluorescence anisotropy in Chap. 2, the maximum likelihood method to obtain distance trajectories from single-molecule FRET experiments is presented (Chap. 3). Then, the method of MD-simulation is introduced in Chap. 4. In Chap. 5 the method to calculate principal curvilinear coordinates and correlations is presented. Chap. 6 describes simulations of free dyes in different solvents to test the dye and solvent force fields, which are used in Chap. 7, where the simulation of a protein-attached dye is presented. Finally, Chap. 8 summarizes the results of this work and gives an outlook on future challenges.

"It is theory that decides what can be observed."
– Albert Einstein

2

Theory of Fluorescence Anisotropy and Fluorescence Resonance Energy Transfer

In this chapter, we introduce the main concepts of fluorescence anisotropy and fluorescence resonance energy transfer, on which this work is based. First, the basic principle of fluorescence anisotropy experiments is presented, then a simple model to describe the anisotropy of a protein-attached dye is derived. It is then explained, how the fluorescence anisotropy can be obtained from molecular dynamics simulations. Finally, the theoretical basis of distance determination between two dyes by fluorescence resonance energy transfer is presented.

2.1 Fluorescence Anisotropy

Fluorescence anisotropy, which was first described by Perrin,[95] is based on the observation that when a small fluorescent molecule is excited with plane-polarized light, the emitted light is largely depolarized because molecules tumble rapidly in solution during their fluorescence lifetime.

In a time-resolved fluorescence anisotropy experiment an ensemble of dyes is excited using a polarized laser pulse, as shown in Fig. 2.1. Those dyes are excited, that have their transition dipole moment oriented roughly parallel to the exciting laser pulse, since the probability of excitation is proportional to $\cos^2 \omega$, where ω is the angle between the transition dipole moment of the

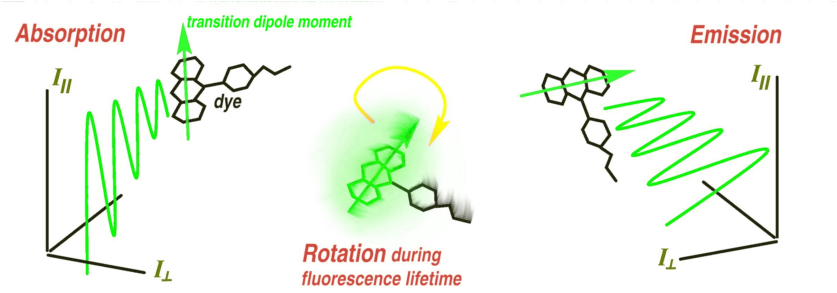

Figure 2.1: Fluorescence anisotropy experiment. The dye is excited by a polarized laser pulse. The dye then undergoes rotational diffusion during its fluorescence lifetime. Finally, the dye emits a photon in a possibly changed orientation. Two polarization detectors yield the parallel ($I_\parallel$) and perpendicular ($I_\perp$) part of the emitted light, with respect to the incident polarization ($I_\parallel$).

dye and the polarization of the incident light. This process is called *photoselection*. During their fluorescence lifetime of typically few nanoseconds, the dyes might undergo rotational diffusion. The dyes then emit the light in a possibly changed orientation. The resulting rotation of the polarization plane is detected by two polarization detectors, which yield two intensity signals; $I_\parallel$ for the parallel and $I_\perp$ for the perpendicular part of the emitted light (cf. Fig. 2.1).

The fluorescence anisotropy $r(t)$ at time t after excitation of the fluorophore is defined as

$$r(t) = \frac{I_\parallel(t) - I_\perp(t)}{I_\parallel(t) + 2I_\perp(t)} \quad , \tag{2.1}$$

where $I_\parallel(t)$ and $I_\perp(t)$ are the parallel and perpendicular fluorescence intensities, respectively, with respect to the field vector of the exciting light pulse. Assuming an ensemble of fluorophores with random isotropic initial orientations, $r(t)$ is given by[96,97]

$$r(t) = \frac{2}{5} \langle P_2 \left[\mu_a(s) \cdot \mu_e(s+t)\right] \rangle_s \tag{2.2}$$

where $\mu_a(t)$ and $\mu_e(t)$ are normalized vectors oriented along the absorption and emission dipole moments, respectively. Here, assuming a sufficiently ergodic MD-trajectory, the ensemble average $\langle\,\rangle_s$ will be approximated by a time-average. $P_2(x) = \frac{1}{2}(3x^2 - 1)$ is the second-order Legendre polynomial. In the simplest case of isotropic rotational diffusion of a fluorophore, the anisotropy shows a mono-exponential decay to zero with a decay time ϕ,

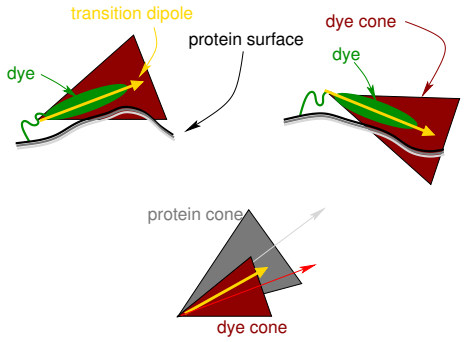

Figure 2.2: *Cone-in-a-cone* model. The two upper figures indicate a motion, which is due to the protein flexibility, superimposed to the standard *wobbling-in-a-cone* model. The lower figure visualizes, how the protein flexibility is described in the *cone-in-a-cone* model by the protein cone (gray), within which the dye cone (red) freely diffuses.

the rotational correlation time, which directly depends on the rotational diffusion coefficient. If the motional freedom of a dye is restricted, e. g., when attached to a protein, the anisotropy will typically not decay to zero. A common model to describe such restricted rotational diffusion is the *wobbling-in-a-cone model*,[97] where the transition dipole is assumed to diffuse freely inside a cone, as shown in the upper part of Fig. 2.2. In this case, the anisotropy $r(t)$ can be approximated by

$$r(t) = r_0 \left[(1 - A_\infty) \mathrm{e}^{-t/\phi} + A_\infty \right] , \tag{2.3}$$

where $r_0 = 0.4\, P_2(\cos\lambda)$, with λ being the angle between the absorption and emission dipole moment. A_∞ is a parameter describing the degree of motional restriction and is therefore related to the (half-)cone angle θ_{max} by

$$A_\infty = \frac{r_\infty}{r_0} = \left[\frac{1}{2} \cos\theta_{max} (1 + \cos\theta_{max}) \right]^2 . \tag{2.4}$$

Note that a large value of A_∞ corresponds to a small cone angle, and for isotropic diffusion, A_∞ vanishes. Assuming an isotropic overall tumbling motion of the whole dye-protein complex with a rotational correlation time ϕ_G, the anisotropy of the protein-attached dye is then given by

$$r(t) = r_0 \left[(1 - A_\infty) \mathrm{e}^{-t/\phi} + A_\infty \right] \mathrm{e}^{-t/\phi_G} . \tag{2.5}$$

Furthermore, the local flexibility of the protein is an additional source of reorientation of the dye. A simplified description, which also includes the

protein flexibility is shown in Fig. 2.2. The dye wobbles in the dye-cone, while the protein surface is changing its orientation and thereby reorienting the dye-cone itself. If the local dye motion and the protein dynamics can be assumed to be uncoupled, this effect can be accounted for by considering a second decay factor,[81]

$$r(t) = r_0 \left[(1 - A_1)\mathrm{e}^{-t/\phi_1} + A_1\right] \left[(1 - A_2)\mathrm{e}^{-t/\phi_2} + A_2\right] \mathrm{e}^{-t/\phi_G} , \qquad (2.6)$$

which can be interpreted as a wobbling in a cone, which itself wobbles in a further cone (see lower part of Fig. 2.2), hence this model is referred to in this text as the *cone-in-a-cone model*.

For the analysis of fluorescence anisotropy experiments, often a sum of exponentials is used,

$$r(t) = \sum_i B_i \mathrm{e}^{-t/\varphi_i} , \qquad (2.7)$$

which however cannot be intepreted directly as independent decay components, that contribute to the anisotropy.

2.1.1 Fluorescence anisotropy from the simulation

The change in the orientation of the transition dipole moment $\mu(t)$ of the dye leads to the decay of the anisotropy $r(t)$, as described by Eq. 2.2. The absorption dipole moment $\mu_a(t)$ in the coordinate frame of the dye is obtained from the CIS calculation of the dye, as described in Sec. 4.3, and is oriented along the long-axis of the three-ring system of the chromophore shown as the red arrow in Fig. 4.2. A difference between the absorption and emission dipole moment, described by the angle λ, leads to an overall reduction of the anisotropy $r(t)$ by a factor $P_2(cos\lambda)$, where P_2 is the second-order Legendre polynom. The initial anisotropy at t=0 is therefore $r_0 = 0.4\ P_2(cos\lambda)$. The measured initial anisotropy is 0.37, which corresponds to an angle $\lambda = 10^{\mathrm{O}}$. Since we are only interested in the shape of $r(t)$ and not in the initial anisotropy r_0, we set the transition dipole moment $\mu(t) = \mu_a(t)$ and plot $r(t)/r_0$, which is thus normalized to 1. The orientation of this transition dipole moment vector is calculated for each snapshot of the MD trajectory from the instantaneous orientation of the dye yielding a trajectory of the transition dipole vector $\mu(t)$. Then the anisotropy $r(t)$ is calculated using Eq. 2.2 from a time average of $\mu(t)$.

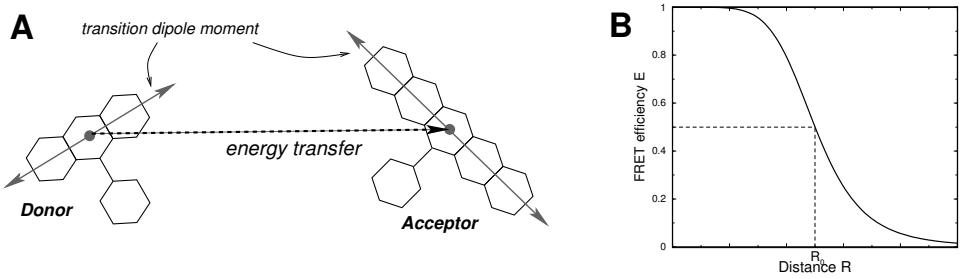

Figure 2.3: Fluorescence resonance energy transfer. (A) Excitation energy is transferred from the donor to the acceptor dye. (B) The transfer efficiency E depends on the distance R between the two dyes. The efficiency is most sensitive to distance changes near the Förster radius R_0.

2.2 Fluorescence Resonance Energy Transfer

Fluorescence Resonance Energy Transfer (FRET) was first described by Förster in 1948.[16] It is a powerful tool to measure distances between two dyes, a donor and an acceptor, in the range of 10–75 Å (see Fig. 2.3 A).[20] The excitation energy is transferred from the donor to the acceptor via an induced dipole–induced dipole interaction. The transfer efficiency E (see Fig. 2.3 B) is given by

$$E = \frac{1}{1 + (R/R_0)^6} \quad , \tag{2.8}$$

where R is the distance between the dyes and R_0 is the Förster radius, which denotes the distance at which 50% of the energy is transferred to the acceptor. Figure 2.3 B indicates, that the highest resolution of distance determination is achieved, if the distance R is close to R_0, since then the transfer efficiency E is most sensitive to distance changes. R_0 depends on the particular properties of the dyes as well as on the relative orientation:

$$R_0^6 = (8.79 \times 10^{-25})\kappa^2 n^{-4} \phi_d J_{da} \quad , \tag{2.9}$$

where κ^2 is the orientation factor (discussed in more detail below), n is the refractive index of the medium between the two dyes, which is generally assumed to be 1.4 for proteins,[98] ϕ_d is the quantum yield of the donor, which is defined as the ratio of the number of photons emitted to the number absorbed, and J_{da} represents the overlap integral of the donor emission spectrum with the acceptor absorption spectrum, illustrated in Fig. 2.4 A

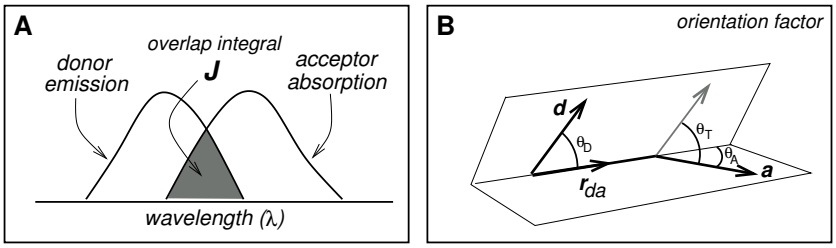

Figure 2.4: (A) Overlap integral (gray area) of the donor emission spectrum with the acceptor absorption spectrum. (B) Visualization of the angles used to define the orientation factor κ^2. $\mathbf{d}$ and $\mathbf{a}$ are the transition dipole moments of the donor and acceptor, respectively, and $\mathbf{r}_{da}$ is the normalized vector connecting the two dyes.

and defined by

$$J_{da} = \int_0^\infty f_D(\lambda)\epsilon_A(\lambda)\lambda^4 d\lambda \quad , \tag{2.10}$$

where $\epsilon_A(\lambda)$ is the molar extinction coefficient of the acceptor and $f_D(\lambda)$ is the fluorescence spectrum of the donor normalized on the wavelength scale

$$f_D(\lambda) = \frac{F_{D\lambda}}{\int_0^\infty F_{D\lambda}(\lambda)d\lambda} \ , \tag{2.11}$$

where $F_{D\lambda}$ is the donor fluorescence per unit wavelength interval.[98] All parameters appearing in Eq. 2.9 can be determined experimentally except for the orientation factor κ^2, which is illustrated in Fig. 2.4 B and defined by

$$\kappa^2 = [\mathbf{d}\cdot\mathbf{a} - 3\,(\mathbf{d}\cdot\mathbf{r}_{da})(\mathbf{a}\cdot\mathbf{r}_{da})]^2 \quad , \tag{2.12}$$

where $\mathbf{d}$ and $\mathbf{a}$ are the transition dipole moments of the donor and acceptor, respectively, and $\mathbf{r}_{da}$ is a normalized vector connecting the two dyes. An equivalent alternative definition is

$$\kappa^2 = (\cos\theta_T - 3\,\cos\theta_D\cos\theta_A)^2 \quad , \tag{2.13}$$

where the angles θ_T, θ_D, and θ_A are defined in Fig. 2.4 B.

The transfer efficiency therefore depends on both the distance and the relative orientation. Thus, in general, the distance cannot be directly obtained by measuring the transfer efficiency. To overcome this problem, usually, one attaches the fluorescent dyes to the biomolecules via long flexible linkers.

The highly flexible dyes then ensure an averaging of dye orientations, which in this case leads to κ^2=2/3. The transfer efficiency then depends only on the donor–acceptor distance.

The FRET efficiency can be obtained by measuring either the fluorescence intensities or the fluorescence lifetimes of the donor with and without the acceptor, which is expressed by

$$E = 1 - \frac{I_{da}}{I_d} = 1 - \frac{\tau_{da}}{\tau_d} \quad , \tag{2.14}$$

where I_{da} and I_d are the measured intensities in the presence and absence of the acceptor, respectively, and τ_{da} and τ_d are the fluorescence lifetimes in the presence and absence of the acceptor.

In the case of single-molecule measurements, the determination of the FRET efficiency becomes more complicated due to a limited number of detected photons, which is the topic of the next chapter.

"Wo viel Licht ist, ist starker Schatten."
– Johann Wolfgang von Goethe

3

Maximum Likelihood Trajectories from FRET experiments

Recently, time-resolved FRET experiments have matured to a level that allows one to record arrival times of *individual* photons from *single* molecules.[11,20,23,35,50,99–101] From the arrival times, fluorescence intensity *variations*, $I_{\mathrm{D}}(t)$ and $I_{\mathrm{A}}(t)$, are obtained,[23,46,47] which, using Eq. (2.8), allow one to track distance changes $R(t)$ between the two dyes, and hence to monitor conformational motions of the studied biomolecule.[46,47]

In the conventional analysis, the required FRET intensities are computed from photon counts in time windows[23,50] (cf. also Ref. 51). For a typical window size of 1 ms, however, the small number of only 10 . . . 50 photons per window[23] implies considerable statistical uncertainty ('shot noise'[52]) and thus limits the time resolution for $R(t)$. Furthermore, the choice of the window size is somewhat arbitrary and only guided by the requirement to trade off shot noise and time resolution. Finally, the traditional method saliently assumes a uniform *a priori* probability for the FRET *intensities* (rather than for the distances). Therefore, and contrary to what one might intuitively assume at first sight, the traditional method cannot be considered a model-free approach. Rather, because the distance R depends non-linearly on the intensities, Eq. (2.8), the assumed uniform intensity distribution transforms into a non-uniform distance distribution,

$$p(R) = \frac{\left(\frac{R}{R_0}\right)^5}{\left[1+\left(\frac{R}{R_0}\right)^6\right]^2}\,. \tag{3.1}$$

This distribution is centered at the Förster radius and has a half width of

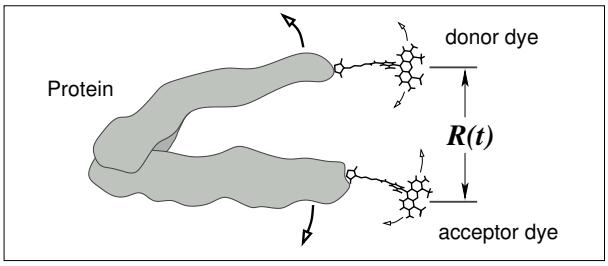

Figure 3.1: Typical single molecule FRET experiment. A donor and an acceptor dye molecule are attached to a protein that exhibits conformational dynamics. By probing the inter-dye distance trajectory $R(t)$, measurement of the FRET efficiency provides time-resolved information on the dynamics of the studied protein (arrows).

about $\frac{1}{3}R_0$, implying preferred distances near R_0; it describes the unjustified bias introduced by the conventional analysis.

In many cases where only limited or noisy data are available, the maximum-likelihood approach has been successfully applied.[36–42] In this chapter, we develop a maximum-likelihood theory to reconstruct $R(t)$ from the photons recorded in single molecule FRET measurements. In particular, we aim at calculating the time-dependent probability distribution $P(R,t|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\})$ for the distance R during a measurement of length ΔT, given that n_{D} photons from the donor dye have been recorded at times t_i^{D}, $i = 1 \ldots n_{\mathrm{D}}$, and n_{A} acceptor photons at times t_i^{A}, $i = 1 \ldots n_{\mathrm{A}}$. Finally, we will extract an effective diffusion coefficient for the biomolecular motion from the FRET data. As an example, the method will be applied to a recorded photon burst from a FRET measurement of donor and acceptor dyes attached to the flexible domains of the neuronal fusion protein syntaxin-1a.[23]

3.1 Theory

To that aim, in a first step we consider a statistical ensemble of distance trajectories, $\{R(t)\}$, and compute for each *full trajectory* the conditional probability $P[R(t)|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}]$ that $R(t)$ is realized for the given photon registration times. Assuming Bayesian statistics, this probability is given by the *a priori* probability $P[R(t)]$ for each trajectory and the conditional probability that the n_A+n_D photons are observed at the measured time instances for given trajectory,

$$P[R(t)|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}] \propto P[R(t)]P[\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}|R(t)]\,. \tag{3.2}$$

To evaluate these two distributions, the time interval ΔT is discretized into N bins $[\tau_{j-1}, \tau_j]$, $j = 1, \ldots, N$, and subsequently $N \to \infty$ is considered. The time discretization $\tau := \tau_j - \tau_{j-1} = \Delta T/N$ is always chosen fine enough such that not more than one photon per interval $[\tau_{j-1}, \tau_j]$ is recorded.

For a discretized trajectory $R_1, \ldots, R_N$, where R_j is the distance at time $\frac{1}{2}(\tau_{j-1} + \tau_j)$, the conditional probability to observe the recorded photon pattern $E_1, \ldots, E_N$ is

$$P[E_1, \ldots, E_N | R_1, \ldots, R_N] = \tau^{n_A + n_D} \prod_{j=1}^{N} w_j \,, \tag{3.3}$$

where the probabilities w_j are chosen according to which of the three possible events E_j [donor-photon is recorded ('D'), acceptor-photon is recorded ('A'), or no photon is recorded ('0')] occurs during $[\tau_{j-1}, \tau_j]$,

$$w_j = \begin{cases} I_D(R_j)[1 - \tau I_A(R_j)] & \text{for 'D'}, \\ I_A(R_j)[1 - \tau I_D(R_j)] & \text{for 'A'}, \\ [1 - \tau I_D(R_j)][1 - \tau I_A(R_j)] & \text{for '0'}. \end{cases} \tag{3.4}$$

Here, $I_{\mathrm{A}}(R_j)$ and $I_{\mathrm{D}}(R_j)$ are specified from Eq. (2.8), and the required (average) total intensity $I_0 = I_{\mathrm{A}}(t) + I_{\mathrm{D}}(t) = (n_{\mathrm{A}} + n_{\mathrm{D}})/\Delta T$ is estimated from the recorded number of photons. Note that for the $n_D + n_A$ events 'D' and 'A', the w_j denote probability *densities*, which have to be scaled by τ to obtain the desired probabilities, hence the prefactor in Eq. (3.3).

For the *a priori* probability $P[R(t)] \propto \lim_{N\to\infty} P[R_1, \ldots, R_N]$, we assume that $R(t)$ results from a one-dimensional diffusion process with effective diffusion coefficient D. This is realistic, e.g., for the overdamped millisecond opening and closure domain motions of the solvated macromolecule at hand.[23] The discretized version is a random walk with transition probabilities

$$g_{j+1|j} \propto \frac{1}{\sqrt{4\pi D\tau}} \exp\left[-\frac{(R_{j+1} - R_j)^2}{4D\tau}\right] . \tag{3.5}$$

Note that this implies that all possible distances are assigned equal *a priori* probabilities, which is reasonable if the energy landscape that governs the distance distribution is unknown. If there is additional information on the energy landscape, this can be incorporated into $g_{j+1|j}$ in a Smoluchowsky-type generalization. Note also that two or three dimensional diffusion of the dyes can be described in a similar manner by an appropriate effective energy landscape that accounts for the projection of the higher-dimensional diffusion onto the one-dimensional distance coordinate $R(t)$.

Thus, $P[R_1,\ldots,R_N]=\prod_{j=2}^{N} g_{j|j-1}$, and Eq. (3.2) reads

$$P[R_1,\ldots,R_N|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}] \propto w_1 \prod_{j=2}^{N} g_{j|j-1} w_j \,. \tag{3.6}$$

In a second step the probability distribution for the *distance* R_k at times $(\tau_{k-1}+\tau_k)/2$ is calculated by integration over all other distances,

$$P(R_k|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}) \propto \int\cdots\int dR_1\ldots dR_{k-1}dR_{k+1}\ldots dR_N P[R_1,\ldots,R_N|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}]\,. \tag{3.7}$$

Using Eq. (3.6) and rearranging integrals, one obtains

$$P(R_k|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}) \propto F_k w_k B_k \tag{3.8}$$

with

$$\begin{aligned} F_k &= \int dR_{k-1} g_{k|k-1} w_{k-1} \int dR_{k-2}\cdots\int dR_1 g_{2|1} w_1, \\ B_k &= \int dR_{k+1} g_{k+1|k} w_{k+1} \int dR_{k+2}\cdots\int dR_N g_{N|N-1} w_N. \end{aligned} \tag{3.9}$$

The above two equations obey the recursion relations

$$\begin{aligned} F_k &= \int dR_{k-1} g_{k|k-1} w_{k-1} F_{k-1}\,, \\ B_k &= \int dR_{k+1} g_{k+1|k} w_{k+1} B_{k+1}\,, \end{aligned} \tag{3.10}$$

which, in the continuum limit (i.e., $\tau\to 0$, $\tau_j\to t$, and $r_k\to r$), transform into forward and backward Schrödinger-type equations that resemble generalized diffusion equations for $F_k\to F(r,t)$ and $B_k\to B(r,t)$,

$$\left.\begin{aligned} \partial_t F(R,t) &= \lim_{\tau\to 0}\left\{\partial_R^2\left[(1+\tau W_\tau(R,t))F(R,t)\right]\right. \\ &\quad \left.+\left[W_\tau(R,t)+\tau\partial_\tau W_\tau(R,t)\right]F(R,t)\right\}, \\ \partial_t B(R,t) &= -\lim_{\tau\to 0}\left\{\partial_R^2\left[(1+\tau W_\tau(R,t))B(R,t)\right]\right. \\ &\quad \left.+\left[W_\tau(R,t)+\tau\partial_\tau W_\tau(R,t)\right]B(R,t)\right\}, \end{aligned}\right\} \tag{3.11}$$

where, to ensure convergence, w_k has been written in the form $w_k = 1+\tau W_\tau(R,t)$. For the derivation of Eqs. (3.11), the recursion relations Eqs. (3.10) have been expanded in τ up to first order, using $\partial_\tau g_{k|k-1} = D\partial^2_{R_{k-1}} g_{k|k-1} = D\partial^2_{R_k} g_{k|k-1}$, and partial integration in R, noting that $F(R,t)$ and $B(R,t)$ as well as their derivatives with respect to R vanish for $R\to\pm\infty$.

Solving Eqs. (3.11) yields, after normalization, the desired probability distribution to find the distance R at time t,

$$P(R,t|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\}) \propto F(R,t)[1+\tau W_\tau(R,t)]B(R,t)\,. \tag{3.12}$$

By combining the three definitions for w_j, Eq. (3.4), into one expression using a Gaussian limit representation for the δ-function, $\delta(t-t') = \lim_{\tau\to 0} h_\tau(t-t')$, with

$$h_\tau(t-t') = \frac{1}{\sqrt{2\pi}\tau}\exp\left[-\frac{(t-t')^2}{2\tau^2}\right], \tag{3.13}$$

and neglecting higher orders of τ, one obtains

$$W_\tau(R,t) = [I_D(R)-1]\sum_{j=1}^{n_D} h_\tau(t-t_j^D) + [I_A(R)-1]\sum_{j=1}^{n_A} h_\tau(t-t_j^A) - I_0\,. \tag{3.14}$$

With this expression, Eqs. (3.11) reads

$$\begin{aligned}\partial_t F(R,t) = \lim_{\tau\to 0}\Bigg\{\int dR' g(R-R',\tau)\partial_{R'}^2\Bigg[F(R',t)\Bigg(1+\tau[I_D(R')-1]\sum_{j=1}^{n_D} h_\tau(t-t_j^D)+\\ \tau[I_A(R')-1]\sum_{j=1}^{n_A} h_\tau(t-t_j^A)\Bigg)\Bigg]+\\ \int dR' g(R-R',\tau)F(R',t)\Bigg[\frac{I_D(R')-1}{\tau^2}\sum_{j=1}^{n_D}(t-t_j^D)^2 h_\tau(t-t_j^D)+\\ \frac{I_A(R')-1}{\tau^2}\sum_{j=1}^{n_A}(t-t_j^A)^2 h_\tau(t-t_j^A) - I_0\Bigg]\Bigg\}\,. \end{aligned}\tag{3.15}$$

A similar expression is obtained for $B(r,t)$. For times t, for which no photon arrives, Eq. (3.15) simplifies to

$$\begin{aligned}\partial_t F(R,t) &= D\partial_R^2 F(r,t) - I_0 F(R,t)\,,\\ \partial_t B(R,t) &= -D\partial_R^2 B(R,t) + I_0 B(R,t)\,,\end{aligned} \tag{3.16}$$

with solutions that propagate in time according to

$$\begin{aligned}F(R,t) &= e^{-I_0(t-t')}\int dR' F(R',t')\exp\left[-\frac{(R-R')^2}{4D(t-t')}\right]\\ B(R,t) &= e^{I_0(t'-t)}\int dR' B(R',t')\exp\left[-\frac{(R-R')^2}{4D(t'-t)}\right]\end{aligned} \tag{3.17}$$

for $t > t'$ and $t < t'$, respectively. To also include the photon arrival times t_j, note that

$$\lim_{\tau\to 0}(t-t_j)^2 h_\tau(t-t_j)/\tau^2 = \lim_{\tau\to 0} h_\tau(t-t_j) + \lim_{\tau\to 0}\tau^2\partial_t^2 h_\tau(t-t_j) = \delta(t-t_j)\,, \tag{3.18}$$

where the second term is $\propto \partial_t^2 \delta(t-t_j)$ and is dropped, because $\int_{-\epsilon}^{\epsilon} \delta''(x)dx = 0$. This gives rise to additive singularities in Eqs. (3.16) of the form $F(R,t)[(I_D(R)-1)]\delta(t-t_j)$, due to which $F(R,t)$ and $B(R,t)$ exhibit discontinuities at all t_j,

$$\left.\begin{aligned} \lim_{t\to(t_j^D)^+} F(R,t) &= I_D(R) \lim_{t\to(t_j^D)^-} F(R,t), \\ \lim_{t\to(t_j^A)^+} R(R,t) &= I_A(R) \lim_{t\to(t_j^A)^-} F(R,t), \\ \lim_{t\to(t_j^D)^-} B(R,t) &= I_D(R) \lim_{t\to(t_j^D)^+} B(R,t), \\ \lim_{t\to(t_j^A)^-} B(R,t) &= I_A(R) \lim_{t\to(t_j^A)^+} B(R,t). \end{aligned}\right\} \tag{3.19}$$

Eqs. (3.17) and (3.19) are the main result of this chapter. Starting with the boundary condition $F(R,0) = 1$, Eqs. (3.17) and (3.19), when alternatingly applied, propagate $F(R,t)$ in time from one photon arrival to the next. Similarly, starting from $B(R,\Delta T) = 1$, $B(R,t)$ is propagated in reverse time direction, which, by using Eq. (3.12), yields $P(R,t|\{t_i^D,t_i^A\})$ for all times t. Note that, from Eqs. (3.19), the discontinuities in $F(R,t)$ and $B(R,t)$ cancel in Eq. (3.12), such that $P(R,t|\{t_i^D,t_i^A\})$ is non-differentiable, but continuous also at $t = t_j$.

3.2 Results and Discussion

As an example, Fig. 3.2(b-d) shows the application of our theory to the 230 photon arrival times (wedges) from a 10 ms single molecule photon burst recorded in a FRET measurement, for which donor and acceptor dyes have been covalently linked to the flexible domains of the neuronal fusion protein syntaxin-1a,[23] as sketched in Fig. 3.1. Three different diffusion coefficients D have been chosen. Each of the three plots shows, gray-shaded, the time dependent distance distribution $P(R,t|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}}\})$, together with the average distance (bold) and 1σ intervals (dashed). As expected from Eq. (2.8), larger distances are obtained for higher donor and lower acceptor photon intensities. For comparison, Fig. 3.2(a) shows the traditional method, which directly uses Eq. (2.8) with intensities and error bars evaluated in successive time bins,[102] here of 0.5 ms width.

Apparently, the choice of D is critical. For small values, the distance can change only slowly. Therefore, it does not fully reflect the significant intensity fluctuations encoded in the recorded photon arrival times, and rather yields smooth trajectories with small amplitude. For very small values (below $0.01 \times 10^{-14}\,\mathrm{m^2/s}$), the distance distribution becomes time indepen-

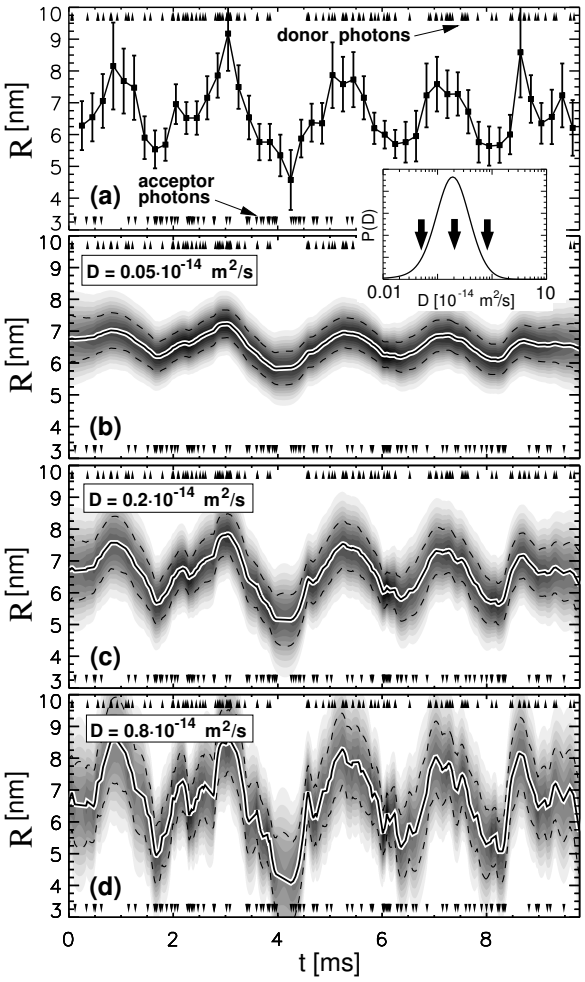

Figure 3.2: (a) Intensity-based calculation of donor/acceptor distances $R(t)$ from a set of 230 photon arrival times (wedges) with $R_0 = 6.5\,\mathrm{nm}$[23] using Eq. (2.8); intensities are obtained from 0.5 ms bins. (b-d) Time dependent distance probability distributions $P(R, t|\{t_i^D, t_i^A\})$ (gray-shaded) calculated from the same set for three different diffusion coefficients D. Also shown are average distance trajectories (bold) and 1σ intervals (dashed). The inset shows the (normalized) likelihood $P(D)$ as a function of D; three arrows denote the three chosen values for D.

dent and approaches the distance given by the average intensities (data not shown). Increasing D entails fluctuations of correspondingly increased frequencies. These fluctuations arise from both intensity fluctuations due to actual distance variations and (undesirable) probability fluctuations due to the broadening of $F(R,t)$ and $B(R,t)$ between subsequent photons. As can be seen from Eqs. (3.17), the latter become relevant for $4D > I_0\sigma^2$, where σ is the width of $P(R,t|\{t_i^D,t_i^A\})$. The lower panel in Fig. 3.2 shows an example for which, due to the large D chosen, the data are apparently over-fitted. In between these two limiting cases, an optimal value for D is expected to provide the best description of the data [Fig. 3.2(c)].

That optimal value was determined by calculating the agreement between the obtained time-dependent distance distribution and the measured photon arrival times as a function of the chosen D. Such type of cross-validation underlies, e.g., the *free* R value used to assess the accuracy of macromolecular X-ray structures.[103] In a similar spirit, one photon k was excluded from the FRET data, and a new distance distribution

$$P_k(R_k) \equiv P_k(R_k,t_k|\{t_i^{\mathrm{D}},t_i^{\mathrm{A}},i\neq k\}) \tag{3.20}$$

was obtained for the arrival time t_k of the excluded photon. Using this distribution, the likelihood $P_k(D)$ for the actually observed photon k was determined for varying D,

$$P_k(D) \propto \int_0^\infty dR_k P_k(R_k) I_{\mathrm{D/A}}(R_k)\,, \tag{3.21}$$

with $I_{\mathrm{D/A}}$ chosen according to the type of the excluded photon. Assuming that for different photons k chosen to be omitted, the obtained likelihoods $P_k(D)$ are statistically independent, one obtains from the maximum of the (normalized) joint likelihoods $P(D) \propto \prod_k P_k(D)$ (inset of Fig. 3.2) a diffusion coefficient $D = 0.2\times10^{-14}\,\mathrm{m}^2/\mathrm{s}$ that describes the measured photon arrival times best. In the figure, no scale for $P(D)$ is given to avoid its erroneous interpretation as the (absolute) probability that D is the correct diffusion constant.

Clearly, the fewer photons are available, the less information on $R(t)$ can be obtained. As an extreme case, Fig. 3.3(a) shows the result of our analysis with only every fourth photon from the original data used. As expected, the distance distribution becomes broader, and only some of the features seen in Fig. 3.2 remain. Yet, despite the very small number of photons used (58), our analysis still reveals a statistically significant distance fluctuation at the 1σ level. This finding suggests that a correspondingly improved time resolution can be achieved by our method.

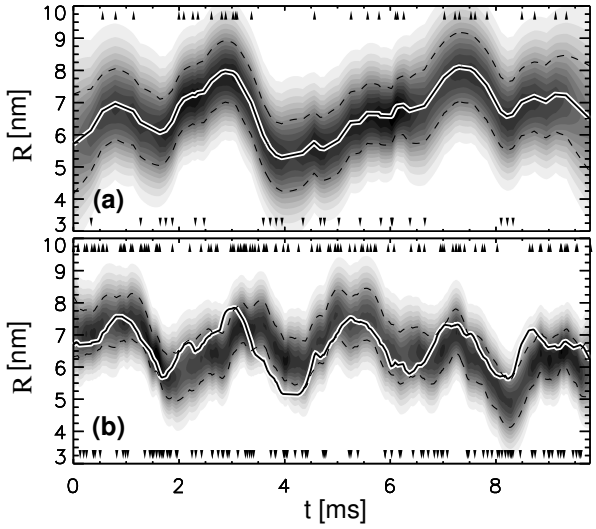

Figure 3.3: (a) Distance distribution for a reduced set of 58 photons (wedges) and $D = 0.2 \times 10^{-14}\,\mathrm{m}^2/\mathrm{s}$; notation as in Fig. 3.2. (b) Re-calculated distance distribution (gray-shaded) for a hypothetical set of 230 photons (wedges) that has been calculated from the original average trajectory in Fig. 3.2(c), also shown in bold here; $D = 0.2 \times 10^{-14}\,\mathrm{m}^2/\mathrm{s}$. The dashed lines denote the 1σ interval for the re-calculated distance distribution.

To check whether the width of the calculated distance distribution correctly describes the actual statistical uncertainty, we have finally used the average trajectory calculated from the original data [thick line in Fig. 3.2(c)] to create a new (hypothetical) set of 230 random photon arrival times obeying Eq. (2.8). Thus, for these data, the underlying trajectory is known. From that set, a new distance distribution was re-calculated and compared with the correct trajectory [Fig. 3.3(b)]. As can be seen, most of the correct trajectory (bold) stays within the 1σ-range of the re-calculated distance distribution, thus showing the reliability of our method.

We have developed a theory that enables reconstruction of nanometer distance trajectories from single molecule single photon FRET recordings. In contrast to the commonly used method of window averaging, the full single photon information is used, and rigorous error bounds are obtained. Furthermore, the method is expected to be robust with respect to variation of the excitation intensity I_0, e.g., due to diffusion of the particle through the laser focus. In addition, our approach allows to extract an effective diffusion constant from the FRET recordings and thus avoids the usual *ad hoc* choice of an averaging interval for the determination of intensities. Finally, the

likelihood approach avoids the severe bias of usual distance determination due to the salient assumption of uniform *a priori* probabilities for the FRET *intensities*, which implies, via Eq. (2.8), preferred distances near R_0. A software package that implements this theory ('FRETtrace') can be downloaded from the web-page of the author.

"Force is not a remedy."
– John Bright

4

Molecular Dynamics Simulation Method

The MD-method has been extensively described in the literature; a good review gives Ref.[104] In this chapter the basic principle is sketched, followed by a description of the used algorithms to improve the efficiency of the calculations, to treat the system boundaries, and couplings to a heat and a pressure bath. Finally, the parametrization of the Alexa488 dye, which will be used later in this work, is described.

4.1 Principle

The goal of molecular dynamics (MD) simulations is to describe the atomic motions of molecular systems containing about 10^3 to 10^6 interacting atoms. The exact treatment of this problem requires the solution of the time-dependent Schrödinger equation. However, this is even for small systems of more than ten atoms computationally too expensive. To be able to describe larger systems, like a protein in its solvent environment, basically three approximations are necessary.

The first approximation is based on the fact, that electrons move much faster than nuclei, due to their much smaller masses. Therefore, the electronic degrees of freedom can be separated from the degrees of freedom of the nuclei, which is called Born-Oppenheimer approximation.[105,106] The resulting time-independent Schrödinger equation for the electrons can then be solved for fixed nuclei positions. This yields an effective potential, which

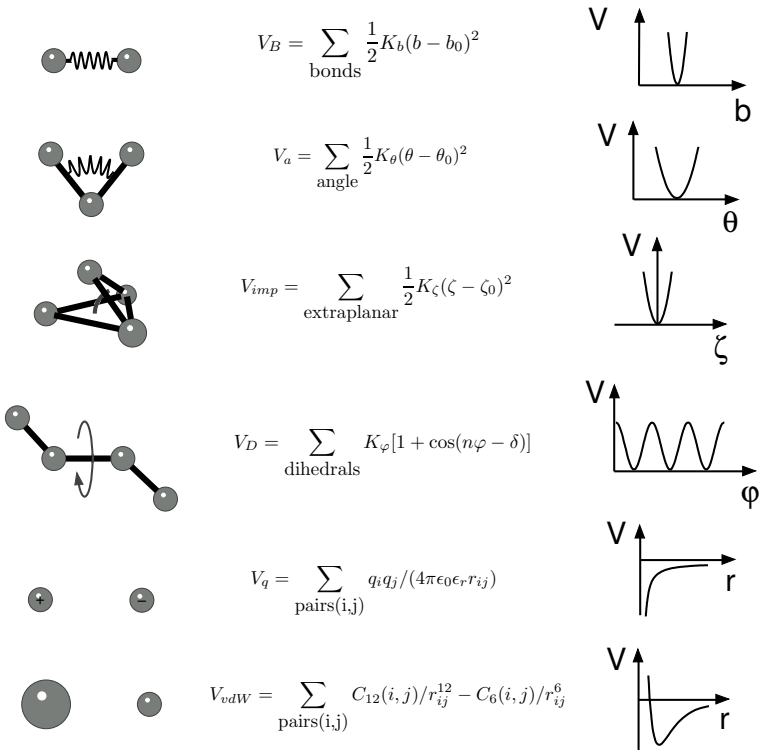

Figure 4.1: Energy terms, which constitute a molecular force field for use in molecular dynamics simulations: (A) Bond-stretching and (B) bond-angle vibrations, (C) out-of-plane motions, (D) dihedral angles, (E) van der Waals, and (F) Coulomb interactions. The interatomic interactions are illustrated on the left, the corresponding energy terms are shown in the middle, and plotted on the right.

depends only on the nuclei positions, and which describes the influence of the electron dynamics on the nuclei motion.

In the second step, this effective potential is approximated by a semi-empirical molecular force field, which comprises a large number of functionally simple energy terms. These energy terms are shown in Fig. 4.1 and include bond-stretching (A), bond-angle (B), out-of-plane (C), and dihedral-angle (D) potentials, which approximately describe the properties of covalent bonds, and which therefore are called *bonded* interactions. Furthermore, the long-range *non-bonded* interactions are also considered: The Lennard-Jones[107] potential (E) models the Pauli repulsion, which prevents atoms from penetrating each other, and induced dipole interactions, collectively termed *van der Waals interactions*. The electrostatic interaction between charged atoms is described by the Coulomb potential (F). The parameters used in

such force fields, like, e. g., equilibrium bond-lengths b_0, force constants K_b, partial charges, and van der Waals radii, are obtained from experiments as well as from quantum-chemical calculations in a self-consistent way.[108] The parameterization is done for small and simple molecules regarded as the building blocks of large molecules; in the case of proteins the parameters are optimized for individual amino acids and short peptides. Most experimental data used for force field development come from x-ray crystallography, IR- and NMR-spectroscopy.[104] In this way, many different force fields has been developed (CHARMM,[109] GROMOS,[110] AMBER,[111] MM3,[112] CFF,[113] SPASIBA,[114] etc.) for the description of different classes of molecules, like, e. g., proteins, DNA, or carbohydrates. In the end, the usage of such empirical force fields is justified by its ability to reproduce and to predict experimental results.[115]

In the third approximation the dynamics of the nuclei is described classically instead of solving the time-dependent Schrödinger-equation. All atoms are thus described as point masses, which move in the given force field $\mathbf{F}(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ according to the *Newtonian equations of motion*

$$m_i \frac{d^2\mathbf{r}_i(t)}{dt^2} = \mathbf{F}_i(\mathbf{r}_1, \ldots, \mathbf{r}_N) = -\nabla_i V(\mathbf{r}_1, \ldots, \mathbf{r}_N)\,, \tag{4.1}$$

where m_i and $\mathbf{r}_i$ is the mass and the position of the i-th nucleus ($i = 1, \ldots, N$), and $\mathbf{F}_i$ the force on atom i and N the number of atoms. $V(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ denotes the used force field.

The classical description is appropriate, if the thermal energy is distinctly larger than the energy gaps between neighbored quantum states of the system. For the harmonic oscillator, as an example, this requirement means that if $k_B\,T \gg \Delta E = hf$ (k_B: Boltzmann constant; T: temperature; ΔE: energy difference of neighbored states; h: Planck constant; f: vibrational frequency) the classical description is sufficient. Motions at a characteristic timescale of a few picoseconds and longer, thus, allow to be described classically. However, the fast motions in molecular systems, like bond-stretching vibrations with a typical frequency of $f \approx 30 - 60\text{ps}^{-1}$, cannot be accurately treated by classical mechanics. Furthermore, processes occuring at low temperatures also cannot be adequately described, since the quantum mechanical character of the low-energy nuclei motions becomes more pronounced.

The conformational motions of interest in this work take place on picosecond or longer timescales and therefore can be described well within classical mechanics.

4.2 Computing Trajectories

The simulations done in this work were performed using the GROMACS software package.[116] The algorithms and methods used by this software will be introduced in the following.

4.2.1 Integration Method

A prerequisite for describing the dynamics of a molecule by means of the Newtonian equations of motion is a proper initial structure and initial atomic velocities. Structures with an atomic resolution are usually obtained from x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. Protein structures can be found in the Brookhaven Protein Data Bank,[117,118] which contain up to now more than 26 000 structures. Small simple molecules can be build using molecule editors.[119–122]

Based on these initial conditions, Newton's equations of motion are numerically solved iteratively in small time steps Δt. In the present work, the *leap-frog* algorithm[123] was employed. The advantage of this method is its numerical stability and, in contrast to the Runge-Kutta method,[124] that the expensive force calculation is done only once per integration step.

The algorithm calculates positions $\mathbf{r}$ at time t and velocities $\mathbf{v}$ at time $t-\frac{\Delta t}{2}$,

$$\mathbf{v}(t+\frac{\Delta t}{2}) = \mathbf{v}(t-\frac{\Delta t}{2}) + \frac{\mathbf{F}(t)}{m}\Delta t \tag{4.2}$$

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \mathbf{v}(t+\frac{\Delta t}{2})\Delta t\ . \tag{4.3}$$

The time step Δt has to be much smaller than the period of the fastest vibration in the system, which is typically the bond-stretching motion of the hydrogen atoms of 10–20 fs. Therefore, the time step is usually chosen to 1 fs. Larger time steps are possible by using constraints, which will be described in Sec. 4.2.5.

4.2.2 Solvent Environment

Solvent molecules strongly affect the properties of proteins and polymers.[125–128] In this work, the solvent molecules are explicitly described in the simulations, although the computation of the solvent dynamics is very expensive; it makes about 80–90% of the computation time. Thus, many models have

been proposed to treat the solvent effects implicitly.[129–135] In the following we will discuss why, in the present work, the explicit description of solvent molecules is indispensable.

Solvent molecules influence the properties of the solute in many different ways. First, polar solvent molecules force the solute to minimize its hydrophobic surface, which may strongly affect the conformation of the solute. This *hydrophobic effect* can be approximated in implicit solvent models by introducing a hydrophobic surface dependent energy term to the force field.[136] Furthermore, the dielectric shielding of molecular charges due to the polarizability of the solvent can be roughly described by using a dielectric coefficient $\varepsilon_r > 1$ ($\varepsilon_{\text{water}} \approx 80$ and $\varepsilon_{\text{methanol}} \approx 30$).

Another, much more computationally expensive, method to describe the dielectric shielding is to solve the linearized Poisson-Boltzmann equation at each point of the simulation system.[127,130,137] Besides these electrostatic effects, the dynamics of the solute is influenced by the viscosity of the solvent. Langevin dynamics[138,139] principally allows to describe this effect by introducing noise and friction forces.[140–142]

However, the translational and rotational diffusion depends on the particular interactions between the solvent and the solute. Since in this work the rotational diffusion of a dye molecule attached to a protein shall be studied in detail, the dye-solvent interactions must be described as accurate as possible. Moreover, the solvent properties in the vicinity of a protein, like, e. g., the viscosity, might significantly differ from its bulk properties.[143–145] Therefore, the usage of explicit solvent molecules in this work is necessary.

4.2.3 System Boundaries

In molecular dynamics simulations the studied system has to be many orders of magnitude smaller than in the experiment to be computationally tractable. To minimize artefacts from the system boundaries, which become important for such small system sizes, appropriate boundary conditions have to be chosen. Different solutions have been suggested to prevent the solvent molecules from evaporation, to counterbalance the arising high pressure due to the surface tension and to avoid preferred orientations of the solvent molecules on the surface.[146–150]

In the present work, periodic boundaries overcome any surface artefacts. For periodic boundaries, the simulation volume represents a space-filling box which is surrounded by translated copies of itself. Possible shapes of

the unit cell can be a cuboid, a dodecahedron, or a truncated octahedron. In a box with periodic boundaries, a molecule that leaves the box on one side, immediately reenters the box on the opposite side. In this way, the simulation system does not have any surface. However, artefacts may arise from the artificial periodicity, since the molecules also interact with their periodic images due to the long-range electrostatic interactions. These artefacts are minimized by increasing the box size.

In the present work, a rectangular box was chosen for all simulations. The choice of the box size is a trade off between minimization of artefacts from the periodic boundaries and minimization of the computational effort.

4.2.4 Temperature and Pressure Coupling

The molecular dynamics calculated by solving the Newtonian equations of motion conserves the total energy of the system (NVE ensemble). Whereas in real systems a molecular subsystem of the size studied in the simulation constantly exchanges energy with its surrounding. To be more close to reality, this energy exchange should therefore be introduced to the simulation. In addition, numerically solving the Newtonian equations leads to discretization and rounding errors, which introduces numerical noise, i. e., random forces, which heat up the system.

It is thus necessary to control the temperature T of the system, which should remain close to a given target temperature T_0. Several methods have been proposed for this purpose.[151–153]

In this work, the *Berendsen thermostat*[153] was used, where the coupling to a heat bath is achieved by correcting the actual temperature according to

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \quad , \tag{4.4}$$

which leads to a strongly damped exponential relaxation of the temperature towards the target temperature T_0 with a time constant τ. The change of the temperature is achieved by rescaling the velocities of each atom every step with a time-dependent factor λ given by

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)} \quad . \tag{4.5}$$

The time constant τ depends on the parameter τ_T

$$\tau = \frac{2C_V\tau_T}{N_f k_B} \quad , \tag{4.6}$$

where N_f denotes the total number of degrees of freedom, C_V the heat capacity, and k_B Boltzmann's constant. In all explicit solvent simulations presented in this work, the target temperature T_0 and the coupling time constant τ were chosen to 300 K and 0.1 ps, respectively.

In addition to the heat bath coupling, real biological systems are subjected to a constant pressure of usually 1 atm. Therefore, in the simulations, isobaric ensembles were generated by using a similar approach as for the temperature coupling. Now the pressure is corrected in each step (*Berendsen barostat*[153]) according to

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_p} \quad . \tag{4.7}$$

The pressure thus relaxes exponentially to the target pressure P_0, which was chosen to 1 atm in all simulations. The pressure correction is achieved by scaling the coordinates by a factor μ, given by

$$\mu = 1 - \frac{\Delta t}{3\tau_p}\kappa(P_0 - P) \quad , \tag{4.8}$$

where κ is the isothermal compressibility of the system.

4.2.5 Improving efficiency

The calculation of a trajectory by molecular dynamics simulations is computationally very expensive. Especially if long-range pair interactions shall be calculated exactly, since then the computational effort would scale quadratically with the number of atoms. For the calculation of a 20 ns trajectory of 50 000 atoms including all interactions exactly, about $6 \cdot 10^{16}$ floating point operations would be necessary. Assuming a typical (actual) computing power of 1.0 Gflops/s (floating point operations per second), the calculation would take about 40 years. In the following, several methods to improve the efficiency of the calculation, which are implemented in the used simulation software GROMACS, are described. All these methods, which include efficient treatment of long-range interactions, enlarging the time step, parallelization, and the use of *compound atoms*, allow the abovementioned 20 ns trajectory to be calculated in only a few months.

Efficient calculation of non-bonded forces

The most time consuming part of MD simulations is the calculation of electrostatic and van der Waals interactions (cf. Fig. 4.1), since all atomic pair interactions have to be considered; the computation thus scales as $\mathcal{O}(N^2)$ with the number N of atoms.

A simple method to improve the efficiency is to introduce a *cut-off* function,[109] where only interactions of atoms that are closer than a specified *cut-off* distance (typically 10–12 Å) are considered. This method reduces the cost to $\mathcal{O}(N)$. This is a good approximation for the short-range van der Waals potential, whereas for the long-range electrostatic potential the neglect of interactions with distant atoms has been shown to cause artefacts in the structure and dynamics of proteins.[154–160]

To avoid such artefacts, we used the *Ewald method*,[161] which was first introduced to describe long-range interactions of the periodic images in crystals and which is therefore particularly useful to describe systems with periodic boundaries.

Here the slowly converging sum of the electrostatic potential

$$V_{\mathrm{q}} = \sum_{\{\mathbf{n}\}} \sum_{\alpha\beta} \frac{q_\alpha q_\beta}{4\pi\epsilon_0\epsilon_{\mathrm{r}} r_{\alpha\beta,\mathbf{n}}} \tag{4.9}$$

$$= V_{\mathrm{dir}} + V_{\mathrm{rec}} + V_{\mathrm{o}} \quad , \tag{4.10}$$

where $\mathbf{n}$ is the box-vector, is replaced by two fast converging sums V_{dir} in the direct and V_{rec} in the reciprocal space and a constant term V_{o}. In this way, relatively small *cut-off* distances of about 1 nm in direct space and 10 wavenumbers in reciprocal space can be chosen. The computational cost of the *Ewald summation* still scales with $\mathcal{O}(N^2)$. An improved extension of the *Ewald summation* is the *particle-mesh Ewald* (PME)[162,163] method, which is implemented in GROMACS. It uses the *fast Fourier transform* (FFT) for the calculation of the reciprocal sum and was used in all explicit solvent simulations described in the present work.

Increasing the integration timestep

As has already been discussed, the maximum integration time step is limited by the smallest oscillation period found in the simulated system, which is typically due to bond-stretching vibrations. However, these bond-stretching vibrations are in the quantum-mechanical ground state and are therefore better represented by a *constraint* than by a harmonic potential. All bond lengths are thus constrained using the LINCS algorithm,[164] which, after an unconstrained integration step, rescales the bond lengths to their equilibrium lengths. The next fastest motions that remain, are the bond-angle vibrations with a typical period of about 20 fs. Thus, the integration time step can be increased to 2 fs, which fortunately leads to both a more correct description of the dynamics and an improved efficiency.

Parallelization

The computation time of an MD simulation can be further decreased by performing the calculation on several processors in parallel. This is achieved by distributing the atoms among the processors such that the data transfer, which is mainly due to the long-range interactions, is minimized, while on the other hand, the workload of each processor is maximized. In this work, all simulations were done using the two processors of double processor PC's, which have a very high data transfer rate due to the shared memory architecture.

Compound atoms

Since the interactions of nonpolar hydrogen atoms with its surrounding is only weak, it has been shown, that the implicit description of these hydrogen atoms by so called *compound atoms* does not significantly affect the physical properties of the system.[165] Therefore, the hydrogen atom and the heavy atom, to which it is bound, is merged to a *compound atom* with modified partial charges and van der Waals parameters. In this work, e. g., the methyl group (CH_3) in the methanol solvent is described by such a *compound atom*, which reduces the total system size by a factor of two. In contrast, the polar hydrogens interact strongly with its surrounding, in particular via the important hydrogen bonds, and are therefore described explicitly.

4.2.6 Minimization and Equilibration

The aim of minimization and equilibration is to generate a state of the simulation system with those atomic positions and velocities, which are close to the equilibrium at a specified temperature and where the energy is equally distributed among all degrees of freedom.

Molecular dynamics simulations usually uses initial protein structures determined by x-ray crystallography or NMR spectroscopy. The experimental structures have a limited resolution of typically 1.0 to 3.0 Å. Thus, these structures often suffer from deformations of bonds or bond-angles or strong van der Waals overlaps. A simulation starting from such a slightly perturbed structure, would quickly destabilize the whole system. To avoid this artefact, all systems were energy minimized prior to starting the dynamics simulation, using a steepest descent method to reach the nearest local minimum on the energy surface and thus relaxing all deformations.

During the subsequent equilibration, the system is coupled to a heat bath, as described in Sec. 4.2.4, which should bring the system to the desired

temperature where it further relaxes. If the system reaches an equilibrium state is monitored by plotting relevant observables, like the *root mean square deviation*, as described in the next section. If the drift of such relevant observables is absent or small the system is considered to be equilibrated.

4.2.7 Relevant observables

To decide, whether a simulation system is equilibrated, besides the individual energy terms, often the *root mean square deviation* (*rmsd*) of the actual structure to a reference structure is considered

$$rmsd \;=\; \min_{\{T,R\}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - x_i^0)^2 + (y_i - y_i^0)^2 + (z_i - z_i^0)^2\right]}\;, \qquad (4.11)$$

where x_i, y_i, z_i and x_i^0, y_i^0, z_i^0 are the cartesian coordinates of atom i of the actual and the reference structure, respectively, and $\{T, R\}$ is the set of all translations and rotations.

The mean square fluctuation (*rmsf*) of an atom i during the simulation time $T = M \cdot \Delta t$

$$rmsf_i \;=\; \min_{\{T,R\}} \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left[(x_i(t_j) - \bar{x}_i)^2 + (y_i(t_j) - \bar{y}_i)^2 + (z_i(t_j) - \bar{z}_i)^2\right]}\;, \qquad (4.12)$$

gives information on the local flexibility of a protein ($\bar{x}_i, \bar{y}_i, \bar{z}_i$ are the mean positions of atom i during the time T).

4.3 Parameterization of the Alexa488 dye

In this work, MD simulations of the fluorescent dye Alexa488 (C5 maleimide, *Molecular Probes*) were carried out. Since there are no force field parameters available for this molecule, they had to be developed, which is described in the following. Because the motional restriction of the dye due to the protein is mainly determined by steric hindrances and electrostatic interactions, we paid particular attention to those force field parameters which sensitively affect these quantities, i. e., van der Waals parameters and partial charges. The van der Waals parameters are almost independent of the chemical environment and are thus taken for corresponding atom types from the Gromacs force field. The aliphatic chain, the linker region of the dye, is already

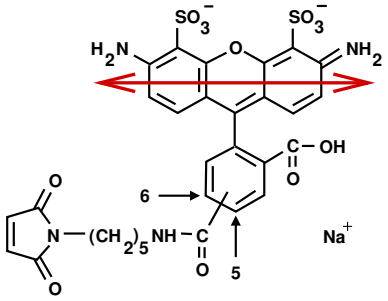

Figure 4.2: Chemical structure of the fluorescent label Alexa488. The red arrow indicates the transition dipole moment.

parametrized in the Gromacs force field, therefore all these already existing force field parameters were used for the linker. In the following, thus only the headgroup of the dye is considered, where the linker is replaced by a methyl group approximately retaining the chemical environment of the headgroup. For the quantum chemical calculation of the partial charges it is important to note, that the experiment observes the dye during the time between absorption and emission, therefore the dye parameters for the first excited state have to be determined. The most accurate method for the calculation of partial charges uses density functional theory (DFT), which is however only able to describe the electronic ground state. We therefore decided for the following strategy: Firstly, the ground state of this molecule was calculated using density functional theory (DFT) implemented in the DMol program[166] with the DNP basis set and the BLYP functional. Then both, the ground and first excited state were calculated using CIS/STO-3G with the GAUSSIAN program (ab-initio)[167] and also PM3 (semi-empirical) with the program MOPAC.[168] Atomic charges from all calculations were obtained by fitting to the electrostatic potential (ESP).[169] The differences of the charges between the ground and first excited state obtained from the ab-initio and semi-empirical calculations were averaged and added to the charges from the DFT calculation. To assure compatibility to the protein force field, all 20 amino acids were calculated using the same DFT method as for the dye and compared to the partial charges assigned to the amino acids in the Gromacs force field. From that, a mean scaling factor of 0.7 was obtained, which was then used to scale the dye charges. Subsequently, the charges were shifted to retain the correct total charge of 2e. To account for the symmetry of the dye, the charges were finally symmetrized. The geometry obtained from the ab-initio calculation was used to derive the equilibrium force field parameters of the bond lengths and angles. The force

constants describing the chemical bonds in the headgroup of the dye, which are rather responsible for the molecular vibrations than for the interaction with the protein, were defined according to chemically similar groups from the Gromacs force field. The Alexa488 dye is only available as a 5,6-isomere (cf. Fig. 4.2); we only used the 5-isomere in the simulation, assuming only minor dependence of the dye dynamics on the choice of the isomere.

"My curves are not crazy."
– Henri Matisse

5

Principal Curvilinear Coordinates and Correlations

In this chapter we develop a method to calculate principal curvilinear coordinates of molecular ensembles. This method will also allow, in the case of the protein-attached dye, to determine the mode of the protein motion, that is mostly correlated with the motion of the dye. The results of this calculation are shown in Chap. 7.

Structural ensembles of biomolecules obtained by molecular dynamics (MD) simulations or Monte-Carlo calculations generally comprise a large amount of data with a complex high-dimensional structure. Their efficient and adequate analysis is a prerequisite for gaining physical insight into dynamics, thermodynamics, and biological function from the ensemble. For this purpose, within the framework of statistical mechanics, the structure of an N-atomic biomolecule, e. g., a protein, is often described by a single point in the $3N$-dimensional configurational space. Typical ensembles thus contain $10^3 - 10^5$ structures in a $10^2 - 10^4$ dimensional configurational space.

Large extensions of the ensemble 'cloud' (see Fig. 5.1 A) in this configurational space represent large conformational motions. The largest conformational motion can be described by a principal coordinate, which is the degree of freedom along which the ensemble has the largest extension.

The established and widely used approach is the principal component analysis[93,94] (PCA) to obtain the principal coordinate in the configurational space. To this end, the covariance matrix $\mathbf{C} = \left\langle (\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T \right\rangle$ of cartesian fluctuations of the system is calculated from the ensemble of protein structures $\mathbf{x} \in \mathbb{R}^{3N}$ and subsequently diagonalized. The obtained eigenval-

ues describe the variances of the ensemble along the corresponding eigenvectors. The eigenvectors with large eigenvalues thus describe the essential motions, i. e., those which contribute most to the atomic fluctuations, and in this sense are principal coordinates of the protein. It has been shown that for a protein usually around 1% of the eigenvectors account for 90% of the total root mean square fluctuations of the protein,[94] which means the dimension of the ensemble can be drastically reduced without loosing the important structural features of the protein.

In the following, we focus only on a single principal coordinate, i. e., on the first eigenvector which corresponds to the largest eigenvalue and which maximizes the variance of the ensemble along this eigenvector. Furthermore, this eigenvector has the property that it also minimizes its root mean square deviation from the ensemble. Therefore, the motion along this (linear) principal coordinate is best correlated with the ensemble, with respect to all other possible linear coordinates.

It is remarkable that this approach is quite successful despite the fact that the essential coordinates obtained by PCA are linear. However, typical conformational motions in proteins are often more complex and frequently involve rotations of domains around hinge axes or dihedral angles of the protein backbone, which suggests that curvilinear coordinates, or 'principal rotations' should be more suitable to describe complex protein motions by as few degrees of freedom as possible. Especially the dye motion studied in this work is highly nonlinear.

Our goal here is to determine a mode of motion, that is 1. nonlinear and 2. best correlated to a specified subspace of the configurational space. The PCA is in both respects not sufficient for this purpose. In this chapter we will therefore develop a new method, which particularly accounts for these two requirements. The strategy for that is to design a method to calculate principal *curvilinear* coordinates from molecular ensembles and then, after that, to extend this method to also allow for the calculation of the mode of motion, that is correlated with a configurational subspace.

Roughly, our approach identifies a pre-specified number of prototypic structures (PS) $\mathbf{a}_j$, that characterize the main shape of the ensemble $\mathbf{x}_i$ (black circles in Fig. 5.1 A). The aim is to position these PS along the largest extension of the ensemble, such that moves along them capture the principal motions of the protein. Then, the PS can be used to construct a curvilinear coordinate, e.g. using a cubic spline function (dashed line in Fig. 5.1 A).

The PS are calculated as local ensemble averages, i. e., each structure of

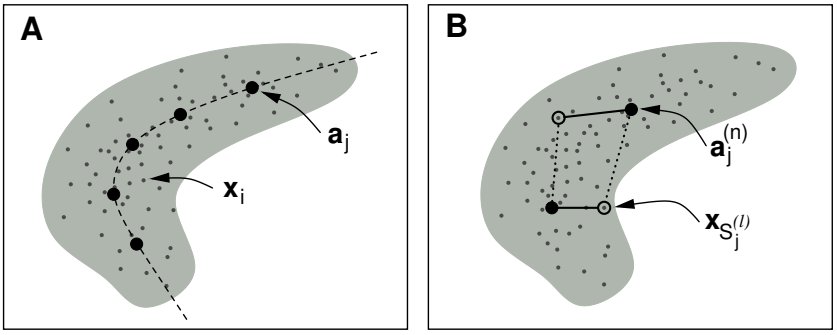

Figure 5.1: Two-dimensional sketch of how to derive curvilinear principle coordinates. (A) An ensemble of protein structures $\{\mathbf{x}_i\}$ (gray dots) is approximately described by a specified number of PS $\mathbf{a}_j$ (black dots) suitably positioned along the main extension of the ensemble and thus capturing the essential conformational motions of the protein. A cubic spline (dashed line) through the PS $\mathbf{a}_j$ yields the principal curvilinear coordinate. (B) Two possible assignments (solid and dotted lines) between randomly chosen $\mathbf{x}_i$ (hollow dots) and the PS $\mathbf{a}_j^{(n)}$ (solid dots) at the n-th iteration are shown. The assignment is chosen such that the mean square distance is minimized (solid lines).

the ensemble contributes to the averages with a different weight, which depends on its position in the ensemble. This approach is similar in spirit to clustering and vector quantization algorithms,[170] which are used for data reduction, data compression, and pattern recognition.[171–173] But in contrast to cluster algorithms, the goal here is not a minimal-error-representation of the ensemble, but, rather, to construct a curvilinear essential coordinate. Whereas cluster centers are represented by *points*, a coordinate is described by a *direction*. Thus, to obtain an essential coordinate, a different algorithm is required, which will be developed here.

In the subsequent theory section, we will first define the local ensemble averages (or PS). For a test case we will show how this new approach compares to the PCA. Then, an efficient algorithm to compute the PS is presented and analyzed. After that, the extension to determine modes of motion correlated with a configurational subspace is presented. The efficient algorithm is then applied to two example distributions, which is described in the results section. First, a two-dimensional artificial ensemble is used as an illustrative example and then, principal curvilinear coordinates of a protein ensemble are calculated using different numbers of PS.

5.1 Theory

We assume an ensemble of M protein structures $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ obtained by an MD-simulation or Monte-Carlo calculation (gray dots in Fig. 5.1). Since one is generally interested in the internal configurational motions only, we further assume that the ensemble is fitted to a reference structure, in order to eliminate the global translations and rotations of the molecule.[174]

The main idea of this approach, which distinguishes it from traditional cluster algorithms, is that a large number of randomly chosen k-tuples of structures probe the *shape* of the ensemble and, therefore, by averaging over all these crude estimates, the 'average' shape of the ensemble can be obtained. The k PS $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$, which shall represent the main shape of the ensemble, as shown in Fig. 5.1 A, are determined iteratively. We denote the iteration steps by a superscript. As a starting point, random structures are chosen for the $\mathbf{a}_j^{(0)}$. This first guess is then refined until the $\mathbf{a}_j^{(n)}$ are converged. In each iteration step n, the updated $\mathbf{a}_j^{(n)}$ is calculated as an average over an ensemble of k-tuples, which is generated from the original ensemble of protein structures in two steps; 1) selection and 2) permutation. First, one of the $\binom{M}{k}$ possible selections $S_1^{(l)}, \ldots, S_k^{(l)}$ $(l = 1, \ldots, \binom{M}{k})$ is randomly chosen, which yields a k-tuple $\{\mathbf{x}_{S^{(l)}_{\Pi_l(1)}}, \ldots, \mathbf{x}_{S^{(l)}_{\Pi_l(k)}}\}$ of structures (hollow circles in Fig. 5.1 B). Then, in the second step, from the $k!$ possible assignments of the $\mathbf{x}_i$ to the $\mathbf{a}_j^{(n)}$ (dotted lines), we choose the one (solid lines), that minimizes the mean square distance to the assigned points $\mathbf{a}_j^{(n)}$, i. e., for each selected k-tuple l, the permutation Π_l is determined, that fulfills the following assignment condition

$$\sum_{j=1}^{k} |\mathbf{a}_j^{(n)} - \mathbf{x}_{S^{(l)}_{\Pi_l(j)}}|^2 \stackrel{!}{=} \min \quad . \tag{5.1}$$

A crucial issue here is that, apparently, the choice of the permutation depends on the actual estimate (iteration step) for the PS $\mathbf{a}_j^{(n)}$. This assignment procedure is repeated for a large number of k-tuples, and improved PS $\mathbf{a}_j^{n+1}$ are calculated as an average over all structures $\mathbf{x}_i$, that were previously assigned to the corresponding $\mathbf{a}_j^n$,

$$\mathbf{a}_j^{(n+1)} = \left\langle \mathbf{x}_{S^{(l)}_{\Pi_l(j)}} \right\rangle_l \quad , j = 1, \ldots, k \; . \tag{5.2}$$

We call this method LMLA, since Localized Mean structures are calculated by using a Linear Assignment of tuples of structures. In the calculation of

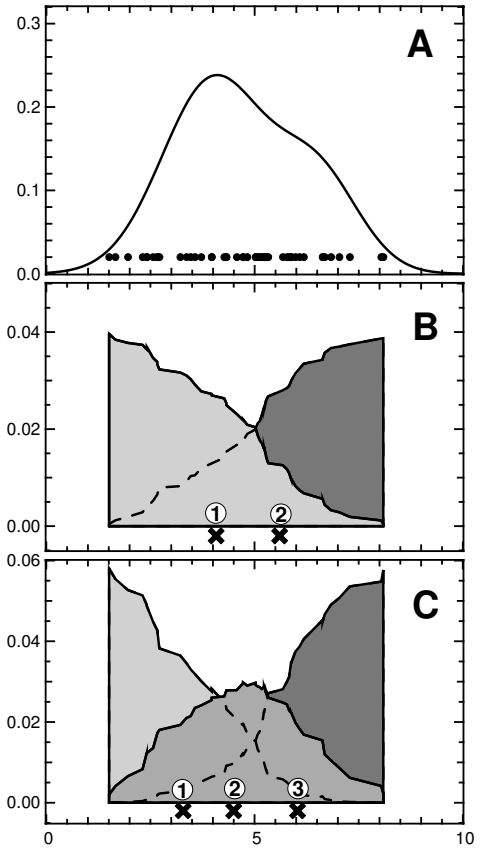

Figure 5.2: Example of weight functions corresponding to prototypic points from the LMLA-method. (A) An ensemble of points $\mathbf{x}_i$ (black dots) is distributed according to a given 1-dimensional probability density (solid line). (B) and (C) show two and three weight functions $p_j(x_i)$, which correspond to the two and three prototypic points a_j (black dots) and which are used in Eq. 5.3.

this average, each structure $\mathbf{x}_i$ appears as often as it has been assigned to $\mathbf{a}_j^n$, which accounts for multiple selections. Note that the average structure of the ensemble is equal to the average structure of the PS $\langle \mathbf{a}_j \rangle_j = \langle \mathbf{x}_i \rangle_i$, since each structure in a k-tuple is selected the same number of times. The iteration is repeated until the $\mathbf{a}_j^{(n)}$ are converged, e. g. , until $|\mathbf{a}_j^{(n+1)} - \mathbf{a}_j^{(n)}| < \epsilon$. Typically only three to ten steps are necessary to yield sufficiently converged PS.

We were not able to show rigorously that the PS converge to a unique solution in any case, but extensive numerical tests suggest that this is actually the case. In particular, even for quite different starting values $\mathbf{a}_j^{(0)}$ we never observed that the PS got trapped within a local minimum.

Fig. 5.2 illustrates the multiple selections mentioned above for a simple one-dimensional ensemble of 50 points $\{\mathbf{x}_i\}$ (black dots Fig. 5.2 A), which are distributed according to an arbitrary probability distribution (solid line). For this ensemble, the above described iteration was carried out for two and

three prototypic points. The obtained two and three points $\mathbf{a}_j$ are shown as crosses in Fig. 5.2 B and C, respectively. The functions $p_j(\mathbf{x}_i)$, shown in Fig. 5.2 B and C, are obtained by counting how often the structure $\mathbf{x}_i$ has been selected and assigned to $\mathbf{a}_j$. After normalization $\sum_i p_j(\mathbf{x}_i) = 1$, the function $p_j(\mathbf{x}_i)$ gives the probability that $\mathbf{x}_i$ is assigned to $\mathbf{a}_j$, if $\mathbf{x}_i$ is part of a randomly chosen k-tuple of points. Eq. 5.2 can thus be written as a weighted sum

$$\mathbf{a}_j^{(n+1)} = \sum_i \mathbf{x}_i p_j(\mathbf{x}_i) \quad . \tag{5.3}$$

The weight functions $p_j(\mathbf{x}_i)$ indicate how each point $\mathbf{x}_i$ contributes in Eqs. 5.3 and 5.2 to the calculation of the prototypic points $\mathbf{a}_j$.

5.1.1 Comparison of LMLA with conventional PCA

For the special, most simple case of just two PS, our approach yields a linear principal coordinate and, therefore, should behave similar to the conventional PCA. In this subsection we analyze if this is actually the case. Only if the two methods are sufficiently similar, our LMLA approach with more than two PS can be considered a proper generalization of the PCA. For the PCA, the first eigenvector (the one with the largest eigenvalue) maximizes the variance of the ensemble along this eigenvector.[175] An equivalent formulation is that the PCA minimizes the root mean square deviation of the ensemble from the first eigenvector.

To identify the quantity that is optimized by our LMLA method, we focus on the special case of a two-dimensional configurational space. We assume a given probability distribution $\rho(\mathbf{x})$, which describes the ensemble of structures considered above. It will be convenient to combine the two PS $\mathbf{a}_1$ and $\mathbf{a}_2$ into one four-dimensional vector $\mathbf{a} = (a_1, a_2, a_3, a_4)$ (see Fig. 5.3 A). For the general case, a k-tuple in N-dimensional space is described by a single vector in (Nk)-dimensional space.

The aim is now to calculate the stationary points $\mathbf{a}^{(\infty)} = \langle \mathbf{x}_\rho \rangle$ by appropriately averaging over the given distribution ρ. To that end, recall the assignment condition, Eq. 5.1, which in this two-dimensional case reduces to

$$\begin{aligned}(x_1 - a_1)^2 + (x_2 - a_2)^2 + (x_3 - a_3)^2 + (x_4 - a_4)^2 \\ \leq (x_1 - a_3)^2 + (x_2 - a_4)^2 + (x_3 - a_1)^2 + (x_4 - a_2)^2 \quad .\end{aligned} \tag{5.4}$$

Each $\mathbf{x} = (x_1, x_2, x_3, x_4)$, that does not fulfill this equation, is permuted, i. e. , mapped onto the permuted structure. Since only two PS are used, there

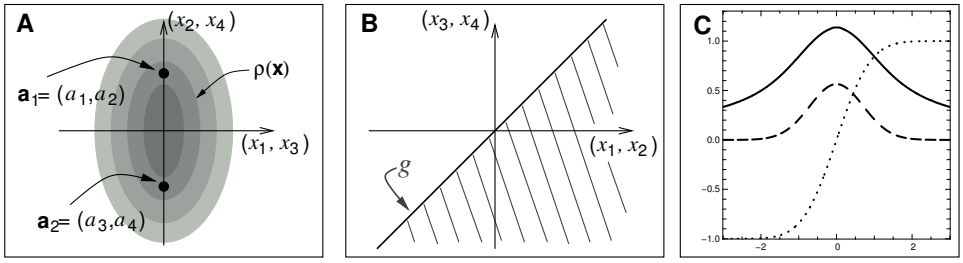

Figure 5.3: (A) Sketch of a two-dimensional probability distribution $\rho(x)$ with two prototypic points obtained by the LMLA-method, which are positioned along the largest extension of the distribution. (B) Schematic plot of the integration volume (hatched area) defined in Eq. 5.4 and used in Eq. 5.5. Permutation of two points $\mathbf{x}_1$ and $\mathbf{x}_2$ means mirroring with respect to the dividing hypersurface g, which depends on the prototypic points $\mathbf{a}_1$ and $\mathbf{a}_2$. (C) A one-dimensional probability distribution (dashed line) and its corresponding weight function $w(x)$ (solid line), according to Eq. 5.13. The dotted line visualizes the integral $\int_{-x_2}^{x_2} dx_4 \rho_2(x_4)$ appearing in Eq. 5.13.

are only two permutations possible. Therefore, all $\mathbf{x}$ that fulfill Eq. 5.4 lie in a half-space of the four-dimensional configuration space, which is schematically shown as the hatched area in Fig. 5.3 B.

If the prototypic points are assumed to be already converged, i. e., if n is sufficiently large, such that $\mathbf{a}_j^{(n+1)} \approx \mathbf{a}_j^{(n)}$, then the calculation of the average corresponding to Eq. 5.2 in this case is

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = 2 \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \rho(x_1, x_2) \int_{-\infty}^{\infty} dx_3 \int_{-\infty}^{\beta} dx_4 \rho(x_3, x_4) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \quad (5.5)$$

where the factor 2 is due to the normalization of ρ and to the fact that the integration is carried out only in the half-space, which is defined by the integration limit β

$$\beta \equiv (x_1 - x_3) \frac{a_1 - a_3}{a_2 - a_4} + x_2 \, . \quad (5.6)$$

The integration limit β is obtained by solving Eq. 5.4 for x_4.

For further simplification, we assume ρ to factorize, i. e., $\rho(x_1, x_2) = \rho_1(x_1)\rho_2(x_2)$, and both, ρ_1 and ρ_2 to be even, i. e., $\rho_1(x) = \rho_1(-x)$ and $\rho_2(x) = \rho_2(-x)$. It is shown in the Appendix I that then $a_2 + a_4 = 0$ and $\beta = x_2$ and

$a_1 = a_3 = 0$, such that Eq. (5.5) simplifies to

$$\begin{pmatrix} 0 \\ a_2 \\ 0 \\ a_4 \end{pmatrix} = 2 \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \rho_1(x_1)\rho_2(x_2) \int_{-\infty}^{\infty} dx_3 \int_{-\infty}^{x_2} dx_4 \rho_1(x_3)\rho_2(x_4) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \tag{5.7}$$

which yields

$$\begin{aligned} a_2 = 2 & \left(\int_{-\infty}^{\infty} dx_1 \rho_1(x_1) \right) \left(\int_{-\infty}^{\infty} dx_3 \rho_1(x_3) \right) \\ & \times \left(\int_{-\infty}^{\infty} dx_2 \rho_2(x_2) x_2 \int_{-\infty}^{x_2} dx_4 \rho_2(x_4) \right) \\ = 2 & \int_{0}^{\infty} dx_2 \rho_2(x_2) x_2 \int_{-x_2}^{x_2} dx_4 \rho_2(x_4) \; . \end{aligned} \tag{5.8}$$

The transformation of the integration limits is described in Appendix II. Since $a_2 = -a_4$ and ρ_2 is even,

$$a_2 - a_4 = 2a_2 = 4 \int_{0}^{\infty} dx_2 \rho_2(x_2) x_2 \int_{-x_2}^{x_2} dx_4 \rho_2(x_4) \tag{5.9}$$

$$\Rightarrow a_2 = \int_{0}^{\infty} dx_2 \rho_2(x_2) x_2 \int_{-x_2}^{x_2} dx_4 \rho_2(x_4) \tag{5.10}$$

This is the result of the LMLA approach, that is to be compared with the result from the PCA,

$$\sigma = \int_{-\infty}^{\infty} dx_2 \; \rho_2(x_2) \; x_2^2 \; . \tag{5.11}$$

To interpret Eq. 5.10, we assume ρ to have a shape similar to a gaussian function, i. e., ρ is centered around its average value, with the maximum near the center, and with a compact support. This assumption does not involve severe restrictions, since ensembles of protein structures are often gaussian-like distributed, at least in a first approximation. Then function ρ does not vary much in the center region and therefore the right integral in Eq. 5.10 becomes

$$\int_{-x_2}^{x_2} dx_4 \; \rho_2(x_4) \sim x_2 \qquad \text{for small } x_2, \tag{5.12}$$

and the integral in Eq. 5.10 becomes the variance, and therefore yields the same result as the PCA. For large x the integral in Eq. 5.12 becomes independent of x_2, i. e., large extensions are less weighted in Eq. 5.10 than in the normal variance. By introducing a weight function

$$w(x) = \frac{1}{x_2} \int_{-x_2}^{x_2} dx_4 \rho_2(x_4) \ , \tag{5.13}$$

Eq. 5.10 then becomes

$$a_2 = \int_{-\infty}^{\infty} dx_2 \ \rho_2(x_2) x_2^2 w(x_2) \tag{5.14}$$

In this respect, our approach, in Eq. 5.10, represents a generalized variance. For $w(x) = 1$ the expression for the normal variance is recovered. Fig. 5.3 C shows the relationship between $\rho(x)$, which is in this example a gaussian distribution, and the corresponding weight function $w(x)$. The weight function $w(x)$ is approximately constant around the center of the gaussian distribution and decays to zero in the more distant region. For the determination of the principal curvilinear coordinate, our approach concentrates more on the shape of the center regions of the ensemble than the PCA does.

5.1.2 An efficient algorithm

The iteration to determine the PS, as described above, requires the calculation of a multidimensional average value in each step. The number of selections in Eq. 5.2 is $\binom{M}{k}$, which becomes very large due to the combinatorial explosion, even for typical ensembles using only few PS. Thus, the calculation of the sum in Eq. 5.2 is often intractable. We therefore present, as an alternative, a faster yet approximative algorithm, which refines the PS and calculates the average according to Eq. 5.2 simultaneously. This new algorithm generally yields a solution with sufficient accuracy for defining a suitable principal curvilinear coordinate. The idea is that the average over the selected k-tuples is now calculated stepwise and the actual guess for the PS is updated after each step, instead of calculating the complete average for one refinement step of the PS, as was done previously. Fig. 5.4 shows two steps of this algorithm for the two-dimensional case using only two PS $\mathbf{a}_1$ and $\mathbf{a}_2$.

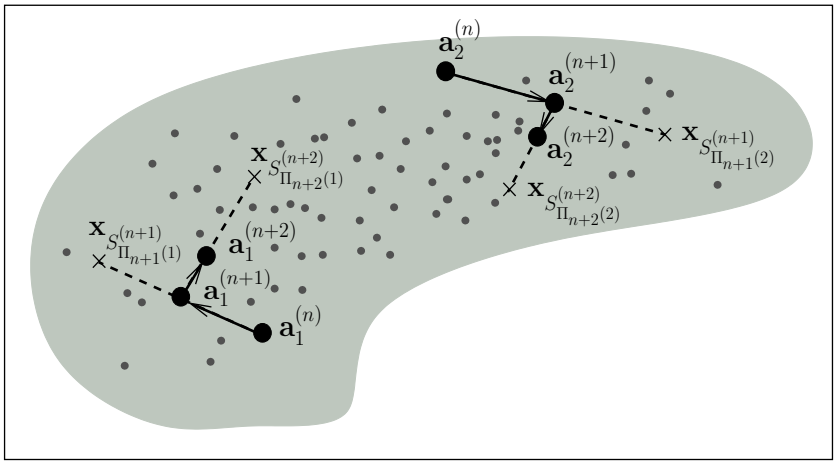

Figure 5.4: Schematic plot of two steps of the efficient LMLA-algorithm for two PS in two dimensions. In the n-th step, the actual guess for the PS is $\mathbf{a}_1^{(n)}$ and $\mathbf{a}_2^{(n)}$. The randomly chosen pair $(\mathbf{x}_{S_1^{(n+1)}}, \mathbf{x}_{S_2^{(n+1)}})$ is assigned to the actual pair of PS to yield the lowest *rmsd*. The new PS $(\mathbf{a}_1^{(n+1)}, \mathbf{a}_1^{(n+1)})$ are obtained by the update formula Eq. 5.16. Note that the stepsize becomes continuously smaller in each step, ensuring the convergence of the PS.

The algorithm in detail reads:

1. Choose a random k-tuple $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ of structures from the ensemble as the inital guess for the PS. Set $n = 1$.

2. Choose another k-tuple $\{\mathbf{x}_{S_1^{(n)}}, \ldots, \mathbf{x}_{S_k^{(n)}}\}$ randomly out of the ensemble.

3. Determine the permutation Π_n, that minimizes the mean square deviation
$$\sum_{j=1}^{k} |\mathbf{a}_j^{(n)} - \mathbf{x}_{S^{(n)}_{\Pi_n(j)}}|^2 \stackrel{!}{=} \min \quad . \tag{5.15}$$

4. Calculate the average for all new $\mathbf{a}_j$
$$\mathbf{a}_j^{n+1} = \frac{n\,\mathbf{a}_j^n + \mathbf{x}_{S^{(n)}_{\Pi_n(j)}}}{n+1} \;, \quad j = 1, \ldots, k \;. \tag{5.16}$$

5. Set $n = n + 1$ and go to step 2 until the PS are converged.

The assignment problem is solved here by using the *Hungarian method*.[176] Note that each $\mathbf{x}_i$ is weighted equally in the mean value and note that here

again the mean value of the PS $\mathbf{a}_j$ is equal to the mean value of all selected structures $\mathbf{x}_{S_j^{(n)}}$:

$$\langle \mathbf{a}_j \rangle_j = \left\langle \mathbf{x}_{S_j^{(n)}} \right\rangle_{n,j} \quad . \tag{5.17}$$

5.1.3 Correlations

We have used the described LMLA-method to determine a principal curvilinear coordinate which best represents the largest conformational motion. This is the mode of motion that is best correlated with the ensemble in the complete configurational space K. To analyze what motions of the protein affect the motion of the dye attached to it, however, we have to answer a somewhat different question, namely: Which collective mode of motion is best correlated with only a subspace of the configurational space?

Recall that in the LMLA-method, the essential step to introduce information on the correlation of the PS with the ensemble is the assignment step (see e. g. Eqs. 5.1 and 5.15). This step assures that mainly 'neighbored', and thus correlated structures are averaged to generate the PS. Now we want the PS to be correlated with the extension of the ensemble only within a specific subspace D. Therefore, the mean square distance in the assignment condition Eq. 5.1 is now calculated in the subspace D,

$$\sum_{j=1}^{k} \left[\sum_{m \in D} \left(a_j^{(n)}(m) - x_{S^{(l)}_{\Pi_l(j)}}(m) \right)^2 \right] \stackrel{!}{=} \min \quad . \tag{5.18}$$

Note that in contrast, the averages in Eqs. 5.2 and 5.16 are of course still calculated in the complete configurational space, since the principal coordinate shall also describe the protein and not only the dye.

For illustration, we discuss two extreme cases (Fig. 5.5): In the first case (Fig. 5.5 A), the ensemble in the complemented space D^c ($D \cup D^c = K$), which here is the one-dimensional x-subspace, is uncorrelated with the ensemble in the subspace D, the yz-space; in the second case (Fig. 5.5 B), it shall be fully correlated. To better visualize the three-dimensional structure of the two sets, projections of the ensembles onto the xz- and yz-subspace are also shown.

In the first case, the depicted ensemble has the largest extension in x-direction. The red vector is the principal coordinate obtained from the conventional LMLA-algorithm using two PS, where the assignment is done

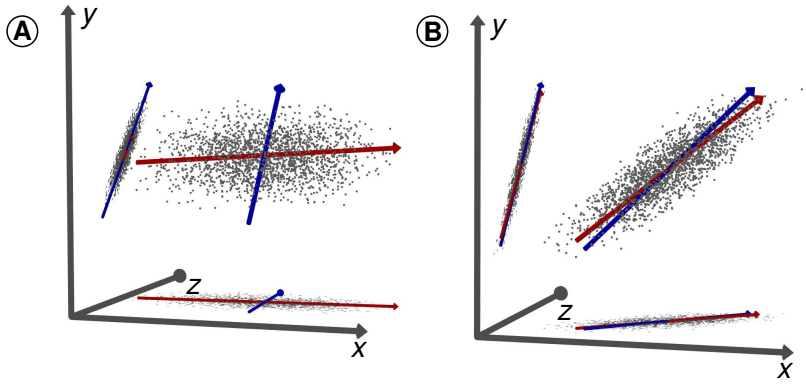

Figure 5.5: Two ensembles showing extreme cases of correlations. Projections of the ensembles onto the xz- and yz-subspace are also shown. In (A), the coordinates y and z are highly correlated, but the x-coordinate is not correlated with the two others. In (B), all three coordinates are highly correlated. The red vectors are the principal coordinates, calculated using the conventional LMLA-algorithm and therefore describe the largest extension of the ensembles. The blue vectors are the principal coordinates, that are best correlated within the yz-subspace.

in the complete configurational space. It therefore points in x-direction, capturing the largest extension and yielding the highest correlation with the given dataset. Obviously, the y- and z-coordinates are highly correlated, whereas the x-coordinates are neither correlated with the y- nor with the z-coordinates. The blue vector describes the mode of motion which is best correlated with the ensemble in the yz-subspace. In the second case, all three coordinates x, y, and z are correlated. The red and blue vectors are calculated like in the previous example. The red vector is therefore again oriented along the largest extension of the ensemble. However, the blue vector now adopts about the same orientation, since all coordinates are highly correlated.

5.1.4 From prototypic structures to curvilinear coordinates

Once the PS are determined, they are used to generate a cubic spline, which represents the principal curvilinear coordinate. To calculate this spline, first, the PS have to be put in order. This is achieved by determining the shortest path connecting all PS in the configurational space. This problem is referred to as the Traveling Salesman Problem,[177,178] a well-known problem in graph theory. If only a small number of PS are used, it is feasible just to calculate the length of all possible paths and then to choose the shortest one. Two

additional points are then constructed by extrapolating linearly the first two and the last two PS, ensuring the spline function generated from this extended set of PS to cover the whole ensemble, as shown in Fig 5.1 A.

5.2 Results

The LMLA-algorithm for the calculation of principal curvilinear coordinates is applied, as a first illustrative example, to an artificial two-dimensional ensemble, built up from three gaussian distributions with different weights (0.5,0.3,0.2), shown in Fig. 5.6 A.

The distribution is represented by an ensemble of 10 000 points, therefore the partial distributions 1,2 and 3 contain 5000, 3000 and 2000 points, respectively. The calculation of the PS was done using 10 000 steps of the LMLA-algorithm. The PS then converged to a relative stepsize of less than 0.01%. As already stated above, the efficient algorithm is an approximation to the exact solution. For the here considered examples, the relative error in the determination of the PS by this efficient algorithm is smaller than 1%.

First, only two PS (crosses) are used, yielding a linear coordinate (solid line). For comparison, the result of the PCA is also shown in Fig. 5.6 A (dashed line). The difference between both vectors illustrates the difference in the weight function $w(x)$, as described in the previous section. In the LMLA-algorithm the distant points in the ensemble are weighted less in the mean calculation. That means, the LMLA-coordinate describes the partial distributions 1 and 2 better than the PCA-coordinate, but gives less weight to the points of the partial distribution 3, since these are more distant to the total mean position of $\rho(x)$. This effect vanishes for symmetric distributions; in this case both methods would result in the same vector.

Fig 5.6 B shows the same ensemble as in Fig. 5.6 A, but now three (crosses) and four (circles) PS are used. Curvilinear principle coordinates are obtained by generating a cubic spline from the set of three (dashed line) and four (solid line) PS, as described in the previous section. Apparently, they describe the ensemble much better than the linear coordinate. The *rmsd* of the ensemble points to the PCA-coordinate is 1.95, whereas it is 1.38 and 1.33 for three and four PS.

In the second example, our algorithm is applied to an ensemble of protein structures. The CONCOORD program[179] was used to calculate an ensemble of 500 structures of the Bovine Pancreatic Trypsin Inhibitor (BPTI),

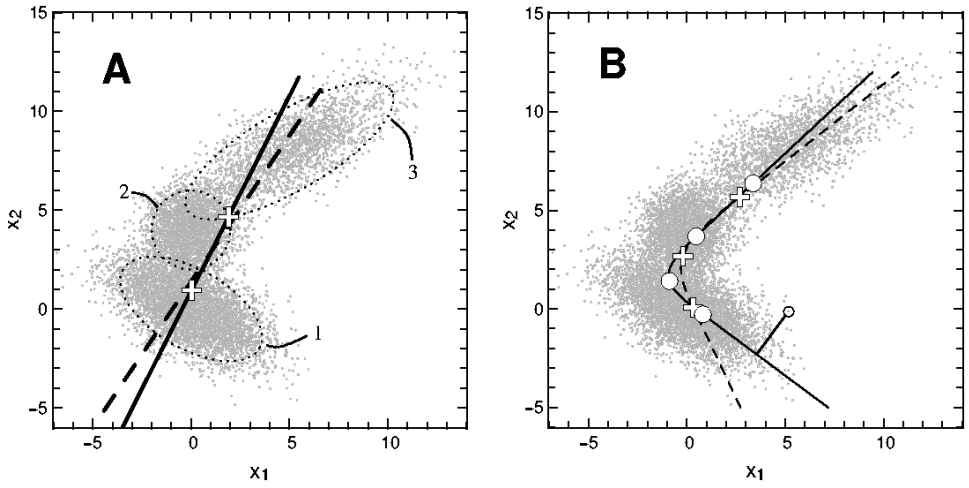

Figure 5.6: (**A**) Comparison between the PCA (solid line) and the LMLA-algorithm (dashed line) in a two-dimensional artificial ensemble, which is built up from three gaussian distributions indicated by the dotted ellipsoids. The obtained PS (crosses) define a linear coordinate (dashed line). (**B**) The use of more probe points leads to a curvilinear principal coordinate, which more accurately describes the ensemble. Three (crosses) and four (circles) PS are used to construct cubic splines (dashed line and solid line, respectively).

which comprises 58 residues. Its conformational space has 1356 dimensions. CONCOORD generates an ensemble of protein structures from a given set of distance restraints, originating, e. g., from hydrogen bonds or hydrophobic contacts. For this ensemble, curvilinear coordinates are calculated as described above using different numbers of PS (2-8). As a measure of how accurately the coordinate describes the ensemble, we calculate the root mean square distance (*rmsd*) of all ensemble structures to the spline. The distance of one ensemble point to the spline is shown in Fig. 5.6 B for illustration. The better the coordinate describes the overall shape of the distribution, the smaller is the perpendicular deviation of the ensemble structures from the spline.

Fig. 5.7 shows the *rmsd* of the BPTI ensemble to the spline functions generated from two to eight PS. The linear vector obtained by two PS shows a slightly larger *rmsd* than the first eigenvector from the PCA (dashed horizontal line), since the PCA exactly minimizes this *rmsd* value. However, three PS already yield a slightly better coordinate, while the advantage becomes more obvious when using more than three PS.

Note that the *rmsd* value is calculated in the high-dimensional conforma-

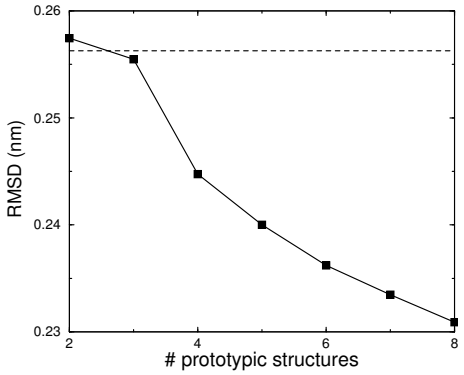

Figure 5.7: The root mean square distance (*rmsd*) from the points of the protein ensemble to a principal curvilinear coordinate obtained by different numbers of PS (black dots). Note that for two PS, the coordinate is linear, whereas all other principal coordinates are curvilinear. The *rmsd* to the linear coordinate obtained by a PCA (dashed line) is also shown for comparison.

tional space and therefore, relatively large changes of the *rmsd* in a small configurational subspace are hidden by the contributions of the other dimensions to the *rmsd*.

Just as an illustration, one could imagine the change of the *rmsd* from 0.231 nm for the LMLA-algorithm using eight PS to 0.256 nm for the PCA to be due to a modification of the ensemble instead of changing the coordinate. As a modification, we choose a variation of the width of the ensemble in just one dimension, the whole *rmsd* change is then due to only this one dimension. For example, if an $3N$-dimensional isotropic gaussian distribution, centered at the origin, with a variance $\sigma = 1$ is extended in one dimension by a factor of two, the *rmsd* from the origin changes by a factor of $\sqrt{(3N+1)/3N}$. For the protein ensemble, the same change in the *rmsd* as above (0.231 nm to 0.256 nm) would be observed, if the width of the ensemble is extended in one dimension by a factor of about 300. That shows the extent to which our principal curvilinear coordinates improve the description of the ensemble, compared to the PCA.

In this work, only a one-dimensional principal curvilinear coordinate has been determined. If a higher dimensional essential conformational subspace of a molecular ensemble is required, we suggest an iterative approach. First, a one-dimensional principal curvilinear coordinate is calculated as described before, *then* the ensemble is projected onto this coordinate, *then* the LMLA-method is again applied to the projected ensemble, yielding a second prin-

cipal curvilinear coordinate, and so on.

5.3 Discussion

For the selected cases analyzed above, our LMLA-algorithm yields a principal curvilinear coordinate of molecular ensembles, which describes the main molecular motion better than the PCA. For a small ensemble of protein conformations, which is e. g. distributed around the native structure, a harmonic approximation of the distribution is often adequate, in which case the linear principal coordinate obtained by PCA is sufficiently accurate. However, for the analysis of bended ensembles describing, e. g., large scale conformational motions of proteins or systems where the nonlinear motions are of particular interest, like the motion of a dye, which is studied in this work further below, the LMLA should, therefore, be particularly useful.

Furthermore, the LMLA-approach offers the possibility to study correlations between motions of different parts of the studied system, like in this work a dye and a protein. This approach is generally applicable to all kind of molecular ensembles to calculate, e. g., correlations of protein-ligand or protein-protein motions. Moreover, this method could also be used to identify the particular mode of motion of a protein, which is correlated with any given property of the protein, like, e. g., radius of gyration, *rmsd*, or water accessible surface.

The LMLA-algorithm differs from conventional clustering algorithms, in that the focus is extracting the *shape* and *extension* of an ensemble, while cluster algorithms typically aim at finding the *positions* of cluster centers. However, it is similar to clustering algorithms in that both reduce a large dataset by representing its main features using a small number of prototypic points or cluster centers, respectively. Although the LMLA-algorithm is not optimized for finding cluster centers, it might also be applicable in the framework of vector quantization for codebook generation. Since many vector quantization algorithms suffer from a local minimum problem, our algorithm might be advantageous in this respect, because in extensive numerical tests our algorithm never got trapped within a local minimum.

5.4 Appendix I

Here we show that, if ρ_1 and ρ_2 are even and $a_2 \neq a_4$, then Eq. (5.7) yields $a_2 + a_4 = 0$ and $a_1 = a_3 = 0$. First note that Eq. 5.5 can also be written as

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = 2 \int_{-\infty}^{\infty} dx_3 \int_{-\infty}^{\infty} dx_4 \, \rho(x_3, x_4) \int_{-\infty}^{\infty} dx_1 \int_{\gamma}^{\infty} dx_2 \, \rho(x_1, x_2) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} . \tag{5.19}$$

The integration limit γ

$$\gamma \equiv (x_3 - x_1) \, \frac{a_1 - a_3}{a_2 - a_4} + x_4 \tag{5.20}$$

is now obtained by solving Eq. 5.4 for x_2. The integration is here carried out over the same half-space as defined by β in Eq. 5.6.

Eqs. 5.5 and 5.19 give for a_2 and a_4

$$\begin{aligned} a_2 \;=\; & \int_{-\infty}^{\infty} dx_1 \rho_1(x_1) \int_{-\infty}^{\infty} dx_2 \rho_2(x_2) x_2 \\ & \int_{-\infty}^{\infty} dx_3 \rho_1(x_3) \int_{-\infty}^{\beta} dx_4 \rho_2(x_4) \end{aligned} \tag{5.21}$$

$$\begin{aligned} a_4 \;=\; & \int_{-\infty}^{\infty} dx_3 \rho_1(x_3) \int_{-\infty}^{\infty} dx_4 \rho_2(x_4) x_4 \\ & \int_{-\infty}^{\infty} dx_1 \rho_1(x_1) \int_{\gamma}^{\infty} dx_2 \rho_2(x_2) \qquad , \end{aligned} \tag{5.22}$$

where β and γ are defined by Eqs. 5.6 and 5.20, respectively. Permutation of x_2 and x_4 and the change of the latter integration limit to the negative range (since ρ_2 is even) leads to

$$\begin{aligned} a_4 = & \int_{-\infty}^{\infty} dx_3 \rho_1(x_3) \int_{-\infty}^{\infty} dx_2 \rho_2(x_2) x_2 \int_{-\infty}^{\infty} dx_1 \rho_1(x_1) \int_{-\infty}^{\delta} dx_4 \rho_2(x_4) \, , \\ & \text{with} \quad \delta = -[(x_3 - x1) \frac{a_1 - a_3}{a_2 - a_4} + x_2] \, . \end{aligned} \tag{5.23}$$

With x_2 for $-x_2$ substituted, the upper integration limit σ becomes β, hence $a_4 = -a_2$. To show that $a_1 = a_3$, the same transformations as above are done for

$$a_1 = \int_{-\infty}^{\infty} dx_1 \rho_1(x_1) x_1 \int_{-\infty}^{\infty} dx_2 \rho_2(x_2) \int_{-\infty}^{\infty} dx_3 \rho_1(x_3) \int_{-\infty}^{\beta} dx_4 \rho_2(x_4) \qquad (5.24)$$

and

$$a_3 = \int_{-\infty}^{\infty} dx_3 \rho_1(x_3) x_3 \int_{-\infty}^{\infty} dx_4 \rho_2(x_4) \int_{-\infty}^{\infty} dx_1 \rho_1(x_1) \int_{\gamma}^{\infty} dx_2 \rho_2(x_2) \qquad , \quad (5.25)$$

which corresponding to Eq. 5.23 yields

$$\begin{aligned} a_3 \quad = \quad & \int_{-\infty}^{\infty} dx_3 \rho_1(x_3) x_3 \int_{-\infty}^{\infty} dx_2 \rho_2(x_2) \\ & \int_{-\infty}^{\infty} dx_1 \rho_1(x_1) \int_{-\infty}^{\delta} dx_4 \rho_2(x_4). \end{aligned} \qquad (5.26)$$

Now the substitution of x_2 for $-x_2$ yields $a_3 = a_1$, which immediately leads to $\beta = x_2$. Using Eq. 5.5, this then gives $a_1 = a_3 = 0$, due to the assumed symmetry of ρ_1.

5.5 Appendix II

Here we show that for $\rho(x)$ an even function, the following equation holds

$$\int_{-\infty}^{\infty} dx \rho(x)\, x \int_{-\infty}^{x} dy \rho(y) = \int_{0}^{\infty} dx \rho(x)\, x \int_{-x}^{x} dy \rho(y)\,. \qquad (5.27)$$

The integral on the left side can be divided into parts:

$$\begin{aligned} \int_{-\infty}^{\infty} dx \rho(x)\, x \int_{-\infty}^{x} dy \rho(y) = & \int_{0}^{\infty} dx \rho(x)\, x \int_{-x}^{-x} dy \rho(y) \\ & + \int_{0}^{\infty} dx \rho(x)\, x \int_{-x}^{x} dy \rho(y) + \int_{-\infty}^{0} dx \rho(x)\, x \int_{-\infty}^{x} dy \rho(y)\,. \end{aligned} \qquad (5.28)$$

Substituting this expression in Eq. 5.27, the integral on the right side of Eq. 5.27 cancels with the second term on the right side of Eq. 5.28, which yields

$$\int_0^\infty dx\rho(x)\, x \int_{-\infty}^{-x} dy\rho(y) + \int_{-\infty}^0 dx\rho(x)\, x \int_{-\infty}^{x} dy\rho(y) = 0 \qquad . \tag{5.29}$$

The second term can be written as

$$\begin{aligned} \int_{-\infty}^0 dx\rho(x)\, x \int_{-\infty}^{x} dy\rho(y) &= -\int_0^{-\infty} dx\rho(x)\, x \int_{-\infty}^{x} dy\rho(y) && (5.30)\\ &= -\int_0^\infty d(-x)\rho(-x)\,(-x) \int_{-\infty}^{-x} dy\rho(y) && (5.31)\\ &= -\int_0^\infty dx\rho(x)\, x \int_{-\infty}^{-x} dy\rho(y)\,. && (5.32) \end{aligned}$$

Substituting this expression into Eq. 5.29 proves Eq. 5.27.

"All good things are wild, and free."
– Henry David Thoreau

6

Fluorescence Anisotropy of a Free Dye

The fluorescence anisotropy is related to the rotational diffusion of a dye. Molecular dynamics (MD) simulations should be able to describe this diffusional motion, from which the anisotropy can be obtained, as described in Sec. 2.1.1. In this chapter, we study if and to what extent MD simulations allow to predict the fluorescence anisotropy of a dye and if the used dye and solvent force fields are appropriate for this purpose. To this aim, simulations of free dyes in methanol and water were carried out and compared to experimental results[180] via the fluorescence anisotropy. For our study, we used the Alexa488 dye, which was also used in the experiment described in Chap. 7. The parametrization of this dye is described in Sec. 4.3. For comparison, we also used the rhodamine 6G dye, for which force field parameters were available.[181]

6.1 Molecular dynamics simulations

MD simulations of the dyes Alexa488 and rhodamine 6G in methanol and different water models were carried out. In all simulations, the dye was free to undergo translational and rotational diffusion within the (periodic) simulation volume. All MD simulations were performed using the GROMACS simulation software.[116] The SPC[182] and SPC/E[183] water models and the methanol parameters, which are included in the GROMACS force field, were used. The Alexa488 and rhodamine 6G systems additionally contain two sodium and one chloride ions, respectively, to use the same ion/dye

ratios as in the experiments. All systems were energy minimized to obtain the starting configuration for the simulations. The solvent and the dye were jointly coupled to an external temperature bath of 300 K with a relaxation time of 0.1 ps.[153] In all simulations the system was weakly coupled to a pressure bath of 1 atm with isotropic scaling and a relaxation time constant $\tau_p =$ 1 ps. Bond lengths were constrained to their equilibrium lengths using the LINCS algorithm.[164] This allows a 2 fs time step for the leapfrog integration scheme. For the Lennard-Jones interactions, a cutoff distance of 1 nm was applied. Electrostatic interactions between charge groups at a distance less than 1 nm were calculated explicitly, and the long-range electrostatic interactions were calculated using the Particle-Mesh Ewald method[162] with a grid spacing of 0.12 nm and a fourth-order spline interpolation. Coordinates of all atoms were saved every 1 ps for further analysis.

A list of all performed simulations is shown in Tab. 6.1. For each simulation, the fluorescence anisotropy decay was calculated, as described in 2.1.1. The rotational correlation time was then obtained by fitting a single exponential function to the anisotropy decay curve.

Dye	Solvent	N	T
1) Alexa488	methanol	1315	2.6 ns
2) Alexa488	SPC water	4673	8.5 ns
3) Alexa488	SPC water	17147	1.5 ns
4) Alexa488 modified charges	SPC water	4673	50 ns
5) Rhodamine 6G	SPC water	4036	5.0 ns
6) Alexa488	SPC/E water	4673	4.5 ns
7) Alexa488, no constraints	SPC water	14073	8 ns
8) Tryptophan	SPC water	4656	7 ns

Table 6.1: List of simulation systems, with the total number N of atoms and the simulation time T.

In simulation 4, the partial charges of the Alexa488 were modified to increase the polarity of the headgroup of the dye. This was achieved by decreasing the partial charges of the inner atoms and increasing the charges of the outer atoms by 10%. This modification intended to reflect the maximum uncertainty expected for the partial charges calculated in 4.3. In simulation 7, no bond length constraints (LINCS) were used. In simulation 8, a tryptophan using the GROMACS force field parameters solvated in SPC water was calculated. The absorption and emission spectra of tryptophan are due to two low-lying excited states ($^{\mathrm{I}}\mathrm{L}_a$ and $^{\mathrm{I}}\mathrm{L}_b$), which have different transition dipole moment orientations. For the comparison of the calculated anisotropy with

the experiment, the absorption and emission were assumed to be solely due to the $^{\mathrm{I}}\mathrm{L}_a$ state.[184]

6.2 Results

The calculated rotational correlation times of the dye for the various systems are shown in Tab. 6.2. The experimental values are also given for comparison. Additionally, Fig. 6.1 shows the calculated anisotropy decay curves of Alexa488 in methanol and water (simulations 1 and 2) as solid lines (green: methanol and blue: water). The rotational correlation times are ϕ= 51 ps for the dye in water and ϕ= 86 ps in methanol (see also Tab. 6.2). Fig. 6.1 also shows an exponential fit to the measured time-resolved fluorescence anisotropies of Alexa488 in aqueous solution and in methanol at 300 K (dashed lines).[180] The measured rotational correlation times are 170 ps and 210 ps in water and methanol, respectively.

System	sim	exp
1) Alexa488 in methanol	86 ps	210 ps[180]
2) Alexa488 in water (SPC)	51 ps	170 ps[180]
3) Alexa488 in large box (SPC)	45 ps	170 ps[180]
4) Alexa488 modified charges (SPC)	60 ps	—
5) Rhodamine 6G in water (SPC)	89 ps	210 ps[185]
6) Alexa488 in water (SPC/E)	67 ps	170 ps[180]
7) Alexa488, no constraints (SPC)	52 ps	170 ps[180]
8) Tryptophan (SPC)	15 ps ($^{\mathrm{I}}\mathrm{L}_a$)	19 ps[186]

Table 6.2: Resulting rotational correlation times from the simulations compared to the corresponding experimental values.

As can be seen, the correlation times in the simulation are generally by about a factor of 3 smaller than in the experiment. This effect of too fast rotational diffusion of small molecules in simulations is well known and has already been discussed in the literature.[88,89,186] It was found that the rotational correlation time of tryptophan in SPC water is by about a factor of two smaller than the experimental value.[88] Furthermore, the self-diffusion coefficients of several solvent models have been shown to be too high.[187,188] Several reasons for this systematic deviation are conceivable: Inappropriate solvent or solute force field parameters, artefacts from the periodic boundaries, the use of bond-length constraints, inadequate treatment of electrostatic interactions, lack of unpolar hydrogens due to their implicit descrip-

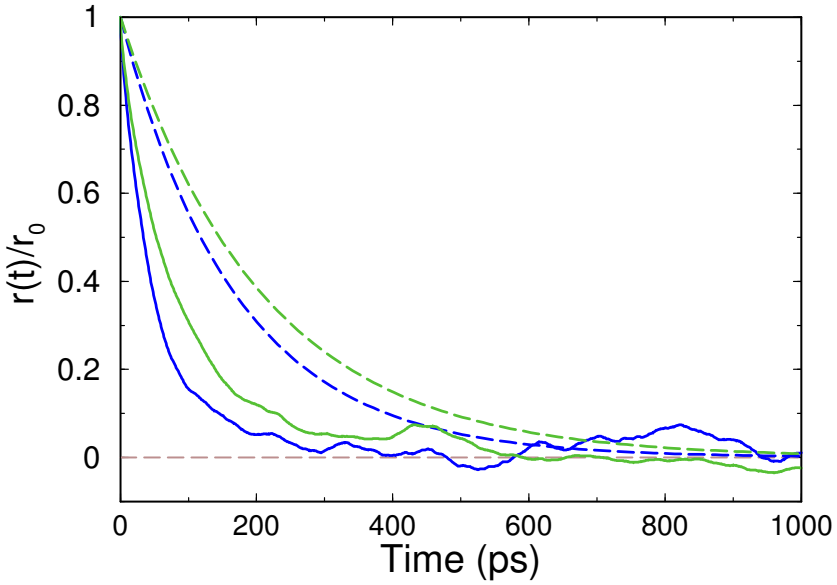

Figure 6.1: Anisotropy decay curves from the simulation (solid lines) and exponential fit curves to the experimental anisotropies (dashed lines) for the dye in water (blue) and methanol (green).

tion by *compound atoms*, and finally the coupling to a heat bath. In the following, we will discuss these possible reasons in more detail.

The solvent properties that mainly influence the rotational diffusion of a solute are supposed to be the viscosity, which is related to the self-diffusion coefficient, and the rotational diffusion of the solvent molecules, which is related to the dipole relaxation time. It has been found that the self-diffusion coefficient for the popular water models SPC, TIP3P, and TIP4P is too high and the dipole relaxation times are too small, indicating a too high mobility of the solvent molecules.[187]

To calculate the self-diffusion coefficient and the dipole relaxation time for the solvent force fields used in this work, we have carried out two 2 ns simulations of 327 methanol molecules and of 895 SPC-water molecules. The self-diffusion coefficients from the simulations are $4.2{\cdot}10^{-5}$ cm^2/s for water and $3.0{\cdot}10^{-5}$ cm^2/s for methanol, and the experimental values are $2.3{\cdot}10^{-5}$ cm^2/s and $2.4{\cdot}10^{-5}$ cm^2/s,[189] respectively. The dielectric relaxation time is 6 ps for water and 13 ps for methanol, whereas the experiment yields 9 ps and 56 ps,[190] respectively. That means, both the translational and the rotational diffusion of the solvent molecules are too fast in the simulation

compared to the experiment, which definitely contributed to the accelerated rotational diffusion of the dye in simulations 1 and 2.

Note, that the rotational correlation time of Alexa488 in SPC/E water (simulation 6) is only slightly higher than in SPC, although the physical properties of SPC/E (self-diffusion coefficient $2.8{\cdot}10^{-5}$ and dipole relaxation time 9.7 ps) agree very well to the experimental values.[187] Thus, the self-diffusion coefficient and the dipole relaxation time are apparently not solely responsible for the correct description of the rotational diffusion of the solute.

To address the question to which extent the rotational diffusion of the dye is sensitive to the dye force field parameters, first, we simulated in addition to the Alexa488 dye, also the rhodamine 6G dye[181] (simulation 5) and the fluorescent tryptophan (simulation 8), whose force field parameters were determined independently from each other in slightly different ways. For both dyes, the anisotropy was calculated and compared to experiment, as for Alexa488.

The calculated rotational correlation times for both excited states of tryptophan, $^{\mathrm{I}}\mathrm{L}_a$ and $^{\mathrm{I}}\mathrm{L}_b$, are 15 ps and 22 ps, respectively, and agree well to the values obtained from MD simulations by Daura et. al.,[88] 14 ps and 20 ps, respectively. For the tryptophan, the $^{\mathrm{I}}\mathrm{L}_a$ value, which is to be compared to the experiment,[88,184] is smaller than the experimental value of 19 ps.[186] For rhodamine 6G the calculated rotational correlation time of 89 ps is also smaller than the experimental value of 210 ps. The results show that also for these two molecules, the rotational diffusion is significantly faster than in the experiment.

The second test of the influence of the dye parameters on its rotational diffusion is done in simulation 4, where the partial charges of Alexa488 are modified, as described in Sec. 6.1. The resulting rotational correlation time of 60 ps indicates that an inaccuracy in the parametrization of the dye cannot explain the discrepancy between simulation and experiment. This is in agreement with Daura et. al., who found almost no sensitivity of the rotational correlation to the total charge of the solute molecule.[88]

Any possible artefacts due to the periodic boundaries should depend on the system size. The influence of the system size is tested in simulation 3. The resulting rotational correlation time of 45 ps does not indicate any effect of the system size and thus, the periodic boundaries are probably not responsible for the mismatch between simulation and experiment.

Simulation 7 tested the influence of using bond-length constraints, but here the result clearly shows no difference between the constrained (simulation

2) and unconstrained simulations (simulation 7). This is in agreement with Fuller and Rowley, who found that the effect of internal model flexibility is small even in polar fluids.[191]

The influence of the treatment of the electrostatic interactions is not studied here. However, it has actually been shown to affect the rotational diffusion, but the magnitude of this influence is much smaller than the effect of the solvent viscosity.[89] The question whether the rotational diffusion of the solute is effected by the method of temperature coupling has been studied by Daura et. al.[88] From simulations of tryptophan in SPC/E water with solute and solvent coupled jointly or individually to a heat bath, they could not observe a major difference.

The nonpolar hydrogens were implicitly described by using *compound atoms*, to ensure compatibility with the GROMACS protein force field, which will be used in the next chapter. The influence of this lack of unpolar hydrogens was not studied here, but is potentially not neglectable.

All these studied possible reasons for the discrepancy between the calculated and measured rotational correlation times suggest that the solvent force field parameters are mainly responsible for the too fast rotational diffusion of the dyes. Unfortunately, the details of this effect are unclear and need to be further investigated. Nevertheless, the differences between the simulation and the experiment for the considered observables are similar for the studied solvent force fields.

Comparison of Alexa488 in water and methanol

As can be seen in Tab. 6.2, the dye shows faster rotational diffusion in water than in methanol, in the simulation as well as in the experiment. This behavior was unexpected, since the viscosity of water [$1.002 \cdot 10^{-3}$ Pa s (at 293 K)] is larger than that of methanol [$0.587 \cdot 10^{-3}$ Pa s (at 293 K)]. Furthermore, the same experiment with fluorescein shows the *expected* behavior: the measured rotational correlation times are 140 ps in methanol and 170 ps in water.

To explain this inverse solvent effect for Alexa488 the structure of the dye in the simulation was analyzed in more detail. It has been observed that the extension of the dye, represented by the distance d defined in Fig. 6.2 A, strongly depends on the solvent. Fig. 6.2 C shows the extension of the dye during both simulations in methanol (green curve) and water (blue curve). The distance d fluctuates between 0.7 nm and 1.6 nm. The corresponding dye conformations are shown in Fig. 6.2 A and B for d=1.6 nm and d=0.7 nm, respectively. The headgroup of the dye is rather stiff, thus the

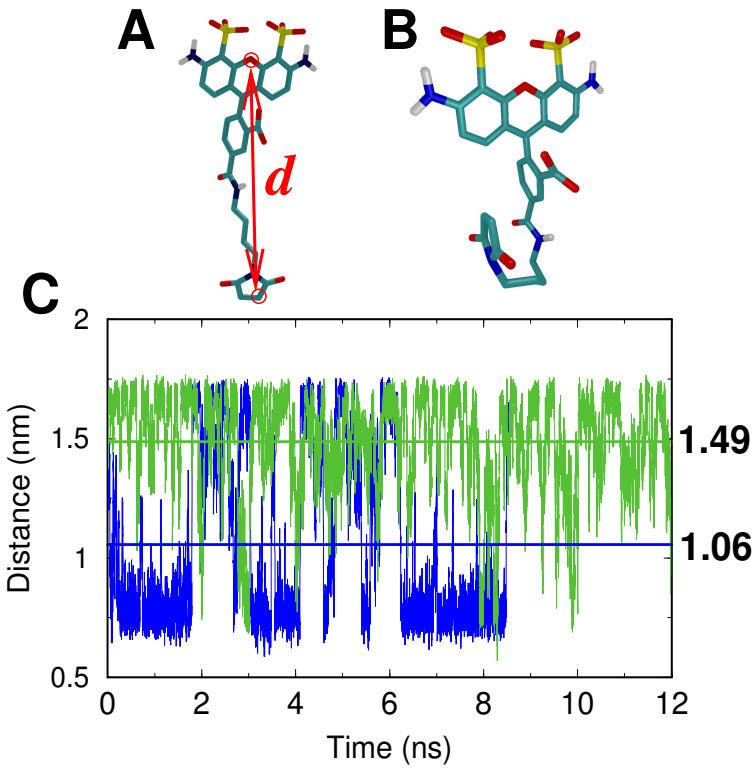

Figure 6.2: Extension of the Alexa488 dye molecule in water and in methanol. The extension is represented by the distance defined in (A) (red arrow). This distance is plotted versus time for the water (blue) and methanol (green) simulations. (B) shows the conformation, the dye mainly adopts in water, which corresponds to a distance of 0.8 nm.

change of the length is only due to the flexible linker, a hydrophobic chain. The average length of the dye in methanol of about 1.5 nm is clearly larger than that in water of about 1.0 nm (thick horizontal lines).

This observation can be interpreted as a minimization of the water exposed hydrophobic surface. In water, the hydrophobic chain minimizes the water accessible surface by coiling up and consequently reducing the effective size of the molecule, which can indeed explain the observed faster rotational diffusion of Alexa488 in water than in methanol. The lack of this flexible hydrophobic chain in fluorescein explains why in this case this inverse solvent effect was not observed.

Summary

In this chapter we studied the dynamics of free dyes, with a particular focus on the Alexa488 dye, which will also be used in the next chapter. The rotational diffusion of the dyes were found to be too fast compared with the experiments. Possible reasons for that were systematically addressed, from which we concluded that the solvent force field parameters were mainly responsible for the too fast rotational diffusion. Details of this effect are not fully understood yet and have to be further studied. Since the discrepancy between simulation and experiment for the considered observables is similar for the methanol and SPC force fields, we assume this deviation to be systematic. Thus, we assume the rotational correlation time of a dye in these methanol and SPC models to be generally too fast by a factor of about three. Nevertheless, the simulations of Alexa488 in methanol and water were able to describe an inverse solvent effect, which showed that MD simulations can really contribute to the interpretation of experimental results.

"If you walk, just walk. If you sit, just sit.
But whatever you do, don't wobble."
– Yunmen

7

Probing Protein Flexibility by Fluorescence Anisotropy

The mobility and dynamics of a protein-attached dye is influenced by the presence of the protein. This change of the dye dynamics is governed by the protein structure and dynamics and can be measured by fluorescence anisotropy experiments, such that the dye can be used as a probe. However, the information gained in these experiments is rather indirect and is usually interpreted in terms of models like the *wobbling-in-a-cone* model described above. Here, our aim is to gain *direct* insight into fluorescence anisotropy experiments using MD simulations that should provide interpretations of the experiments in atomic detail. We particularly address the question which processes influence the reorientational dynamics of the dye and therefore contribute to the observed anisotropy decay, and how to extract information on the protein conformational dynamics from the anisotropy decay curve. Additionally, we ask if and to what extent, vice versa, the attached dye affects the unperturbed protein dynamics, e. g., of flexible loops. Today, this pertubation is commonly and necessarily assumed to be negligible. The present study offers the chance to test this assumption. Here, we studied the Alexa488 dye (C5 maleimide, *Molecular Probes*) (cf. Fig. 4.2) attached to the loop connecting the helices A and B of bacteriorhodopsin (bR). To reduce the system size, only the AB fragment of bR was used for both the simulation and the experiment.

This chapter is organized as follows: First the simulation setup is presented and then, after a description of the used methods, the conformations of the dye on the protein surface are analyzed in Sec. 7.2.1. In Sec. 7.2.2

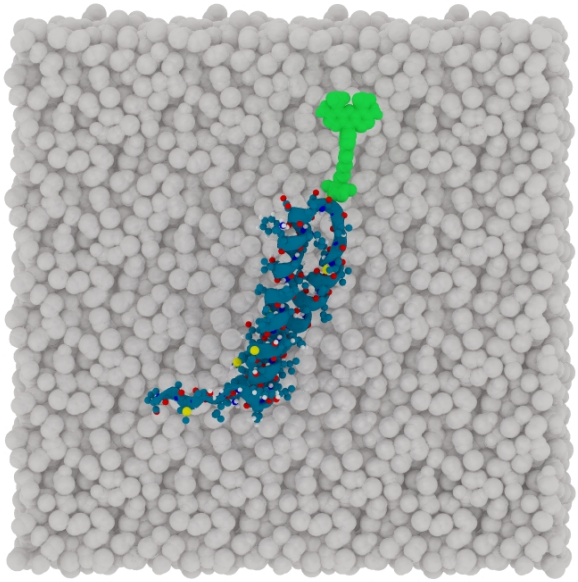

Figure 7.1: Simulation system setup. The AB-helix fragment of bacteriorhodopsin is shown in ribbon representation (blue) overlayed with a ball&stick representation of the protein atoms. The dye, shown in green, is attached to the loop connecting the two helices. The dye-protein system is solvated in methanol (grey spheres). The methanol box is cut out, to visualize the dye and the protein, embedded in the solvent.

the influence of the dye on the protein dynamics is studied by comparing simulations with and without attached dye. The correlations of the dye and protein motions are analyzed in detail in Sec. 7.2.3. In Sec. 7.2.4, the simulated anisotropy decay is compared to the experiment. Finally, we will discuss if and to what extent the *cone-in-a-cone* model, which was presented in Chap. 2, provides an appropriate description.

7.1 Methods

7.1.1 The simulation system

MD simulations of the Alexa488 dye covalently bound to position S35C of the AB-fragment (residues 8–71) of bacteriorhodopsin (bR) solvated in methanol were carried out. This system comprises 18752 methanol molecules, 64 amino acids, 52 dye atoms and two sodium ions, which summarizes to 56 933 atoms in total (see Fig. 7.1).

The simulation was performed using the GROMACS simulation software[116] with the united-atom GROMACS force field, describing non-polar hydrogens implicitly via compound atoms. The methanol parameters were taken from the GROMACS force field. The force field parameters for the dye had been determined as described in section 4.3. The system was energy minimized to obtain the starting configuration for the simulations. All simulation parameters were chosen as for the free dye simulations (cf. section 6.1), except for the coupling to the heat bath; here the solvent was separately coupled to a heat bath of 300 K with a relaxation time of 0.1 ps. It has been shown that the AB-fragment solvated in an organic solvent (methanol/chloroform (1:1)) adopts a conformation similar to its structure in bR,[192] therefore the initial structure of the AB-fragment of bacteriorhodopsin was taken from the crystal structure, PDB entry 1AP9.[193]

7.1.2 Probability distribution of the dye from a vacuum simulation

To sample the conformational space of the dye more efficiently than could be done in room temperature MD simulations with explicit solvent, vacuum simulations at 1000 K with implicit solvent were carried out. The central oxygen atom in the headgroup of the dye (cf. Fig. 4.2) was chosen to represent the conformation of the dye. The simulation then yields a probability distribution p_{high} of this atom at high temperature. Since the electrostatic interaction between the dye and the protein predominantly governs the conformational equilibrium of the dye, it was calculated for each conformation visited in the simulation. To account for the dielectric properties of the thin methanol layer between the dye and the protein, that was left out in the vacuum simulation, a dielectric coefficient of ϵ=10 was used for the calculation of the electrostatic energy.[194] A probability distribution was obtained from the Boltzmann-factor using the calculated electrostatic energy and correcting for the dye distribution p_{high} from the vacuum simulation,

$$p_{ele} \propto \exp[-(E_{ele} - k_B T_{high} \ln p_{high})/(k_B T_{room})] \tag{7.1}$$

with T_{high}=1200 K and T_{room}=300 K. On the one hand, the high temperature and the loss of friction due to the missing solvent drastically improve the sampling. On the other hand, the free energy landscape of the dye is perturbed due to the missing solvent. Therefore, the obtained probability distribution p_{ele} (of the central oxygen atom of the dye), which is shown in Fig. 7.3 C, is expected to yield only a rough estimate for the occupancy.

7.1.3 Correlation analysis

The component of the motion of the dye, that is correlated with the motion of the protein, was calculated using the LMLA-algorithm described in Chap. 5. This algorithm yields a specified number k of so called 'prototypic structures' (PS) in the conformational space, which are positioned along the largest extension of the molecular ensemble, thus capturing the main conformational changes of the system.

In this work, three PS (conformations of the dye-protein system) were calculated using the LMLA-algorithm, and the assignment (see Chap. 5) in Eq. 5.15 was done in the subspace of the dye. These three PS define a collective curvilinear coordinate in the complete configurational space, which is best correlated with the motion of the dye. These PS are called *subspace-correlated* PS, to distinguish them from the PS obtained from the conventional LMLA-algorithm, which are called *complete-space-correlated* PS. This curvilinear coordinate is a nonlinear combination of the cartesian coordinates of all atoms. The magnitude of each component in this nonlinear combination corresponds to its correlation with the motion of the dye (see Sec. 5.1.3). The magnitude of each component is quantified by the B-factor of each atom in the set of the three PS. The B-factors B_i are related to the mean square fluctuations $\langle \Delta \mathbf{r}_i^2 \rangle$ by

$$B_i = \frac{8\pi^2}{3} \langle \Delta r_i^2 \rangle \tag{7.2}$$

In this calculation, a high B-factor is due to a high correlation, but also due to a high flexibility of the atom itself. To account for this effect, the B-factors calculated from the set of the complete-space-correlated PS are substracted from those of the subspace-correlated PS. Thus, these corrected B-factors become independent of the flexibility of the atoms and therefore depend only on the correlation. Note that the obtained B-factors, although being directly related to the correlation of the protein with the dye motion, are only a relative measure for the correlation, i. e., they show which residue motion is more correlated than others, but no absolute correlation value is obtained.

7.1.4 Analysis of depolarization timescales

To study the characteristics of the fluorescence depolarization on different timescales, we calculated a position-dependent contribution to the depolarization of the dye. To this aim, we considered only five degrees of freedom

of the dye trajectory: the position of the dye $\mathbf{x}_m$, represented by the center of mass of the headgroup of the dye, and the normalized vector of the transition dipole moment $\boldsymbol{\mu}(t)$. To gain information on the timescale of the correlations, we compared the trajectory to a smoothed trajectory, where the fast fluctuations were averaged out and only the slow components of the dye dynamics remained. The smoothing, using a gaussian kernel with a standard deviation of σ= 40 ps, was done only for the transition dipole vector, whereas the positions of the dye remained unchanged. This yielded a trajectory $\boldsymbol{\mu}_s(t)$, where the slow components in the dynamics of the dye orientation are emphasized.

In Eq. 2.2 the anisotropy $r(t)$ is calculated as a time-average over $P_2(\boldsymbol{\mu}(t') \cdot \boldsymbol{\mu}(t'+t))$. The anisotropy $r(t)$ is therefore obtained, basically, by comparing all orientations of the dye with its orientations after time t. We define these contributions to the depolarization as

$$\xi(t') \equiv P_2\left[\boldsymbol{\mu}(t') \cdot \boldsymbol{\mu}(t'+t)\right] \quad , \tag{7.3}$$

where for the analysis in this work, a fixed time lag t= 50 ps was chosen. We now assign this contribution $\xi(t')$ to the position of the dye $\mathbf{x}_m(t')$ at time t'. This yields a function $\xi[\mathbf{x}_m(t')]$, which is a position-dependent measure for the mobility of the dye. A large value of ξ means, that the dye does not change its orientation much at $\mathbf{x}_m(t')$ compared to the time $t'+t$, whereas a low value indicates a jump between the orientation at $\mathbf{x}_m(t')$ and $\mathbf{x}_m(t'+t)$. Thus, a value corresponding to the orientational flexibility is assigned to each position of the dye. The same calculation is also done for the smoothed trajectory $\boldsymbol{\mu}_s(t)$, yielding the corresponding function $\xi_s[\mathbf{x}_m(t')]$.

7.1.5 Orientation distribution of the dye

The orientation distribution (cf. Fig. 7.11) is represented as a histogram on the surface of a sphere, built up of cones pointing towards the average center-of-mass position of the headgroup of the dye. An irregular grid consisting of 500 sample directions $\mathbf{g}_j$ was used to approximate the density of orientations p_j with a gaussian kernel

$$p_j = \sum_{i=1}^{n} \exp\left[|\mathbf{r}_i \cdot \mathbf{g}_j|/(2\sigma^2)\right] \quad , \tag{7.4}$$

using a variance $\sigma^2 = 0.025$, where n is the number of frames of the trajectory (10 000 for conformation A and 5000 for conformation B), and $\mathbf{r}_i$ is the

normalized transition dipole vector of the i-th frame. The lengths of the cones were chosen to be proportional to p_j and then scaled to lie in the range [5.0,8.5] Å.

All structures from the trajectory, used to calculate the orientation distribution of the dye in conformation A, were aligned to minimize the *rmsd* of the loop residues 30–42, since here, the local wobbling of the dye in the loop frame is of interest. For the distribution in conformation B the structures from the trajectory were aligned to minimize the *rmsd* of the helical residues.

7.1.6 Statistical error of MD from brownian dynamics

The rotational diffusion of a single normalized vector, which represents the transition dipole moment of the dye in the dye-cone, and the diffusion of the dye-cone in the protein-cone is calculated numerically. This yields a trajectory of the single vector, which is chosen to be of the same length as the MD-trajectory of the dye-protein system (16 ns). From this trajectory the anisotropy decay is calculated as described above (Sec. 2.1.1). The rotational diffusion coefficients and cone angles for the dye- and the protein-cone are chosen such that the parameters obtained by fitting the *cone-in-a-cone* model (Eq. 2.6) to the anisotropy are comparable to the parameters obtained from the MD-simulation in Tab. 7.1. Then 230 trajectories were calculated, which yields distributions of the *cone-in-a-cone* parameters, shown in Fig. 7.9. The diffusion coefficients used in this brownian dynamics simulation are 0.0018 ps^{-1} and 0.0003 ps^{-1} for the diffusion in the dye cone and in the protein cone, respectively. The (half-)cone angles are chosen 45^o and 50^o for the dye and protein cone, respectively.

7.2 Results

7.2.1 Dye conformations

Fig. 7.1 shows the simulation system. It includes the Alexa488 dye attached to the S35C position of the AB-helix fragment (residues 8–71) of Bacteriorhodopsin (bR) solvated by 18 752 methanol molecules and two sodium ions. The total simulation time was 26 ns. Fig. 7.2 shows the *rmsd* of the backbone atoms of the helical part (residues 10–29 and 43–61). The *rmsd* reaches a relatively low mean *rmsd* value of 0.12 nm after only about 20 ps, which indicates that the α-helical structure remains very stable during the

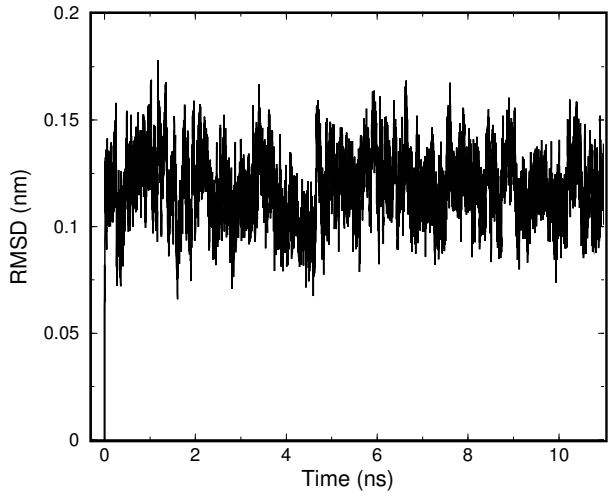

Figure 7.2: Root mean square deviation of the backbone atoms of the helical residues (residues 10–29 and 43–61) from the inital x-ray structure during the first 10 ns of the simulation.

simulation. Furthermore, the root mean square fluctuation (*rmsf*) around the mean structure of about 0.02 nm indicates a low flexibility of the protein. In contrast, the loop connecting the two helices shows a much higher flexibility than the helical regions, which is discussed further below.

The initial dye conformation was chosen to point away from the surface (cf. Fig. 7.1). After about 500 ps, the dye reaches conformation A, shown in Fig. 7.3 A, where it is loosely bound via non-covalent interactions to the protein surface for about 16 ns. It then detaches from the surface and flips back on the other side into conformation B, shown in Fig. 7.3 B, where it stays for the rest of the simulation. In conformation B, the dye is much less flexible than in conformation A, as quantified by the *rmsf* of 0.15 nm and 0.23 nm, respectively. In conformation A, the dye adopts two conformational substates, shown in Fig. 7.4 ('up' and 'down'). During the residence time of 16 ns in conformation A, the dye flips back and forth three times between these two substates, which significantly contributes to the mobility of the dye. This complex and hierarchical motion is apparently insufficiently described by a simple *wobbling-in-a-cone* model.

Unfortunately, the occupancy of conformations A and B cannot be inferred directly from the simulation, since only one transition was observed. As a substitute for this lack of reversibility, high temperature vacuum simulations were carried out to sample the conformational space of the dye more efficiently. From these simulations, a room temperature probability distribution p_{ele} of the dye positions based on the electrostatic dye-protein interactions

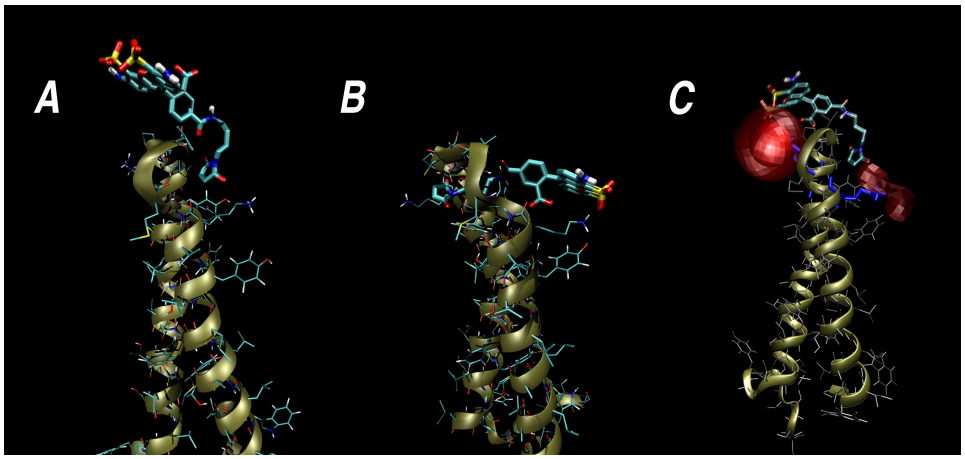

Figure 7.3: The protein from the side view with different dye conformations: The dye adopts conformation A (A) for the first 16 ns of the simulation and shows a relatively high mobility. In conformation B (B), where the dye remains for the rest of the simulation, the dye is much less flexible, since it is more tightly attached to the two helices. (C) The probability distribution of the dye calculated from the vacuum simulations is visualized by red isosurfaces on two different contour levels, which encloses 60% and 90% of the probability density (solid and transparent surfaces, respectively).

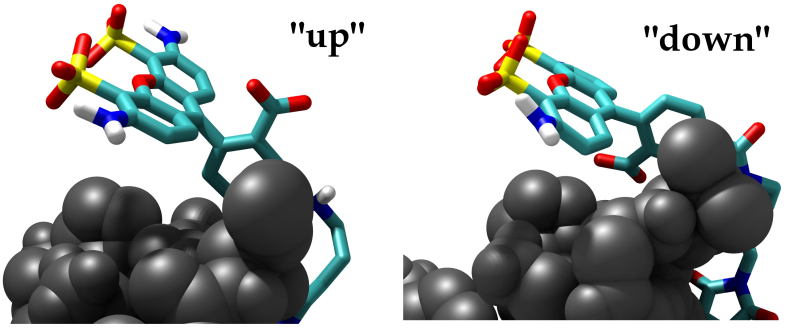

Figure 7.4: Transitions of the dye between two sub-conformations "up" and "down" of conformation A on the nanosecond timescale. The restriction of the mobility of the dye in both conformations is quite similar.

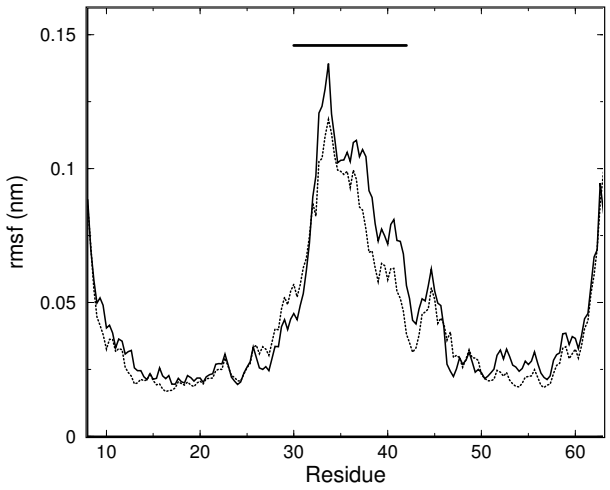

Figure 7.5: Root mean square fluctuations of the protein backbone for the protein with the bound dye (dotted line) and without the dye (solid line). The black bar on top denotes the loop residues.

is derived, as described in 7.1.2. Fig. 7.3 C shows an isosurface of p_{ele} in red. The electrostatic interaction between the dye and the protein suggests that the dye is more probable in conformation A than in B. In agreement with this result is the fact that the negatively charged dye has contact to two lysines (colored blue in Fig. 7.3 C) in conformation A, while it only has contact to one lysine in conformation B. In addition, as will be described further below, the calculated fluorescence anisotropy of the dye in conformation A agrees much better with the measured one than it does in conformation B. Thus, we assume the dye to be most of the time in conformation A.

7.2.2 Influence of the dye on the loop flexibility

To address the question if and to what extent the dye influences the protein conformation and dynamics, we compared the simulation described above to a 5 ns simulation of the same system *without* the dye. We focussed on the change in the flexibility of the protein and in particular of the loop region, where the dye is attached to. The flexibility is quantified by the root mean square fluctuation (*rmsf*) of the backbone, shown in Fig. 7.5 for the protein with bound dye (dotted line) and without the dye (solid line). For the calculation of the *rmsf* the trajectory was fitted onto a reference structure using only the helical residues. The black bar on the top denotes the loop

residues. As can be seen, the overall shape of both curves is quite similar. Only the loop residues show a slight decrease of the flexibility, while the rest of the protein is not affected by the dye. The dye decreases the *rmsf* of the loop residues only by about 15%. Therefore, the assumption that the dye does not influence the protein dynamics is, at least for this case, justified.

7.2.3 Dye-protein correlation

An important question, which can also be answered by simulation is, which region and which motion of the protein is actually probed by the dye. To this aim, we have to determine which mode of the observed protein dynamics correlates best with the dye motion. Fig. 7.6 shows the protein colored according to the correlation with the dye motion, calculated as described in 7.1.3 . The residues which are most correlated with the dye motion are shown in red, blue means no correlation. Note that the obtained measure of correlation is only relative, i. e., we only learn which atoms move more correlated than others. Although the dye is bound to residue 35, the neighboring residues 33 and 34 show the highest correlation, i. e., their flexibility is mainly probed by the experiment.

To see the reason for this, note that the fluorescence anisotropy is determined by the motion of the chromophore of the dye molecule, which is located in the headgroup of the dye. This headgroup is, as observed in the simulation, in close contact and bound via non-covalent interactions to the residues 33 and 34. We therefore conclude that this contact gives rise to the observed high correlation.

In constrast, residue 35 shows only a minor correlation, which was unexpected, since the dye is covalently bound to this residue. This finding suggests that the relatively long flexible linker of the dye impedes any correlation. From this we conclude that influences of the dye on the protein dynamics are mainly due to non-covalent interactions between the headgroup of the dye and the protein.

The region around residue 45, in the middle part of the left helix (helix B) in Fig. 7.6, also shows a significant correlation, whereas the helix connecting this residue to the loop does not show any correlation. This could be explained by assuming that small conformational changes of the helix lead to larger conformational changes of residue 45. Then, the motion of the helix, that is supposed to be correlated with the dye motion, but such small in amplitude that it cannot be identified in Fig. 7.6, could strongly affect the

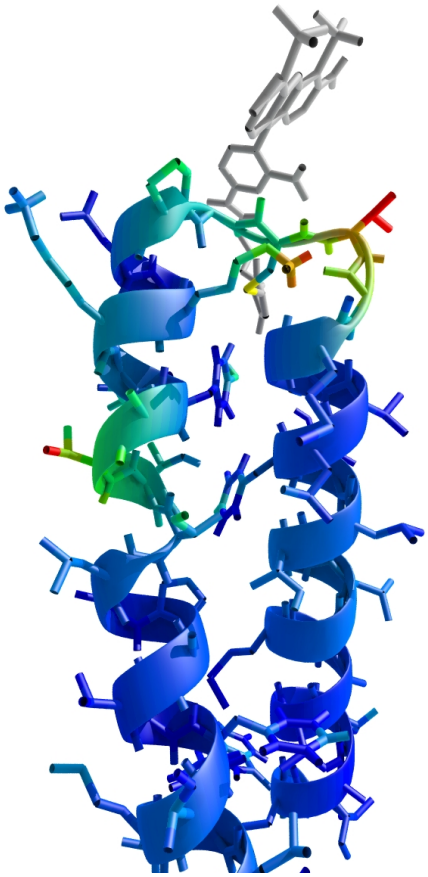

Figure 7.6: The Alexa488 dye attached to the loop in the protein. The protein is colored according to the relative correlation of its motion with the motion of the dye. Atoms showing a high (low) correlation are shown in red (blue).

motion of residue 45. In this way, a small correlation of the helix could be amplified in residue 45.

One goal of this work, as mentioned above, was to identify the component in the anisotropy decay $r(t)$ of the dye that is governed by and, hence, yields information about, the protein dynamics. This component in the anisotropy is due to a component of the dye motion which is correlated with the protein motion. In general, there are different possible processes leading to a depolarization of the fluorescence. Typical sources of depolarization are, e. g., the fast wobbling of the dye with a rotational correlation time of 100-300 ps and the overall tumbling motion of the whole protein including the attached dye with a rotational correlation of a few nanoseconds up to

the microsecond range, depending on the size of the protein.

The timescale of the dye-protein correlation shall be studied in the following, to be able to identify the component in the anisotropy decay that is due to the interaction of the dye with the protein. To this aim, we calculated, as described in Sec. 7.1.4, a position-dependent contribution to the depolarization of the dye for the original trajectory $\xi[\mathbf{x}_m(t')]$, as well as for a smoothed trajectory $\xi_s[\mathbf{x}_m(t')]$, where the fast fluctuations are averaged out and only the slow components of the dynamics remain. Fig. 7.7 shows the result of this calculation. A bottom-up viewing direction is chosen to overlay the dye positions with the loop residues. The ribbon in the foreground depicts the backbone of the protein, colored according to the correlation with the dye motion, as in Fig. 7.6. The "cloud" in the background is built up by all the positions the center of mass of the dye visited during the simulation. This "cloud" in Fig. 7.7 A is colored according to the contribution to the depolarization $\xi(\mathbf{x}_m)$. A red colored region represents a high depolarization, which can be interpreted as a high mobility of the dye at this position. A blue color indicates a low mobility of the dye. The coloring of the position distribution in Fig. 7.7 B is calculated using the smoothed trajectory and therefore shows the contribution to the depolarization of the slow components of the dye dynamics $\xi_s(\mathbf{x}_m)$.

As can be seen, for the original trajectory (Fig. 7.7 A), the mobility of the dye is rather position-independent. In contrast, for the smoothed trajectory, the mobility becomes position-dependent, and the dye shows a higher mobility in the vicinity of those residues of the protein which were previously identified to correlate most with the motion of the dye (red colored backbone). In summary, there is a slow component of the dye dynamics, which is correlated with the protein motion.

To estimate which rotational correlation time corresponds to this slow component, the anisotropy decays for both trajectories were calculated. The fastest decay times from the original and from the smoothed trajectory are 120 ps and 300 ps, respectively. The 120 ps decay time presumably originates from the fast local wobbling of the dye. Apparently, this fast wobbling motion is suppressed in the smoothed trajectory. The slow depolarizing component which is correlated with the protein motion and therefore contains information on the protein dynamics, therefore must correspond to a correlation time of at least 300 ps, or larger, as described in the next section.

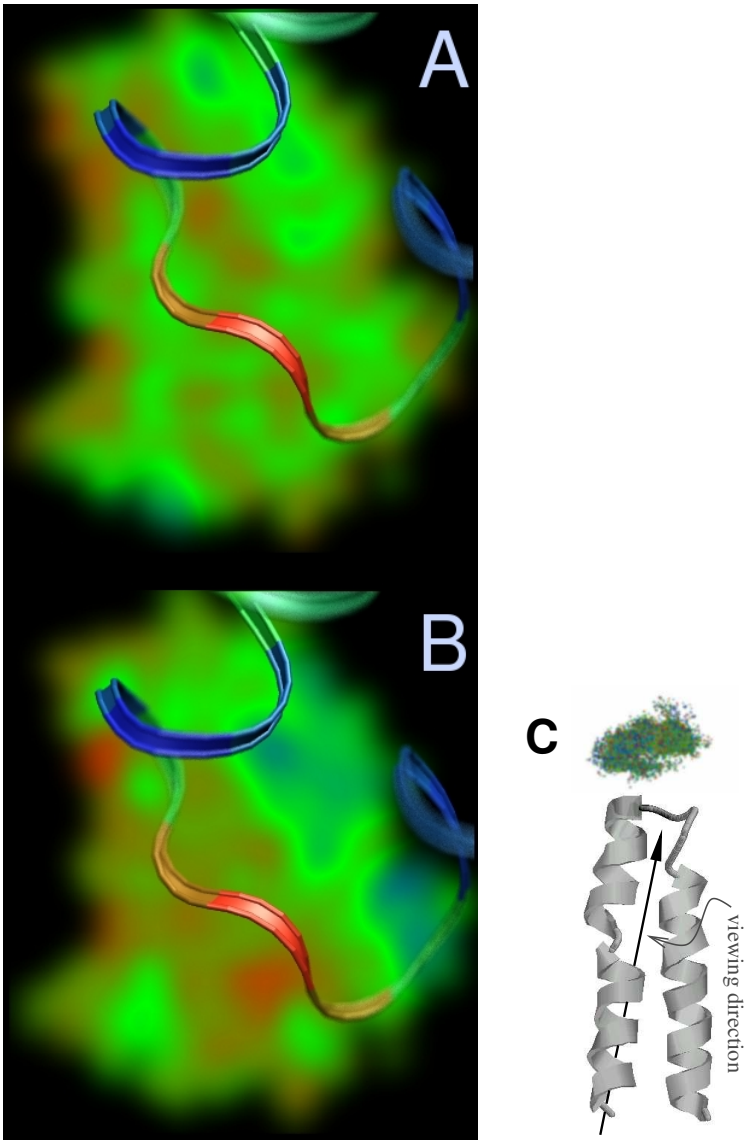

Figure 7.7: Analysis of the timescale of the dye-protein correlation. (A) and (B) The ribbon in the foreground depicts the protein backbone looking from the bottom to the top of the protein, as illustrated in (C). The backbone is colored according to the correlation of the protein motion with the dye motion, as in Fig. 7.6. The 'cloud' in the background shows all the positions that are visited by the dye during the simulation. The coloring of this 'cloud' is according to the contribution to the depolarization, i. e., basically the mobility. Red, green and blue indicate high, mid and low mobility of the dye at a certain region. This coloring is calculated from the original (A), and from a smoothed (B) trajectory.

7.2.4 Comparison of simulation and experiment

In the measured fluorescence depolarization curve, shown in Fig. 7.8 A (solid line), three decay components can be resolved: two faster components (about 300 ps and 800 ps) are assumed to arise from the dye motion relative to the protein, where the slower of these two components is suggested to be influenced by the local protein flexibility, as described in the previous section. The third (slowest) component of about 5 ns originates from the overall tumbling motion of the protein-dye complex.

The three decay components are difficult to spot in Fig. 7.8 A, since this curve is convoluted with the instrument response function (IRF) with a full width at half maximum (FWHM) of 48 ps. The inset in Fig. 7.8 A shows a logarithmic plot of the anisotropy decay. From the fit to the experimental curve using a sum of three exponentials (Eq. 2.7) convoluted with the IRF as the model function, we obtained the parameters shown in Tab. 7.1 ('fit 2'). Since a fit of the *cone-in-a-cone* model (Eq. 2.6), convoluted with the IRF, to the experimental curve was not available, the *cone-in-a-cone* parameters in Tab. 7.1 ('fit 1') are derived by fitting Eq. 2.6 to the three-exponential fit curve. The parameters A_1 and A_2 are related to the (half-)cone angles of the dye- and the protein-cone, respectively (cf. Eq. 2.4), and are given in parenthesis.

Fit 1	exp	sim	Fit 2	exp	sim
A_1	0.35 (46^o)	0.39 (44^o)	B_1	0.221	0.22
ϕ_1	0.35 ns	0.12 ns	φ_1	0.296 ns	0.12 ns
A_2	0.15 (59^o)	0.32 (47^o)	B_2	0.097	0.17
ϕ_2	1.65 ns	0.98 ns	φ_2	0.805 ns	0.82 ns
ϕ_G	5.0 ns	∞	B_3	0.0065	0.01
			φ_3	5.0 ns	∞

Table 7.1: Results from the fits to the experimental and simulated anisotropy decays (cf. Fig. 7.8) using two different model functions. Fit 1 uses Eq. 2.6 which describes the *cone-in-a-cone* model. Values in parenthesis are the corresponding cone angles (Eq. 2.4). Fit 2 uses a sum of three exponentials as in Eq. 2.7.

As discussed above, for the comparison of the anisotropy, the first part of the simulation (about 16 ns), where the dye is in conformation A, was used to calculate the anisotropy decay curve, as described in Sec. 2.1.1. For direct comparison with the experiment, first, the three-exponential model Eq. 2.7 was fitted to the simulated anisotropy (Fig. 7.8 B). The obtained parameters are shown in Tab. 7.1 ('fit 2'). Since we are here not interested in the overall

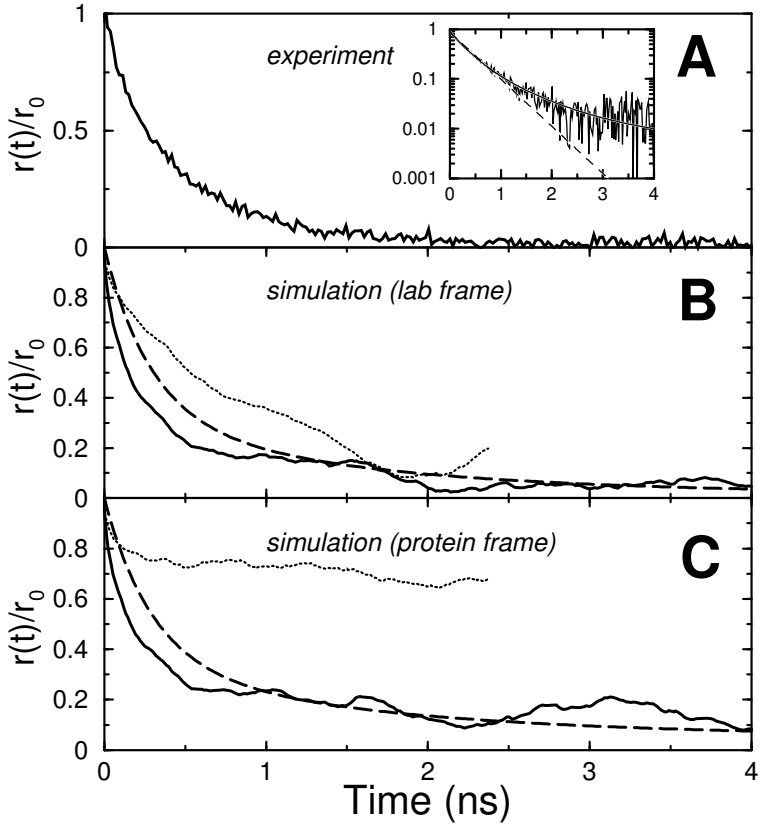

Figure 7.8: Anisotropy decays from the simulation and from the experiment, normalized by the initial anisotropy r_0. (**A**) Measured anisotropy decay (convoluted with the instrument response function). (**B**) Anisotropy of the dye in conformation A (solid line) and in conformation B (dotted line) calculated from the original trajectory. The dashed line shows the fit curve of the *cone-in-a-cone* model to the experimental curve, using the parameters shown in Tab. 7.1. (**C**) Anisotropy in the protein frame for conformation A (solid line) and for conformation B (dotted line), calculated from a trajectory, that is fitted onto a reference structure. The dashed line shows the same fit curve as in (**B**), except for the global rotational correlation time ϕ_G, which is here set to infinity. This curve thus corresponds to the measured anisotropy in the protein frame.

tumbling motion of the dye-protein system, this motion is suppressed by fitting all protein structures to a reference structure. The elimination of the overall tumbling allows for a fit using two instead of three exponentials (cf. Eq. 2.6), which improves the quality of the parameters for the local motion of the dye. The anisotropy calculated from the fitted trajectory was then fitted by Eq. 2.6, setting $\phi_G = \infty$, which corresponds to an infinitely slow global rotational diffusion. The result of this fit is also shown in Tab. 7.1 ('fit 1'). Assuming the local dye motion and the global motion of the protein to be independent, both fits, to the measured and to the simulated anisotropy, should yield the same parameters for the local dye motion, described by the parameters A_1, ϕ_1, A_2, and ϕ_2 from Eq. 2.6.

The agreement of the fit parameters between experiment and simulation is quite good, although again, the rotational diffusion in the simulation is too fast. The speedup in the rotational diffusion in the simulation here is comparable to what was observed for the free dye (cf. Chap. 6). The amplitude A_1, which describes the dye-cone angle in the *cone-in-a-cone* model, matches quite well, indicating the simulation to accurately describe the local wobbling of the dye. The second amplitude A_2 will be analyzed in more detail in the next section.

The correlation time ϕ_G of the global rotation of the protein was obtained from the experiment, as described above, by fitting the *cone-in-a-cone* model (Eq. 2.6) to the anisotropy decay. The simulation offers the chance to calculate the correlation time ϕ_G directly from the rotational diffusion coefficient D_G of the protein by $\phi_G = 1/(6D_G)$. The rotational diffusion coefficient was calculated from the simulation $D_G = 4.3 \cdot 10^{-5}\text{ps}^{-1}$, which yields ϕ_G=3.9 ns. This is in good agreement with the experimental value of 5 ns.

7.2.5 Analysis of the statistical error

The limited length of the calculated trajectory (16 ns), which means a limited statistics in the sampling of the configurational space, causes a statistical error in the calculation of the anisotropy decay and therefore also in the fitted *cone-in-a-cone* parameters.[195]

The straightforward approach to estimate this error would be to calculate many similar MD-trajectories and to analyze the variance of the parameters obtained from each of the trajectories. Unfortunately, the calculation of just one trajectory required already four months of computer time, so this option is not practicable.

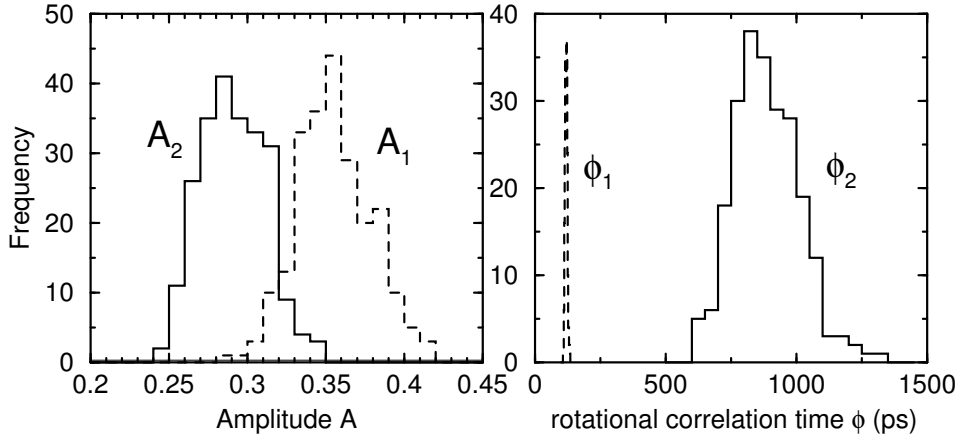

Figure 7.9: Distribution of amplitudes A_1 and A_2 and rotational correlation times ϕ_1 and ϕ_2 from the Brownian dynamics simulations in the *cone-in-a-cone* model to estimate the statistical errors of the obtained parameters.

Instead, we used an approach, which is principally the same as the one suggested before, but differs in the way the trajectories are obtained. Assuming the dye dynamics be sufficiently well described by the *cone-in-a-cone* model, the MD simulation is substituted by a brownian dynamics simulation of the transition dipole moment diffusion in the *cone-in-a-cone* model (cf. Sec. 2.1), as described in Sec. 7.1.6. From this brownian dynamics simulations 230 trajectories were obtained, from which anisotropy decay curves were calculated. These were then fitted by Eq. 2.6 yielding parameters from which histograms are calculated and plotted in Fig. 7.9. The variances of the obtained parameters are measures for their statistical errors.

The error of the fast rotational correlation time $\Delta\phi_1$= 5 ps is significantly smaller than of the slow correlation time $\Delta\phi_2$= 120 ps, since the larger rotational diffusion coefficient leads to a better sampling of the dye-cone. From the errors of the amplitudes $\Delta A_1 = \Delta A_2 = 0.02$ errors of the cone angles $\Delta\theta_1 = \Delta\theta_2 = 1^o$ can be derived.

7.2.6 Anisotropy within the loop frame

The two faster components of the anisotropy decay were above assumed to be due to the local wobbling of the dye on the surface of the protein. It was further assumed that the slower one of these two components is due to the flexibility of the loop to which the dye is attached.

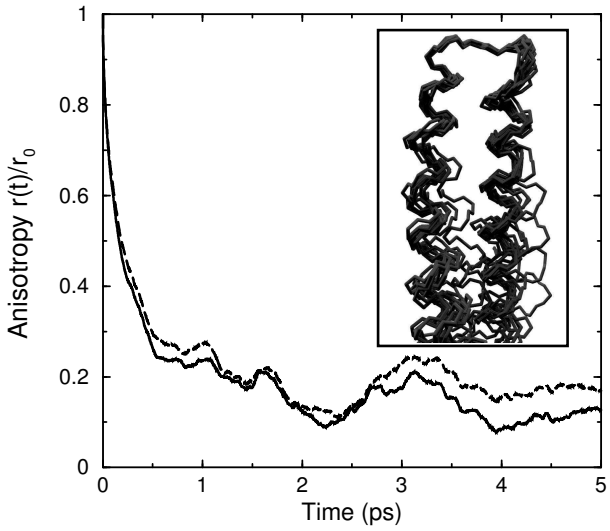

Figure 7.10: Calculated anisotropy of the dye in conformation A in the protein frame (solid line) and in the loop frame (dashed line). The inset shows several superimposed protein structures from the trajectory, which is fitted onto the loop residues.

These assumptions are tested now by determining the motion of the dye relative to the loop. To this aim, the dye coordinates are transformed into the coordinate frame of the loop. This is achieved by fitting all structures from the trajectory onto the loop residues (30–42). The inset in Fig. 7.10 shows several protein snapshots of this fitted trajectory. This trajectory describes the motion of the dye, that is uncorrelated with the motion of the loop.

The anisotropy decay within the loop frame $r_{\mathrm{loop}}(t)$ calculated from this trajectory is shown in Fig. 7.10 (dotted line). The solid line shows the simulated anisotropy decay in the protein frame $r_{\mathrm{protein}}(t)$ (the trajectory was fitted onto the entire protein-dye system), which is the same curve as the solid line in Fig. 7.8 C. The difference between both curves describes the reorientation of the dye due to the loop flexibility. The influence of the loop flexibility is expressed in terms of a decay component in the framework of the *cone-in-a-cone* model by

$$r_{\mathrm{protein}} \approx r_{\mathrm{loop}}(t) \left[(1 - A_{\mathrm{loop}}) e^{-t/\phi_{\mathrm{loop}}} + A_{\mathrm{loop}} \right] . \qquad (7.5)$$

If we assume $r_{\mathrm{protein}}(t)$ to contain only two decay terms, namely a *wobbling-in-a-cone* motion of the dye and the loop flexibility, then $r_{\mathrm{loop}}(t)$ will contain

only one decay term. Equation 7.5 then directly corresponds to Eq. 2.6 from the *cone-in-a-cone* model. That means, A_{loop} and ϕ_{loop} should be the same as A_2 and ϕ_2.

Fitting Eq. 7.5 to the anisotropy in the protein frame $r_{\text{protein}}(t)$ yields A_{loop}=0.77 and ϕ_{loop}=1370 ps. The obtained A_{loop} parameter corresponds in the *wobbling-in-a-cone* model to a cone angle of 23^o (Eq. 2.4), which is in good agreement with the flexibility of the loop observed in the simulation, as will be discussed in the next section.

Obviously, A_{loop} is significantly larger than A_2, which means the corresponding cone angle is smaller, which indicates that there must be a second decay component in the anisotropy in the loop frame $r_{\text{loop}}(t)$, that additionally contributes to the depolarization and which depolarizes on a similar timescale of about one nanosecond.

Since we observed a flipping of the dye between two orientations in conformation A, as described above (cf. Fig. 7.4), which occurs roughly on the same time scale as the slow correlation times ϕ_{loop} and ϕ_2 compared to 3 orientation flip events in 10 ns, we propose this process to be the missing additional component in the anisotropy decay. That means, that there are *two* processes, contributing to the slow component in the experimental anisotropy, the flipping of the dye orientation and the loop flexibility, which both occur at the same timescale and therefore cannot be resolved experimentally. The straight use of the *cone-in-a-cone* model to interpret the measured anisotropy decay would therefore overestimate the absolute amplitude of the loop flexibility. Nevertheless, *changes* in the loop flexibility are observable in the experiment, as has been shown by Alexiev and co-workers.[13]

7.2.7 Orientation distribution of the dye

To test whether the *wobbling-in-a-cone* model is appropriate to describe the local wobbling of the dye, the orientation distribution of the transition dipole moment was calculated for both dye conformations (A and B). Fig. 7.11 shows this orientation distribution, calculated as described in 7.1.5. The distributions both are centered around one distinct maximum, and the width of the distribution in conformation A is broader than in conformation B. The shape of both distributions is well approximated by a gaussian distribution. The standard deviation of the distribution in conformation A and B is 44^o and 29^o, respectively. The cone angles obtained by a strict

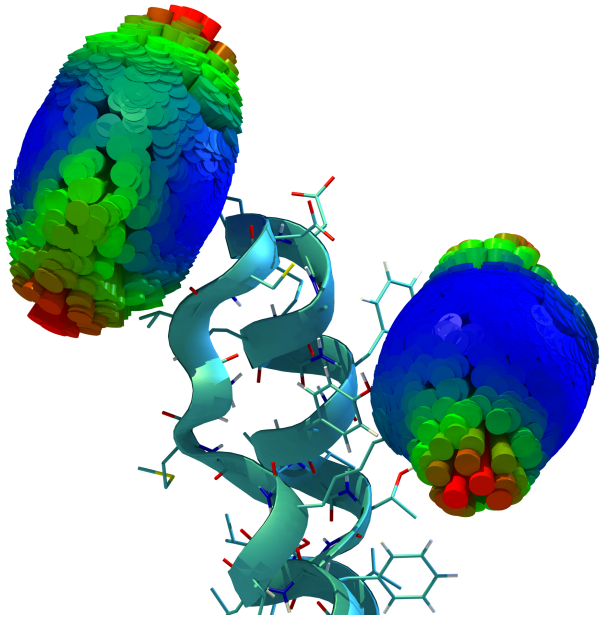

Figure 7.11: Orientation distribution of the transition dipole moment of the dye in both conformations A and B, represented by cones placed at the surface of a sphere. The color and the length of the cones denote the frequency that the transition dipole moment adopts a certain direction (red, green, and blue means high, mid, and low frequency, respectively). The two orientation distributions are centered at the mean position of the dye in conformation A or B, respectively.

wobbling-in-a-cone analysis of the simulated anisotropy, 40^o and 27^o for A and B, respectively, agree very well to the effective widths of the gaussian distributions, as has already been shown.[196] That means the cone angle from the *wobbling-in-a-cone* model is able to describe the fast local wobbling of the dye, if interpreted as the effective width of a gaussian distribution.

7.3 Summary of dye/protein simulations

The simulation of the Alexa488 dye attached to the loop within the AB-fragment of bacteriorhodopsin revealed two possible conformations of the dye. It was observed that the mobility of the dye differs significantly in these two conformations. Additionally, the anisotropy determined from the simulation in one of the two conformations (B) (cf. 7.3) is much higher than the measured anisotropy. In contrast, the anisotropy in the other

conformation (A) agrees very well with the experiment. Since the sampling of the electrostatic energy surface (Sec. 7.1.2) shows that conformation A is electrostatically favored, and, therefore, more populated, we obtained good agreement with the experimental data.

By comparison with a second simulation of the protein without the dye, the influence of the dye on the flexibility of the protein was determined and found to be small, because as the only significant difference between the fluctuations of the protein backbone a small decrease ($\approx 15\%$) of the loop flexibility due to the bound dye was observed. This finding supports the usual assumption made in the experiments that the dye does not influence the protein.

To study, vice versa, how the dye motion is influenced by the protein, correlations between the dye and the protein motion were analyzed in more detail, using the LMLA-algorithm (Chap. 5). This calculation revealed those residues that affect the dye motion and which are therefore mainly probed in the experiment. In contrast to what one might intuitively assume, we found that the dye motion does not primarily probe the residue to which it is covalently bound. Instead, it interacts mainly with the two neighboring residues, 33 and 34. This information is crucial for the interpretation of the experiment and cannot be inferred from the expriment alone.

Overall, the agreement between the calculated and the measured anisotropy is very good. The calculated anisotropy in the protein frame shows two decay components of 120 ps and 980 ps. The first decay time is attributed to the rotational diffusion of the dye in the solvent (methanol), although it is slightly slower than for the free dye (86 ps), presumably due to the attachment of the protein. Indeed, such increase of the rotational correlation time of the bound compared to the free dye also has been observed in the experiment (210 ps compared to 300 ps). This effect is probably due to the decreased solvent diffusion in the vicinity of the protein.[144,145,197] Furthermore, the corresponding cone angle from the *cone-in-a-cone* model matches to the cone angle observed within the (moving) coordinate frame of the loop. From this, we conclude that this fast component is due to the local wobbling of the dye.

The second component of 980 ps has initially been attributed to the protein flexibility. This assumption is supported by the analysis of the timescale of the dye-protein correlation, which found, that the motion of the dye, that is due to the protein flexibility, leads to an additional decay component in the anisotropy with a rotational correlation time larger than 300 ps. To

further test this assumption, the anisotropy was calculated in the coordinate frame of the loop, which allowed to directly assess the influence of the loop flexibility onto the anisotropy. It was found that the decay time of the depolarization induced by the loop (1370 ps) is indeed close to the second component, but its amplitude is much smaller compared to the second component and, therefore, the application of the *cone-in-a-cone* model overestimates the corresponding cone angle. Thus, the loop motion alone is obviously not sufficient to account for the second component. Rather, we propose the missing additional contribution to the depolarization on the timescale of about one nanosecond to be due to the transition of the dye between the two conformational substates ("up" and "down"), which we observed in our simulations. Thus, by differentiation between these two processes (the loop flexibility and the conformational transition of the dye) we provided an interpretation that is not accessible by the experiment alone.

Our assignment of the decay components is partly consistent with the assignment in the experiment, which was done by studying different fragments of bR.[13] Besides the free dye, the fluorescence anisotropies of the dye bound to only the loop-fragment, to the AB-fragment, and finally to the complete bR in micelles have been measured. The fast component (local wobbling of the dye) was the same in all experiments, whereas the slowest component (overall tumbling) increased with the size of the fragment. For the AB-fragment and the whole bR, an intermediate decay component was identified, which had been assigned solely to the loop flexibility. Our results showed that this assignment overlooked the contribution from the slow conformational dynamics of the dye. Further evidence for the rotational correlation time of the loop determined in this work and in the fluorescence anisotropy experiment is provided by NMR experiments, which show that the backbone N–H vectors are involved in an intermediate motion close to 1 ns.[198]

So the main result is that actually two processes, loop flexibility *and* conformational dye dynamics, contribute to the measured depolarization decay at the timescale of one nanosecond. Straightforward application of the conventional *cone-in-a-cone* model to the anisotropy decay curve, therefore, overestimates the cone angle of the protein cone, i. e., the loop flexibility. In contrast, we have shown that *cone-in-a-cone* model provides an appropriate and accurate description of the local wobbling of the dye.

"Science... never solves a problem without creating ten more."
– George Bernard Shaw

8

Summary & Discussion

Fluorescence spectroscopy in combination with site-directed fluorescent labeling of biomolecules like proteins or DNA has become a standard tool in molecular biology and biochemistry to study interactions and conformational dynamics of biomolecules. The goal of this work was to contribute, by analysis and simulation, to the structural interpretation of fluorescence anisotropy and fluorescence resonance energy transfer (FRET) experiments at the atomic level. To this aim, different approaches have been developed and applied.

The first step aimed at improving the analysis of single-molecule FRET experiments. Single-molecule methods demand for analysis methods which particularly account for the typically small number of observed events. In the case of single-molecule FRET experiments, the challenge is to determine the distance and distance fluctuations of two fluorescent dyes from only few detected photons.

The usual approach employs window averaging to obtain a time-dependent intensity from the measured single photons. Here the problem is the uncontrolled statistical error due to shot noise, thus, the error bars of the obtained distance trajectory are unknown. In addition, we have shown, however, that this method introduces an artificial bias to the obtained distance, since it assumes a uniform a priori probability of *intensities*.

To overcome these problems, we developed a maximum-likelihood theory to reconstruct distance trajectories from single molecule FRET experiments. This yields a probability distribution of trajectories, from which rigorous error bounds are obtained. In contrast to the conventional window averag-

ing, our theory assumes a uniform a priori probability of *distances*, which in the absence of knowledge on distances before the measurement is the most reasonable assumption.

In this work the focus was on FRET experiments that measure inter- or intra-biomolecular distances. Since the underlying motions, such as translational motions or conformational dynamics of proteins, are known to be diffusive in nature, this extra knowledge is explicitly used in our approach. This allows, in addition, to determine an effective diffusion coefficient for the distance changes observed in the experiment.

Several extensions of this method are conceivable, which, e. g., concern position- and dye-dependent detection efficiencies. Furthermore, if there is already some knowledge about the free energy landscape along the dye-distance affecting conformational motion, it should be possible to include this knowledge into our theory to further enhance the distance determination. The quite recent development of two-step FRET allows to measure the distances between *three* dyes simultaneously,[199] and thus to probe correlations of conformational dynamics.[200] An extension of our theory to also describe this two-step FRET method is straightforward and would be a valuable contribution to the interpretation of such experiments. Finally, since low count rates are a notorious problem of single molecule experiments in general, we expect our approach to be of wide applicability.

In a second contribution, we developed the LMLA-algorithm to compute principal curvilinear coordinates from molecular ensembles. To extract main structural features of an ensemble generated by a molecular dynamics (MD) simulation, the common approach is the principal component analysis (PCA), which yields a *linear* principal coordinate. However, in this work, as in many other situations, the motion of interest was the *rotational* dynamics of a fluorescent dye, which demanded for principal *curvilinear* coordinates. Whereas the principal coordinate obtained by the PCA maximizes the variance of the ensemble along this coordinate, we could show that our approach maximizes a *generalized* variance. This method should also be particularly useful for describing conformational changes in proteins, particularly (nonlinear) rotations of domains around hinge axes or dihedral angles of the protein backbone.

We tested our method using an artificial two dimensional distribution as well as an ensemble of structures of the protein BPTI (bovine pancreatic trypsin inhibitor), generated by the CONCOORD program. We found that the LMLA-algorithm indeed yields a coordinate which describes the ensem-

bles in these test examples significantly better than the conventional PCA. Furthermore, and as an extension to the usual PCA, our method allows to answer the question which mode of motion is best correlated with the motion in a specified conformational subspace. It is this feature which could be used to great advantage in this work to study the dye-protein correlation, and which should be applicable to a wide range of processes like, e. g., protein-protein or protein-ligand interactions.

The third contribution concerned MD simulations of fluorescence anisotropy decay experiments. First, simulations of two different free dyes (Alexa488 and rhodamine 6G) in methanol and in different water models were performed, to show the general ability of MD simulations to calculate the fluorescence anisotropy of a dye and to test the used solvent and dye force field parameters.

These initial tests provided promising evidence, that it is indeed possible to calculate the fluorescence anisotropy of a dye from MD simulations. We were actually able to predict an unexpected inverse solvent effect of the Alexa488 dye: The dye shows a larger rotational correlation time in methanol than in water, although the viscosity of methanol is smaller than that of water. This finding has been confirmed by experiment. Additionally, the MD simulations revealed this effect to be due to the conformational dynamics of the hydrophobic chain in the Alexa488 dye. This example showed that MD simulations can actually help to interpret fluorescence spectroscopy experiments.

In all simulations, the dyes had rotational correlation times that were too small as compared with the experiment, i. e., the rotational diffusion in the simulation is systematically too fast, at least for such small molecules as the dyes under investigation. In contrast, the discrepancy of the rotational correlation times of a protein between simulation and experiment is much smaller. Test simulations suggested that this discrepancy is due to inaccuracies in the solvent force field. The details of this effect are not fully understood, however, and need further investigations.

In close collaboration with U. Alexiev (FU Berlin), who performed the fluorescence anisotropy experiments, we performed simulations of the Alexa488 dye attached to the loop region of the AB-fragment of bacteriorhodopsin solvated in methanol. From these simulations, we were able to determine the conformation of the dye, which was a prerequisite for all following investigations.

A critical assumption for the interpretation of fluorescence spectroscopy ex-

periments is, that the fluorescent label does not influence the dynamics of the protein. This assumption is necessary, but difficult to test. By comparison of the backbone root mean square fluctuations from a simulation of the AB-fragment with and without the dye, we found that the dye changes the dynamics of the protein by approximately 15%, such that the abovementioned assumption, at least for the case at hand, seems justified.

By using our LMLA-algorithm to study the dye-protein correlations, we were able to identify those residues, whose motion is correlated with the dye motion and which are therefore probed in the experiment.

From the MD simulations, the anisotropy of the dye has been calculated and compared to the measured anisotropy. The overall agreement between the calculated and the measured anisotropy decay curves is very good. Our analysis of the anisotropy decay revealed three decay times, which are due to different motions, the fast wobbling of the dye, which is uncorrelated with the protein motion, an intermediate component, which was shown to be partially correlated with the protein motion, and the overall tumbling motion of the protein. The intermediate component in the anisotropy decay, that initially was assumed to be due to the loop flexibility, was shown to be also due to slow conformational dynamics of the dye. An important consequence is that the straightforward application of the *cone-in-a-cone* model would overestimate the amplitude of the loop motion considerably.

If it is that difficult to investigate protein dynamics by fluorescence spectroscopy, why not using different techniques like, e. g., NMR spectroscopy? First, the theories of NMR relaxation and fluorescence anisotropy decay are very similar, and thus, the methods developed in this work can be applied in both fields. Moreover, there are indeed particular advantages of fluorescence spectroscopy experiments that make it an approach complementary to other techniques: The size of the studied biomolecules is not restricted, which is the case in NMR spectroscopy. Furthermore, it can be used to study *single molecule dynamics*, which already provided much insight to structural heterogeneities in many cases.[11,20,47] In addition, fluorescence spectroscopy *in vivo* offers the great chance of monitoring conformational dynamics within a living cell.

The interpretation of fluorescence spectroscopy experiments is difficult, and it often requires knowledge of dye-biomolecule interactions in atomic detail, which is hardly available from the experiments. As has been shown in this work, molecular dynamics simulations are actually able to provide the required information. With this first step we have opened a new applica-

tion field of MD simulations to help interpret experimental results, which can be extended towards the combination of different single-molecule techniques like simultaneous application of atomic force microscopy or patch clamp methods with fluorescence spectroscopy. The growing importance of single-molecule experiments demands for the development of more sophisticated techniques, which poses considerable challenges to experiment and theory. These can be met most successfully by close collaboration between experimentalists and theoreticians, as shown in this work.

"Dankbarkeit ist ein Zeichen edler Seele."
– Aesop

Danksagung

Zum Abschluss möchte ich allen danken, die zum Entstehen und Gelingen dieser Arbeit beigetragen haben.

Mein besonderer Dank gilt Helmut Grubmüller, unter dessen Leitung diese Arbeit entstand. Nach den überaus positiven Erfahrungen während meiner Diplomarbeit, die er auch schon betreut hatte, habe ich mich sehr über das Angebot gefreut, meine Arbeit unter seiner Leitung im Rahmen einer Promotion fortsetzen zu können. In ihm habe ich einen ausgezeichneten Betreuer gefunden, bei dem ich sehr viel gelernt habe, der mir mit seinen Ideen immer wieder neue Wege aufzeigte und der meine (teils noch unausgegorenen) Ideen stets offen aufnahm und in richtige Wege zu lenken vermochte. Auch in der hektischen Phase des Aufbaus der neuen Abteilung am Max-Planck-Institut für biophysikalische Chemie hat er sich immer ausreichend Zeit für Diskussionen und Hilfestellungen genommen. Darüber hinaus hat er unserer Abteilung nicht nur ein hervorragendes materielles Umfeld zur Verfügung gestellt, sondern durch seine unkomplizierte und humorvolle Art auch ein sehr angenehmes Arbeitsklima geschaffen.

Ganz herzlich möchte ich auch Herrn Prof. Dr. Tim Salditt dafür danken, dass er sich freundlicherweise bereit erklärt hat, als Doktorvater die Betreuung der Arbeit zu übernehmen. Er hat durch wertvolle Anregungen zur Verbesserung dieser Dissertation beigetragen.

Die Zusammenarbeit mit den Mitgliedern der Abteilung war sehr fruchtbar und hat mir darüberhinaus großen Spaß gemacht. Für die vielen hilfreichen Diskussionen und die offene, freundschaftliche Atmosphäre bedanke ich mich bei Rainer Böckmann, Benjamin Bouvier, Peer Geisendorf, Bert de Groot, Helmut Grubmüller, Evi Heinemann, Berthold Heymann, Ingo Hoffmann, Jochen Hub, Harshad Joshi, Volker Knecht, Marcus Kubitzki, Carsten Kutzner, Oliver Lange, Frauke Meyer, Matthias Müller, Friedemann 'Thermometerhuhn' Reinhard, Lars Schäfer, Rudolf Schemm, Daniel Seeliger, Lingling Shen, Oliver Slawik und Martin 'Totentanz' Stumpe.

Ein großer Teil dieser Arbeit basiert auf dem Vergleich meiner Simulationen mit Fluoreszenzanisotropie-Experimenten. Diese Experimente wurden von Ulrike Alexiev an der Freien Universität Berlin durchgeführt. Ihr danke ich für diese fruchtbare Zusammenarbeit und viele anregende Diskussionen. Auch mit Claus Seidel habe ich auf diesem Gebiet intensiv und erfolgreich zusammengearbeitet. Für wertvolle Denkanstösse und viele hilfreiche Diskussionen möchte ihm herzlich danken. Ebenso bedanke ich mich bei Filipp Oesterhelt für die vielversprechende Zusammenarbeit bei FRET-Experimenten an RNA und DNA.

Für die Finanzierung meiner Promotion danke ich der Volkswagen-Stiftung und der Max-Planck-Gesellschaft. Die Teilnahme an Tagungen wurde unter anderem durch die Deutsche Forschungsgemeinschaft ermöglicht.

Ein persönlicher Dank gilt meinen großartigen Eltern, die mich zu jeder Zeit unterstützten und einen wesentlichen Teil dazu beigetragen haben, dass ich unbeschwert studieren und promovieren konnte. Zum krönenden Abschluss möchte ich meiner umwerfenden Ehefrau Elfriede danken für all die Liebe, die tollen gemeinsamen Jahre und den ganzen Spaß, den wir immer haben.

Bibliography

[1] A. L. Lehninger, D. L. Nelson, and M. M. Cox. *Biochemie*. Springer, Heidelberg, 2001.

[2] L. Stryer. *Biochemistry*. W. H. Freeman and Company, San Francisco, 1988.

[3] W. Hoppe, W. Lohmann, H. Markl, and H. Ziegler, editors. *Biophysik*. Springer, 1982.

[4] J. W. Jung and W. Lee. Structure-based functional discovery of proteins: Structural proteomics. *J. Biochem. Mol. Biol.*, 37:28–34, 2004.

[5] M. Nilges. Structure calculation from NMR data. *Curr. Opin. Struct. Biol.*, 6:617–623, 1996.

[6] A. T. Brunger and M. Nilges. Computational challenges for macromolecular structure determination by X ray crystallography and solution NMR spectroscopy. *Q. Rev. Biophys.*, 26:49–125, 1993.

[7] J. G. Kempf and J.P. Loria. Protein dynamics from solution nmr — theory and applications. *Cell Biochem. Biophys.*, 37:187–211, 2003.

[8] H. J. Steinhoff. Methods for study of protein dynamics and protein-protein interaction in protein-ubiquitination by electron paramagnetic resonance spectroscopy. *Frontiers in Bioscience*, 7:C97–C110, 2002.

[9] F. Gabel, D. Bicout, U. Lehnert, M. Tehei, M. Weik, and G. Zaccai. Protein dynamics studied by neutron scattering. *Q. Rev. Biophys.*, 35:327–367, 2002.

[10] J.C. Smith. Protein dynamics — comparison of simulations with inelastic neutron-scattering experiments. *Q. Rev. Biophys.*, 24:227–291, 1991.

[11] S. Weiss. Fluorescence spectroscopy of single biomolecules. *Science*, 283:1676–1683, 1999.

[12] T. Ha. Single-molecule fluorescence resonance energy transfer. *Methods*, 25:78–86, 2001.

[13] U. Alexiev, I. Rimke, and T. Pöhlmann. Elucidation of the nature of the conformational changes of the EF-interhelical loop in bacteriorhodopsin and of the helix VIII on the cytoplasmic surface of bovine rhodopsin: A time-resolved fluorescence depolarization study. *J. Mol. Biol.*, 328:705–719, 2003.

[14] B. Schuler, A. Lipman, and W. A. Eaton. Probing the free energy surface for protein folding with single molecule fluorescence spectroscopy. *Nature*, 419:743–747, 2002.

[15] J. Lakowicz. *Principles of Fluorescence Spectroscopy*. Kluwer Academic and Plenum Publishers, New York, 1999.

[16] T. Förster. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.*, 2, 1948.

[17] P. Wahl. Nanosecond pulsefluorimetry. *New Tech. Biophys. Cell Biol.*, 2:233, 1975.

[18] J. Yguerabide. Nanosecond fluorescence spectroscopy of macromolecules. *Methods Enzymol.*, 26:498, 1972.

[19] K. Jr. Kinosita, S. Mitaku, and A. Ikegami. Degree of dissociation of apohemoglobin studied by nanosecond fluorescence-polarization technique. *Biochim. Biophys. Acta*, 393:10, 1975.

[20] S. Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Struct. Biol.*, 7:724, 2000.

[21] T. Förster. Academic Press, New York, 1965.

[22] G. F. Schröder and H Grubmüller. FRETsg : Biomolecular structure model building from multiple FRET experiments. *Comp. Phys. Comm.*, 158:150–157, 2003.

[23] M. Margittai, J. Widengren, E. Schweinberger, G.F. Schröder, S. Felekyan, E. Haustein, M. König, D. Fasshauer, H. Grubmüller, R. Jahn, and C. A. M. Seidel. Single-molecule fluorescence resonance energy transfer reveals a dynamic equilibrium between closed and open conformations of syntaxin 1. *Proc. Natl. Acad. Sci. USA*, 100(26):15516–15521, 2003.

[24] T. Hirschfeld. Quantum efficiency independence of the time integrated emission from a fluorescent molecule. *Appl. Opt.*, 15:3135–3139, 1976.

[25] R. A. Keller, W. P. Ambrose, P. M. Goodwin, J. H. Jett, J. C. Martin, and M. Wu. Single molecule fluorescence analysis in solution. *Applied Spectroscopy*, 50:A12–A32, 1996.

[26] S. M. Nie, D. T. Chiu, and R. N. Zare. Real-time detection of single-molecules in solution by confocal fluorescence microscopy. *Analytical Chemistry*, 67:2849–2857, 1995.

[27] L. Q. Li and L. M. Davis. Rapid and efficient detection of single chromophore molecules in aqueous-solution. *Appl. Opt.*, 34:3208–3217, 1995.

[28] Y. H. Lee, R. G. Maus, B. W. Smith, and J. D. Winefordner. Laser-induced fluorescence detection of a single-molecule in a capillary. *Anal. Chem.*, 66:4142–4149, 1994.

[29] S. M. Nie, D. T. Chiu, and R. N. Zare. Probing individual molecules with confocal fluorescence microscopy. *Science*, 266:1018–1021, 1994.

[30] C. W. Wilkerson, P. M. Goodwin, W. P. Ambrose, J. C. Martin, and R. A. Keller. Detection and lifetime measurement of single molecules in flowing sample streams by laser-induced fluorescence. *Appl. Phys. Lett.*, 62:2030–2032, 1993.

[31] S. A. Soper, Q. L. Mattingly, and P. Vegunta. Photon burst detection of single near-infrared fluorescent molecules. *Anal. Chem.*, 65:740–747, 1993.

[32] W. B. Whitten, J. M. Ramsey, S. Arnold, and B. V. Bronk. Single-molecule detection limits in levitated microdroplets. *Anal. Chem.*, 63:1027–1031, 1991.

[33] S. A. Soper, E. B. Shera, J. C. Martin, J. H. Jett, J. H. Hahn, H. L. Nutter, and R. A. Keller. Single-molecule detection of rhodamine-6g in ethanolic solutions using continuous wave laser excitation. *Anal. Chem.*, 63:432–437, 1991.

[34] E. B. Shera, N. K. Seitzinger, L. M. Davis, R. A. Keller, and S. A. Soper. Detection of single fluorescent molecules. *Chem. Phys. Lett.*, 174:553–557, 1990.

[35] W. E. Moerner and Michel Orrit. Illuminating single molecules in condensed matter. *Science*, 283:1670, March 1999.

[36] F. G. Ball, Y. Cai, J. B. Kadane, and A. O'Hagan. Bayesian inference for ion-channel gating mechanisms directly from single-channel recordings, using markov chain monte carlo. *Proc. R. Soc. London Ser. A-Math. Phys. Eng. Sci.*, 455:2879, 1988.

[37] W. A. Carrington, R. M. Lynch, E. D. W. Moore, G. Isenberg, K. E. Fogarty, and F. S. Fredric. Superresolution 3-dimensional images of fluorescence in cells with minimal light exposure. *Science*, 268:1483, 1995.

[38] T. Dudok de Wit and E. Floriani. *Phys. Rev. E*, 58:5115, 1998.

[39] T. J. Loredo and D. Q. Lamb. Bayesian analysis of neutrinos observed from supernova sn 1987a. *Phys. Rev. D*, 65:063002, 2002.

[40] J. Enderlein. Maximum-likelihood criterion and single-molecule detection. *Appl. Optics*, 34:514, 1995.

[41] C. Zander, M. Sauer, K. H. Drexhage, D.-S. Ko, A. Schulz, J. Wolfrum, L. Brand, C. Eggeling, and C. A. M. Seidel. Detection and characterization of single molecules in aqueous solution. *Appl. Phys. B*, 63:517–523, 1996.

[42] J. Enderlein, P. M. Goodwin, A. Van Orden, W. P. Ambrose, R. Erdmann, and R. A. Keller. A maximum likelihood estimator to distinguish single molecules by their fluorescence decays. *Chem. Phys. Lett.*, 270:464, 1997.

[43] H. Yang and X. S. Xie. Probing single-molecule dynamics photon by photon. *J. Chem. Phys.*, 117, 2002.

[44] H. Yang and X. S. Xie. Statistical approaches for probing single-molecule dynamics photon-by-photon. *Chem. Phys.*, 284:423–437, 2002.

[45] L. P. Watkins and H. Yang. Information bounds and optimal analysis of dynamic single molecule measurements. *Biophys. J.*, 86:4015–4029, 2004.

[46] T. J. Ha, A. Y. Ting, J. Liang, A. A. Deniz, D. S. Chemla, P. G. Schultz, and S. Weiss. Temporal fluctuations of fluorescence resonance energy transfer between two dyes conjugated to a single protein. *Chem. Phys.*, 247:107, 1999.

[47] R. Kühnemuth and C. A. M. Seidel. Principles of single molecule multiparameter fluorescence spectroscopy. *Single Molecules*, 2:251, 2001.

[48] S. A. McKinney, A.-C. Dclais, D. M. J. Lilley, and T. Ha. Structural dynamics if individual holliday junctions. *Nat. Struct. Biol.*, 10:93–97, 2003.

[49] G. F. Schröder and H. Grubmüller. Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer. *J. Chem. Phys.*, 119:9920–9924, 2003.

[50] X. Zhuang, L. E. Bartley, H. P. Babcock, R. Russell, T. Ha, D. Herschlag, and S. Chu. A single-molecule study of rna catalysis and folding. *Science*, 288:2048, 2000.

[51] E. Barkai, Y. J. Jung, and R. Silbey. Time-dependent fluctuations in single molecule spectroscopy: A generalized wiener-khintchine approach. *Phys. Rev. Lett.*, 87:207403, 2001.

[52] C. Eggeling, J. R. Fries, L. Brand, R. Gnther, and C. A. M. Seidel. Monitoring conformational dynamics of a single molecule by selective fluorescence spectroscopy. *Proc. Natl. Acad. Sci. USA*, 95:1556–1561, February 1998.

[53] I. Munro, I. Pecht, and L. Stryer. Subnanosecond motions of tryptophan residues in proteins. *Proc. Natl. Acad. Sci. USA*, 76:56–60, 1979.

[54] L. J. Juszczak, Z.-Y. Zhang, D. S. Gottfried L. Wu, and D. D. Eads. Rapid loop dynamics of yersinia protein tyrosine phosphatases. *Biochemistry*, 36:2227–2236, 1997.

[55] R. Swaminathan, U. Nath, J. B. Udgaonkar, N. Periasamy, and G. Krishnamoorthy. Motional dynamics of a buried tryptophan reveals the presence of partially structured forms during denaturation of barstar. *Biochemistry*, 35:9150–9157, 1996.

[56] K. Doring, W. Beck, L. Konermann, and F. Jahnig. The use of a long-lifetime component of tryptophan to detect slow orientational fluctuations of proteins. *Biophys. J.*, 72:326–334, 1997.

[57] J. E. Hansen, S. J. Rosenthal, and G. R. Fleming. Subpicosecond fluorescence depolarization studies of tryptophan and tryptophanyl residues of proteins. *J. Phys. Chem.*, 96:3034–3040, 1992.

[58] H. R. M. Leenders, J. Vervoort, A. Vanhoek, and A. J. W. G. Visser. Time-resolved fluorescence studies of flavodoxin - fluorescence decay and fluorescence anisotropy decay of tryptophan in desulfovibrio flavodoxins. *Europ. Biophys. J.*, 18:43–55, 1990.

[59] A. J. W. G. Visser, T. Ykema, A. Vanhoek, D. J. Okane, and J. Lee. Determination of rotational correlation times from deconvoluted fluorescence anisotropy decay curves - demonstration with 6,7-dimethyl-8-ribityllumazine and lumazine protein from photobacterium-leiognathi as fluorescent indicators. *Biochemistry*, 24:1489–1496, 1985.

[60] J. R. Lakowicz, B. P. Maliwal, H. Cherek, and A. Balter. Rotational freedom of tryptophan residues in proteins and peptides. *Biochemistry*, 22:1741–1752, 1983.

[61] C. Laboulais, E. Deprez, H. Leh, J. F. Mouscadet, J. C. Brochon, and M. Le Bret. HIV-1 integrase catalytic core: Molecular dynamics and simulated fluorescence decays. *Biophys. J.*, 81:473–489, 2001.

[62] A. J. W. G. Visser, P. A. W. van den Berg, N. V. Visser, A. van Hoek, H. A. van den Burg, D. Parsonage, and A. Claiborne. Time-resolved fluorescence of flavin adenine dinucleotide in wild-type and mutant nadh peroxidase. elucidation of quenching sites and discovery of a new fluorescence depolarization mechanism. *J. Phys. Chem. B*, 102:10431–10439, 1998.

[63] J. Schaffer, A. Volkmer, C. Eggeling, V. Subramaniam, G. Striker, and C. A. M. Seidel. Identification of single molecules in aqueous solution by time-resolved fluorescence spectroscopy. *J. Phys. Chem. A*, 103:331–336, 1999.

[64] Jr. K. Kinosita, R. Kataoka, Y. Kimura, O. Gotoh, and A. Ikegami. Dynamic structure of biological membranes as probed by 1,6-diphenyl-1,3,5-hexatriene: A nanosecond fluorescence depolarization study. *Biochemistry*, 20:4270–4277, 1981.

[65] N. C. Maiti, M. M. G. Krishna, P. J. Britto, and N. Periasamy. Fluorescence dynamics of dye probes in micelles. *J. Phys. Chem. B*, 101:11051–11060, 1997.

[66] E. L. Quitevis, A. H. Marcus, and M. D. Fayer. Dynamics of ionic lipophilic probes in micelles - picosecond fluorescence depolarization measurements. *J. Phys. Chem.*, 97:5762–5769, 1993.

[67] A. Szabo. Theory of fluorescence depolarization in macromolecules and membranes. *J. Chem. Phys.*, 81:150–167, 1984.

[68] R. P. H. Kooyman, M. H. Vos, and Y. K. Levine. Determination of orientational order parameters in oriented lipid-membrane systems by angle-resolved fluorescence depolarization experiments. *Chem. Phys.*, 81:461–472, 1983.

[69] C. Zannoni, A. Arcioni, and P. Cavatorta. Fluorescence depolarization in liquid-crystals and membrane bilayers. *Chem. Phys. Lipids*, 32:179–250, 1983.

[70] K. Kinosita, R. Kataoka, Y. Kimura, O. Gotoh, and A. Ikegami. Dynamic structure of biological-membranes as probed by 1,6-diphenyl-1,3,5-hexatriene - a nanosecond fluorescence depolarization study. *Biochemistry*, 20:4270–4277, 1981.

[71] J. Seelig, L. Tamm, L. Hymel, and S. Fleischer. Deuterium and phosphorus nuclear magnetic-resonance and fluorescence depolarization studies of functional reconstituted sarcoplasmic-reticulum membrane-vesicles. *Biochemistry*, 20:3922–3932, 1981.

[72] G. Lipari and A. Szabo. Effect of librational motion on fluorescence depolarization and nuclear magnetic-resonance relaxation in macromolecules and membranes. *Biophys. J.*, 30:489–506, 1980.

[73] O. F. A. Larsen, I. H. M. van Stokkum, B. Gobets, R. van Grondelle, and H. van Amerongen. Probing the structure and dynamics of a DNA hairpin by ultrafast quenching and fluorescence depolarization. *Biophys. J.*, 81:1115–1126, 2001.

[74] S. Georghiou, T. D. Bradrick, A. Philippetis, and J. M. Beechem. Large-amplitude picosecond anisotropy decay of the intrinsic fluorescence of double-stranded DNA. *Biophys. J.*, 70:1909–1922, 1996.

[75] A. Larsson, C. Carlsson, M. Jonsson, and N. Albinsson. Characterization of the binding of the fluorescent dyes yo and yoyo to DNA by polarized-light spectroscopy. *J. Am. Chem. Soc.*, 116:8459–8465, 1994.

[76] M. D. Barkley and B. H. Zimm. Theory of twisting and bending of chain macromolecules - analysis of the fluorescence depolarization of DNA. *J. Chem. Phys.*, 70:2991–3007, 1979.

[77] R. M. Levy and A. Szabo. Initial fluorescence depolarization of tyrosines in proteins. *JACS*, 104:2073–2075, 1982.

[78] S. Ringhofer, J. Kallen, R. Dutzler, A. Billich, A. J. W. G. Visser, D. Scholz, O. Steinhauser, H. Schreiber, M. Auer, and A. J. Kungl. X-ray structure and conformational dynamics of the HIV-1 protease in complex with the inhibitor SDZ283-910: Agreement of time-resolved spectroscopy and molecular dynamics simulations. *J. Mol. Biol.*, 286:1147–1159, 1999.

[79] P. H. Axelsen, C. Haydock, and F. G. Prendergast. Molecular dynamics of tryptophan in Ribonuclease-T1. *Biophys. J.*, 54:249–258, 1988.

[80] E. R. Henry and R. Hochstrasser. Molecular dynamics simulations of fluorescence polarization of tryptophans in myoglobin. *Proc. Natl. Acad. Sci. USA*, 84:6142–6146, 1987.

[81] T. Ichiye and M. Karplus. Fluorescence depolarization of tryptophan residues in proteins: A molecular dynamics study. *Biochemistry*, 22:2884–2893, 1983.

[82] D. L. Harris and B. S. Hudson. Fluorescence and molecular-dynamics study of the internal motion of the buried tryptophan in bacteriophage-T4 lysozyme - effects of temperature and alteration of nonbonded networks. *Chem. Phys.*, 158:353–382, 1991.

[83] P. H. Axelsen, E. Gratton, and F. G. Prendergast. Experimentally verifying molecular dynamics simulations through fluorescence anisotropy measurements. *Biochemistry*, 30:1173–1179, 1991.

[84] M. Gentin, M. Vincent, J. C. Brochon, A. K. Livesey, N. Cittanova, and J. Gallay. Time-resolved fluorescence of the single tryptophan residue in rat alpha-fetoprotein and rat serum-albumin - analysis by the maximum-entropy method. *Biochemistry*, 29:10405–10412, 1990.

[85] J. P. Duneau, N. Garnier, G. Cremel, G. Nullans, P. Hubert, D. Genest, M. Vincent, J. Gallay, and M. Genest. Time resolved fluorescence properties of phenylalanine in different environments. comparison with molecular dynamics simulation. *Biophys. Chem.*, 73:109–119, 1998.

[86] G. S. Jas, Y. Wang, S. W. Pauls, C. K. Johnson, and K. Kuczera. Influence of temperature and viscosity on anthracene rotational diffusion in organic solvents: Molecular dynamics simulations and fluorescence anisotropy study. *J. Chem. Phys.*, 107:8800–8812, 1997.

[87] G. S. Jas, E. J. Larson, C. K. Johnson, and K. Kuczera. Microscopic details of rotational diffusion of perylene in organic solvents: Molecular dynamics simulation and experiment vs Debye-Stokes-Einstein theory. *J. Phys. Chem. A*, 104:9841–9852, 2000.

[88] X. Daura, R. Suter, and W. F. van Gunsteren. Validation of molecular simulation by comparison with experiment: Rotational reorientation of tryptophan in water. *J. Chem. Phys.*, 110:3049–3055, 1999.

[89] P. Mark and L. Nilsson. A molecular dynamics study of tryptophan in water. *J. Phys. Chem. B.*, 106:9440–9445, 2002.

[90] K. Lim and J. N. Herron. Molecular-dynamics of the antifluorescein-4-4-20 antigen-binding fragment. 1. computer-simulations. *Biochemistry*, 34:6962–6974, 1995.

[91] G. R. Marshall. Peptide interactions with G-protein coupled receptors. *Biopolymers*, 60:246–277, 2001.

[92] A.-O. Colson, J. H. Perlman, A. Smolyar, M. C. Gershengorn, and R. Osman. Static and dynamics role of extracellular loops in G-protein coupled receptors: A mechanism for sequential binding of thyrotropin-releasing hormone to its receptor. *Biophys. J.*, 74:1087–1100, 1998.

[93] A. E. Garcia. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.*, 68:2696–2699, 1992.

[94] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993.

[95] F. Perrin. La fluorescence des solutions. Polarisation. Vie moyenne des molécules dans l'état excité. *J. de Phys.*, 7:390–401, 1926.

[96] T. Tao. Time-dependent fluorescence depolarization and Brownian rotational diffusion coefficients of macromolecules. *Biopolymers*, 8:609–632, 1969.

[97] K. Kinosita, S. Kawato, and A. Ikegami. A theory of fluorescence polarization decay in membranes. *Biophys. J.*, 20:289–305, 1977.

[98] B. W. van der Meer, G. Cooker, and S.-Y. Chen. In *Resonance Energy Transfer: Theory and Data*. VCH Publishers, New York, 1994.

[99] T. J. Ha, A. Y. Ting, J. Liang, W. B. Caldwell, A. A. Deniz, D. S. Chemla, P. G. Schultz, and S. Weiss. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc. Natl. Acad. Sci. USA*, 96:893, 1999.

[100] N. L. Goddard, G. Bonnet, O. Krichevsky, and A. Libchaber. Sequence dependent rigidity of single stranded dna. *Phys. Rev. Lett.*, 85:2400, 2000.

[101] X. Zhuang, H. Kim, M. J. B. Pereira, H. P. Babcock, N. G. Walter, and S. Chu. Correlating structural dynamics and function in single ribozyme molecules. *Science*, 296:1473, 2002.

[102] P. J. Rothwell, S. Berger, O. Kensch, S. Felekyan, M. Antonik, B. M. Wöhrl, T. Restle, R. S. Goody, , and C. A. M. Seidel. Multiparameter single-molecule fluorescence spectroscopy reveals heterogeneity of hiv-1 reverse transcriptase: primer/template complexes. *Proc. Natl. Acad. Sci. USA*, 100:1655–1660, 2003.

[103] A. T. Brünger. Free r-value - a novel statistical quantity for assessing the accuracy of crystal-structures. *Nature*, 355:472–475, 1992.

[104] W. F. van Gunsteren and H. J. C. Berendsen. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.*, 29:992–1023, 1990.

[105] P. W. Atkins. *Molecular Quantum Mechanics*. Oxford University Press, Oxford, 2nd edition edition, 1983.

[106] H. Haken and H. C. Wolf. *Molekülphysik und Quantenchemie*. Springer-Verlag, Berlin, 1991.

[107] J. E. Lennard-Jones. *Proc. Roy. Soc. London Ser. A*, 106:463, 1924.

[108] A. Warshel. *The consistent force field and its quantum mechanical extensions*. Plenum Press, New York, 1977.

[109] 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. *J. Comp. Chem.*, 4:187–217, 1983.

[110] W. F. van Gunsteren and H. J. C. Berendsen. *Groningen Molecular Simulation (GROMOS) Library Manual*. Biomos, Groningen, 1987.

[111] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.*, 7:230–252, 1986.

[112] N. L. Allinger, Y. H. Yuh, and J.H. Lii. Molecular mechanics. The MM3 force fields for hydrocarbons. *J. Am. Chem. Soc.*, 111:8551–8566, 1989.

[113] K. Rasmussen. How to develop force fields: An account of the emergence of potential energy functions for saccharides. *J. Mol. Struct.*, 395-396:81–90, 1997.

[114] P. Derreumaux, M. Dauchez, and G. Vergoten. The structures and vibrational frequencies of a series of alkanes using the SPASIBA force-field. *J. Mol. Struct.*, 295:203–221, 1993.

[115] P. H. Hünenberger and Wilfred F. van Gunsteren. Empirical classical interaction functions for molecular systems. In W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors, *Computer Simulation of Biomolecular Systems, theoretical and experimental applications*, volume 3, pages 3–82. Kluwer/Escom, Dordrecht, The Netherlands, 1997.

[116] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. Gromacs: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[117] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. J. Tasumi. The protein data bank: A computer-based archival file for molecular structures. *J. Mol. Biol.*, 112:557, 1977.

[118] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[119] Molecular Simulations Inc. *QUANTA97*. University of York, York, England, 1986–1998.

[120] Molecular Simulations Inc. *INSIGHTII(97.2)*. University of York, York, England, 1986-1998.

[121] Cray Research Inc. *UniChem 3.0*. 655 Lone Oak Drive, Eagan, Minnesota 55151, 1997.

[122] Reto Koradi, Martin Billeter, and Kurt Wüthrich. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.*, 14:51–55, 1996.

[123] R. W. Hockney and S. P. Goel. Quiet high-resolution computer models of a plasma. *J. Comp. Phys.*, 14(2):148–158, 1974.

[124] E. Hairer, S. P. Nørsett, and G. Wanner. Solving Ordinary Differential Equations I. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1987.

[125] M. F. Perutz. Electrostatic effects in proteins. *Science*, 201(4362):1187–1191, 1978.

[126] A. Warshel and S. T. Russel. Calculations of electrostatic interactions in biological systems and in solution. *Q. Rev. Biophys.*, 17:283–422, 1984.

[127] M. K. Gilson and B. H. Honig. Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies and conformational analysis. *Proteins*, 4(1):7–18, 1988.

[128] W. F. van Gunsteren, F. J. Luque, D. Timms, and A. E. Torda. *Molecular Mechanics in Biology: From Structure to Function, Taking Account of Solvation*, pages 849–863. Annual Reviews Inc., 1994.

[129] D. Eisenberg and A. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[130] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrikson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, 112:6127–6129, 1990.

[131] M. Orozco, C. Alhambra, X. Barril, J. M. López, M. A. Busquets, and F J. Luque. Theoretical methods for the representation of solvent. *J. Mol. Model.*, 2:1–15, 1996.

[132] B. Roux and T. Simonson. Implicit solvent models. *Biophys. Chem.*, 78(1-2):1–20, 1999.

[133] C. J. Cramer and D. G. Truhlar. Implicit solvation models: Equilibria, structure, spectra, and dynamics. *Chem. Rev.*, 99(8):2161–2200, 1999.

[134] M. Orozco and F. J. Luque. Theoretical methods for the description of the solvent effect in biomolecular systems. *Chem. Rev.*, 100(11):4187–4225, 2000.

[135] T. Simonson. Macromolecular electrostatics: continuum models and their growing pain. *Curr. Opin. Struct. Biol.*, 11(2):243–252, 2001.

[136] F. Fraternali and W. F. van Gunsteren. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.*, 256:939–948, 1996.

[137] M. E. Davis and J. A. McCammon. Electrostatics in biomolecular structure and dynamics. *Chem. Rev.*, 90(3):509–521, 1990.

[138] Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: Fifty years after Kramers. *Rev. Mod. Phys.*, 62(2):251–341, Apr. 1990.

[139] R. Kubo, M. Toda, and N. Hashitsume. *Statistical Physics II, Nonequilibrium Statistical Mechanics*. Springer Series in Solid-State Sciences. Springer-Verlag, Berlin, 2nd edition, 1995.

[140] R. M. Levy, M. Karplus, and J. A. McCammon. Diffusive Langevin dynamics of model alkanes. *Chem. Phys. Lett.*, 65(1):4–11, 1979.

[141] Ram Brustein, Shlomo Marianer, and Moshe Schwartz. Langevin memory kernel and noise from Lagrangian dynamics. *Physica A*, 175:47–58, 1991.

[142] Mark E. Tuckerman and Bruce J. Berne. Stochastic molecular dynamics in systems with multiple time scales and memory friction. *J. Chem. Phys.*, 95(6):4389–4396, September 1991.

[143] G. Otting, E. Liepinsh, and K. Wuthrich. Protein hydration in aqueous-solution. *Science*, 254:974–980, 1991.

[144] A. R. Bizzarri and S. Cannistraro. Molecular dynamics of water at the protein-solvent interface. *J. Phys. Chem. B*, 106:6617–6633, 2002.

[145] S. M. Bhattacharyya, Z. G. Wang, and A. H. Zewail. Dynamics of water near a protein surface. *J. Phys. Chem. B*, 107:13218–13228, 2003.

[146] A. J. Stace and J. N. Murrell. Molecular-dynamics and chemical reactivity - computer study of iodine atom recombination under high-pressure conditions. *Macromolecules*, 33(1):1–24, 1977.

[147] K. A. Sharp. Inclusion of solvent effects in molecular mechanics force fields. In W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson, editors, *Computer Simulation of Biomolecular Systems, theoretical and experimental applications*, volume 2, pages 147–160. Escom, Leiden, The Netherlands, 1993.

[148] M. Berkowitz and J. A. McCammon. Molecular dynamics with stochastic boundary conditions. *Chem. Phys. Lett.*, 90:215, 1982.

[149] C. L. Brooks III and M. Karplus. Deformable stochastic boundaries in molecular dynamics. *J. Chem. Phys.*, 79:6312–6325, 1983.

[150] R. Kossmann. Entwicklung eines effektiven Randpotentials für Molekulardynamiksimulationen wäßriger Lösungen. Master's thesis, Ludwig-Maximilians-Universität München, 1997.

[151] S. Nosé. A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.

[152] W. G. Hoover. Canonical dynamics — equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[153] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with a coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[154] Charles L. Brooks III, B. Montgomery Pettitt, and Martin Karplus. Structural and energetic effects of truncating long ranged interactions in ionic and polar fluids. *J. Chem. Phys.*, 83(11):5897–5908, December 1985.

[155] Richard J. Loncharich and Bernard R. Brooks. The effects of truncating long-range forces on protein dynamics. *Proteins*, 6:32–45, 1989.

[156] H. Schreiber and O. Steinhauser. Cutoff size does strongly influence molecular-dynamics results on solvated polypeptides. *Biochemistry*, 31(25):5856–5860, 1992.

[157] M. Saito. Molecular dynamics simulations of proteins in water without the truncation of long-range Coulomb interactions. *Mol. Sim.*, 8:321–333, 1992.

[158] M. Saito. Molecular dynamics simulations of proteins in solution - artifacts by the cutoff approximation. *J. Chem. Phys.*, 101(5):4055–4061, 1994.

[159] M. Saito. Molecular-dynamics free-energy study of a protein in solution with all degrees of freedom and long-range coulomb interactions. *J. Phys. Chem.*, 99(46):17043–17048, 1995.

[160] X. Rozanska and C. Chipot. Modeling ion-ion interaction in proteins: A molecular dynamics free energy calculation of the guanidinium-acetate association. *J. Chem. Phys.*, 112(22):9691–9694, 2000.

[161] P. P. Ewald. Die Berechnug optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, IV(64):253–287, 1921.

[162] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[163] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. The smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577, 1995.

[164] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.

[165] W. F. van Gunsteren and M. Karplus. Effects of constraints on the dynamics of macromolecules. *Macromolecules*, 15:1528–1544, 1982.

[166] Molecular Simulations, Inc. *DMol Version 960, Density Functional Theory electronic structure program.*, 1996.

[167] M. J. Frisch et al. *Gaussian 98, Revision A.5.* Gaussian, Inc., Pittsburgh PA, 1998.

[168] J. P. Stewart. *MOPAC 5.0.* Stewart Computational Chemistry.

[169] U. C. Singh and P. A. Kollman. An approach to computing electrostatic charges for molecules. *J. Comp. Chem.*, 5:129–145, 1984.

[170] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.

[171] A. Gersho and R. M. Gray. *Vector quantization and signal processing.* Kluwer Academics Publisher, Boston, MA, 1992.

[172] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-2:129–137, 1982.

[173] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.

[174] S.J. Remington and B.W. Matthews. A systematic approach to the comparison of protein structures. *J. Mol. Biol.*, 140:77–99, 1980.

[175] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, London, 1979.

[176] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[177] G. Reinelt. TSPLIB–a traveling salesman problem library. *ORSA J. Comput.*, 3:376–384, 1991.

[178] A. Schrijver. On the history of combinatorial optimization (till 1960). *Preprint*. electronic publication *http://homepages.cwi.nl/ lex/files/histco.ps*.

[179] R. M. Scheek A. Amadei G. Vriend B. L. de Groot, D. M. F. van Aalten and H. J. C. Berendsen. Prediction of protein conformational freedom from distance constraints. *PROTEINS: Struct. Funct. Gen.*, 29:240–251, 1997.

[180] U. Alexiev. personal communication.

[181] A. C. Vaiana, A. Schulz, J. Wolfrum, M. Sauer, and J. C. Smith. Molecular mechanics force field parametrization of the fluorescent probe rhodamine 6g using automated frequency matching. *J. Compt. Chem.*, 24:632–639, 2003.

[182] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, and J. Hermans. In B. Pullman, editor, *Intermolecular Forces*. Reidel, Dordrecht, Netherlands, 1981.

[183] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91:6269–6271, 1987.

[184] Y. Hu and G. R. Fleming. *J. Chem. Phys.*, 94:3857, 1990.

[185] J. Schaffer. *Charakterisierung von Einzelmolekülen durch selektive Fluoreszenzspektroskopie*. PhD thesis, Georg-August-Universität Göttingen, 2000.

[186] L. X.-Q. Chen, R. A. Engh, and G. R. Fleming. Reorientation of tryptophan and simple peptides: onset of internal flexibility and comparison with molecular dynamics simulation. *J. Phys. Chem.*, 92:4811–4816, 1988.

[187] D. van der Spoel, P. J. Van Maaren, and H. J. C. Berendsen. A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *J. Chem. Phys.*, 108:10220–10230, 1998.

[188] E. J. W. Wensink, A. C. Hoffmann, P. J. van Maaren, and D. van der Spoel. Dynamic properties of water/alcohol mixtures studied by computer simulation. *J. Chem. Phys.*, 119:7308–7317, 2003.

[189] Z. J. Derlacki, A. J. Easteal, A. V. J. Edge, and L. A. Woolf. Diffusion coefficients of methanol and water and the mutual diffusion coefficient in methanol-water solutions at 278 and 298 k. *J. Phys. Chem.*, 89:5318–5322, 1985.

[190] R. L. Smith Jr., S. B. Lee, H. Komori, and K. Arai. Relative permittivity and dielectric relaxation in aqueous alcohol solutions. *Fluid Phase Equilibria*, 144:315–322, 1998.

[191] N. G. Fuller and R. L. Rowley. The effect of model internal flexibility upon NEMD simulations of viscosity. *Int. J. Thermophys.*, 21:45–55, 2000.

[192] V. Y. Orekhov, K. V. Pervushin, and A. S. Arseniev. Three-dimensional structure of (1-71)bacterioopsin solubilized in methanol/chloroform and SDS micelles determined by ^{15}N-^{1}H heteronuclear NMR spectroscopy. *Eur. J. Biochem.*, 219:571–583, 1994.

[193] E. Pebay-Peyroula, G. Rummel, J. P. Rosenbusch, and E. M. Landau. X-ray structure of bacteriorhodopsin at 2.5 Angstroms from microcrystals grown in lipidic cubic phases. *Science*, 277:1676–1681, 1997.

[194] S. A. Hassan, F. Guarnieri, and E. L. Mehler. A general treatment of solvent effects based on screened coulomb potentials. *J. Phys. Chem. B.*, 104:6478–6489, 2000.

[195] S. Lee and M. Karplus. Brownian dynamics simulations - statistical error of correlation-functions. *J. Chem. Phys.*, 81:6106–6118, 1984.

[196] K. Kinosita, A. Ikegami, and S. Kawato. On the wobbling-in-a-cone analysis of fluorescence anisotropy decay. *Biophys. J.*, 37:461–464, 1982.

[197] R. Abseher, H. Schreiber, and O. Steinhauser. The influence of a protein on water dynamics in its vicinity investigated by molecular dynamics simulation. *PROTEINS: Struct., Funct. Gen.*, 25:366–378, 1996.

[198] V. Y. Orekhov, K. V. Pervushin, and A. S. Arseniev. Backbone dynamics of (1-71)bacteriorhodopsin studied by two-dimensional ^{1}H-^{15}N NMR spectroscopy. *Eur. J. Biochem.*, 219:887–896, 1994.

[199] H.M. Watrob, C.-P. Pan, and M. D. Barkley. Two-step FRET as a structural tool. *J. Am. Chem. Soc.*, 125:7336–7343, 2003.

[200] S. Hohng, C. Joo, and T. Ha. Single-molecule three-color FRET. *Biophys. J.*, 87:1328–1337, 2004.