

# MULTISCALE MODELING OF BIOMOLECULAR SYSTEMS

BY

LORANT JANOSI

DISSERTATION

Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in Physics  
in the Graduate College of the  
University of Missouri-Columbia

DECEMBER 2007

COMMITTEE ON FINAL EXAMINATION\*

Ioan Kosztin  
Gabor Forgacs  
Dong Xu  
Carlos Wexler  
Shi-Jie Chen

\* The Certificate of Committee Approval is not part of dissertations deposited electronically through the Graduate College of the University of Missouri-Columbia. A copy of the original Certificate of Committee Approval is permanently archived with the student's record.

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

MULTISCALE MODELING OF BIOMOLECULAR SYSTEMS

presented by Lorant Janosi,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

---

Professor Ioan Kosztin

---

Professor Gabor Forgacs

---

Professor Dong Xu

---

Professor Carlos Wexler

---

Professor Shi-Jie Chen

*To Oana, my parents, Claudius and Anca.*

# Acknowledgments

I would like to thank Dr. Ioan Kosztin for his guidance, constant dedication and tremendous support, my collaborators Dr. Thorsten Ritz and Harindar Keer, the members of my committee and my colleagues, for their various contributions to this work. I would also like to acknowledge the financial support by grants from: University of Missouri Research Board, the Institute for Theoretical Sciences (a joint institute of Notre Dame University and Argonne Natl. Lab.), and the Department of Energy (DOE Contracts W-7405-ENG-36, W-31-109-ENG-38). Supercomputer time was provided by an AAB grant from NCSA. Generous computational resources were also provided by the University of Missouri Bioinformatics Consortium.

# Table of Contents

<b>List of Tables</b> . . . . .	<b>v</b>
<b>List of Figures</b> . . . . .	<b>vi</b>
<b>List of Abbreviations</b> . . . . .	<b>xi</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
<b>2 Theoretical prediction of optical and spectral properties of pigment-protein complexes</b> . . . . .	<b>6</b>
2.1 Introduction . . . . .	7
2.2 Optical Transitions at Finite Temperature . . . . .	11
2.2.1 Linear Absorption and Lineshape Function . . . . .	12
2.2.2 Circular Dichroism . . . . .	17
2.3 Simulations of Optical Spectra . . . . .	18
2.3.1 Building the LH2 Rings . . . . .	18
2.3.2 Molecular Dynamics Simulations . . . . .	19
2.3.3 Quantum Chemistry Calculations . . . . .	19
2.4 Energy Gap Density of States . . . . .	22
2.5 Linear Absorption Spectrum . . . . .	27
2.6 Circular Dichroism . . . . .	32
2.7 Conclusions . . . . .	35
<b>3 Calculating free energy profiles and diffusion coefficients from non-equilibrium molecular dynamics simulations</b> . . . . .	<b>37</b>
3.1 Introduction . . . . .	38
3.2 Reaction Coordinates and Potential of Mean Force . . . . .	42
3.3 Harmonic Guiding Potential . . . . .	43
3.4 Stiff Spring Approximation . . . . .	44
3.5 PMF Using Umbrella Sampling and WHAM . . . . .	44
3.6 The Transient Fluctuation Theorem and Jarzynski's Equality . . . . .	46
3.7 PMF from Unidirectional SMD Pulls . . . . .	47
3.8 PMF from Forward and Reverse SMD Pulls . . . . .	50
3.9 Generalized Acceptance Ratio Method . . . . .	52
3.10 Helix-to-Coil Transition of Deca-Alanine . . . . .	54
3.10.1 System Modeling and SMD simulations . . . . .	54
3.10.2 Potential of Mean Force . . . . .	56

3.11	Conclusions . . . . .	62
<b>4</b>	<b>Ion and molecular transport through protein channels . . . . .</b>	<b>64</b>
4.1	Introduction . . . . .	65
4.2	Potassium Ion Transport through Gramicidin-A Channel . . . . .	66
4.2.1	Modeling of Gramicidin-A Channel . . . . .	68
4.2.2	SMD Simulations . . . . .	69
4.2.3	Potential of Mean Force . . . . .	70
4.2.4	Conclusions . . . . .	75
4.3	Glycerol Transport through GlpF Channel . . . . .	76
4.3.1	Modeling of GlpF Channel . . . . .	77
4.3.2	SMD Simulations . . . . .	80
4.3.3	Potential of Mean Force and Diffusion Coefficient . . . . .	82
4.3.4	Mean First Passage Time . . . . .	86
4.3.5	Channel Permeability . . . . .	87
4.3.6	Conclusions . . . . .	89
<b>5</b>	<b>In silico prediction of the oligomerization state of light harvesting complexes . . . . .</b>	<b>90</b>
5.1	Introduction . . . . .	91
5.2	Interaction Between Two LH2 Subunits . . . . .	95
5.2.1	Reaction Coordinates $R$ and $\theta$ and PMF $U(\theta, R)$ . . . . .	95
5.2.2	System Modeling and MD Simulations . . . . .	97
5.2.3	Equilibrium MD Simulations of MOLI and ACI Dimers . . . . .	99
5.2.4	Forced Detachment of the Subunits Using SMD . . . . .	101
5.2.5	Calculation of PMF $U(\theta, R)$ . . . . .	105
5.2.6	Preferred Angle . . . . .	107
5.3	Prediction of the LH2 Ring Size . . . . .	108
5.3.1	System Modeling and MD Simulations . . . . .	109
5.3.2	Energy Minimization . . . . .	112
5.3.3	Free MD Simulation . . . . .	112
5.3.4	Potential of Mean Force $U(\theta)$ . . . . .	117
5.4	Conclusions . . . . .	120
	<b>References . . . . .</b>	<b>123</b>
	<b>Vita . . . . .</b>	<b>136</b>

# List of Tables

5.1	Average separation distances at which the most important inter-residue links between two LH2 subunits of MOLI are broken in SMD simulations. .	104
5.2	Average separation distances at which the most important inter-residue links between two LH2 subunits of ACI are broken in SMD simulations. . .	104
5.3	The total number of atoms, the number of lipids and of water molecules, of the built dimer systems at initial angles $\theta_0 = \{36^\circ, 40^\circ, 45^\circ\}$ . . . . .	110
5.4	Angle values determined from the three approaches for the nine systems: (i) final angle value $\theta_{final}$ after energy minimization, (ii) mean value of $\theta$ (averaged over the last 7 ns) and the peak position of its distribution, obtained from the 10 ns free MD simulations, and (iii) angle value $\theta_{min}$ corresponding to the global minimum of the PMF $U(\theta)$ . All angles are reported in degrees. . . . .	114

# List of Figures

2.1	(a) Side view of an octameric LH2 complex from <i>Rs. molischianum</i> embedded in a fully solvated POPC lipid bilayer. The transmembrane helices of the apoprotein subunits are shown as cylinders (cartoon representation) and are colored by residue type; dark (light) colors represent hydrophilic (hydrophobic) residues. For clarity only the BChl macrocycles are shown and the front half of the lipids are not shown. The clearly visible B800 (B850) ring is surrounded mostly by polar and charged (nonpolar) protein residues. (b) Tilted side view of the quantum system formed by the optically active B800 and B850 macrocycles that form rings oriented parallel to the surface of the membrane. . . . .	9
2.2	Normalized DOS, $\mathcal{N}(\omega)$ , for (a) B800, and (b) B850 BChls in LH2 of <i>Rs. molischianum</i> computed as binned histograms of the corresponding $Q_y$ excitation energy time series obtained from combined MD/QC simulations. Whether the charge fluctuations of the BChls' environment are included (solid lines) or not (dashed line) makes an important difference in $\mathcal{N}(\omega)$ only for B800. In (b) the DOS of the B850 excitons is shown as a thick solid line. . . . .	23
2.3	Average transition dipole moments $\langle d_J \rangle$ corresponding to the $J = 1, \dots, 16$ B850 excitonic states. Both $\langle d_J \rangle$ and the corresponding error bars are expressed relative to the mean dipole moment of individual B850s. . . . .	24
2.4	Absorption spectrum $I_{DOS}(\omega)$ of LH2 for <i>Rs. molischianum</i> calculated as a combined DOS of B800 BChls and B850 excitons weighted by the corresponding dipole strengths (solid line). $I_{DOS}(\omega)$ was blueshifted by 20 meV in order to overlay its B850 peak with the corresponding one in the experimental OD spectrum [31] (dashed line). . . . .	26
2.5	Normalized autocorrelation function $C(t)/C(0)$ of the energy gap fluctuations $\delta E(t) = E(t) - \langle E \rangle$ for individual B800 (dashed line) and B850 (solid line) BChls, calculated using Eq. (2.24). The mean square energy gap fluctuations are $C_{B800}(0) = 3.16 \times 10^{-3} \text{ eV}^2$ and $C_{B850}(0) = 8.68 \times 10^{-4} \text{ eV}^2$ . . . . .	27
2.6	Spectral density function $J(\omega)$ for B800 (dashed line) and B850 (solid line) obtained according to Eq. (2.12). . . . .	29
2.7	Lineshape functions $\bar{A}_{B800}(\Delta\omega)$ (dashed line) and $\bar{A}_{B850}(\Delta\omega)$ (solid line). . . . .	30
2.8	Computed (solid line) and experimental (dashed line) absorption spectra (in arbitrary units) of the BChl aggregate in <i>Rs. molischianum</i> LH2. The computed spectrum has been blue shifted by 20 meV for best match. . . . .	31



2.9	Mean rotational strength of the excitonically coupled B800 (circles) and B850 (rectangles) BChls as a function of the corresponding excitonic energies. The purpose of the thin lines is to guide the eye. . . . .	33
2.10	(a) CD spectrum contributions due to B800 (dashed line) and B850 (solid line) BChls. (b) Comparison between the computed (solid line) and experimental (dashed line) CD spectrum of the BChl aggregate in <i>Rs. molischianum</i> LH2. . . . .	34
3.1	Gaussian forward (long-dashed), reverse (dotted) and dissipative (solid) work distribution functions within the <i>stiff spring approximation</i> . The shaded region in $P_{F/R}(W)$ is the one sampled in F/R SMD pulls. The tail region of $P_d(W)$ corresponding to negative dissipative work is also highlighted. . . . .	41
3.2	(a) Cartoon representation of deca-alanine. The reaction coordinate $R$ is defined as the distance between the first ( $CA_1$ ) and last ( $CA_{10}$ ) $C_\alpha$ atoms, i.e., the end-to-end distance of the peptide. The spring, with elastic constant $k$ , connecting $CA_1$ and $CA_{10}$ corresponds to an elastic guiding potential $V(R;t) = (k/2)[R - R_0(t)]^2$ that can be used to cycle deca-alanine between the (b) folded and (c) unfolded (completely stretched) conformations. In (b) and (c) the backbone (sidechain) atoms are shown in cartoon (CPK) representation. The hydrogen bonds that stabilize the $\alpha$ -helix are also shown in configuration (b). . . . .	55
3.3	Potential of mean force (PMF) of deca-alanine as a function of the reaction coordinate $R$ . The different curves were obtained with the FR method by employing different simulation and PMF calculation protocols described in the text. . . . .	57
3.4	Comparison between the PMFs $U(R)$ obtained using the FR method (thick solid line) and the cumulant approximation of the JE corresponding to the ten forward (dashed line) and reverse (dot-dashed line) SMD trajectories, respectively. The thin solid line corresponds to the exact PMF. The upper (lower) panels correspond to a uniform pulling speed of 1 Å/ns (0.1 Å/ns). The PMFs in the right panels were determined by dividing the 20 Å pulling distance into ten equidistant segments (the system being equilibrated in each of the end points of the individual segments), while the PMFs in the left panels were determined by considering the entire pulling distance as a single segment. . . . .	58
3.5	Histogram of the distribution functions (thin solid lines) of the forward ( $W_F$ ) and reverse ( $W_R$ ) works along the segment $R \in (17, 21)$ Å. Although the histograms seem to be Gaussian (dashed lines) they are not identical as predicted by the FR method (see text for details). . . . .	59

3.6	(a) Variation of the number of hydrogen bonds in deca-alanine during the quasi-equilibrium ( $v = 10^{-4}$ Å/ns) F and R pulls. (b) The PMF $U(R)$ calculated as the external work done during the quasi-equilibrium F (dashed line) and R (solid line) pulls. The discrepancy between the two PMFs is most likely due to the difference on how the H-bonds are formed and destroyed during the forced folding and unfolding processes, respectively, as indicated in the inset snapshots of the peptide. Dark (light) color corresponds to the R (F) process. . . . .	61
4.1	(a) Gramicidin A channel ( $\alpha$ -helix dimer colored in green) in POPE lipid bilayer (grey), solvated in water (van der Waals representation). Potassium ions (blue), and the oxygen (red) and hydrogen (white) atoms of water molecules are represented as spheres; (b) Cross-section of the gramicidin A channel. The $K^+$ ion (blue) and the water molecules move single file inside the pore. . . . .	67
4.2	Comparison of PMFs obtained for systems with fully restrained protein backbones, but with different electrolyte concentrations: $\sim 10$ mM (1 $K^+$ and 1 $Cl^-$ ) (thin solid line) and 200 mM (21 pairs of ions) (dotted line); PMFs of systems with z-restrained backbones, 200 mM electrolyte, but with different pulling protocols: along the z-direction only (dashed line) and along the channel axis (thick solid line). . . . .	70
4.3	Results from the “FR method” when calculating the potential of mean force $U(z)$ and the corresponding diffusion coefficient $D(z)$ (velocity of pulls is $v = 15$ Å/ns): (a) individual (thin lines) and mean (thick line) work obtained from forward pulls; (b) individual (thin lines) and mean (thick line) work obtained from reverse pulls (explanations in the text); (c) $U(z)$ (dashed line) and its symmetrized version (solid line); (d) mean dissipative work $\overline{W}_d(z)$ for the two intervals, $z \in (-15, 0)$ Å (dotted line) and $z \in (0, 15)$ Å (dashed line), and their arithmetic mean (solid line). $z = 0$ Å corresponds to the middle of the channel and $z = 15$ Å to its extremities. The linear slope of $\overline{W}_d(z)$ yields constant $D = 10.34$ Å <sup>2</sup> /ns. . . . .	71
4.4	Comparison of (a) potential of mean force and (b) mean dissipative work, obtained by applying on 10 SMD pulls the “FR method” (thick solid line) and the Jarzynski equality with the cumulant approximation (CA) (thick dashed line). The CA was employed in both forward (dotted lines) and reverse (dash-dotted lines) directions. . . . .	72
4.5	Potential of mean force (PMF) calculated using various pulling protocols for the FR method (black lines) compared to PMFs obtained from umbrella sampling (US) method ([2] (red line) and [1] (blue line)) and from Poisson-Nernst-Planck theory combined with MD simulations (PMFPNP) [111] (green bulleted line). . . . .	74

4.6	Cartoon representation of the simulated GlpF tetramer containing the glycerol molecules: the four GlpF channels in the middle are colored in green, red, blue and orange; the glycerol molecules inside the channels are colored in magenta; the lipids are shown in tan colored licorice representation; the water is represented in light blue spheres. The left panel is a top view of the system, while the panel on the right-hand side represents the side view of two of the channels. . . . .	78
4.7	(a) Definition of the angle $\psi$ . Let $\mathbf{a} = \mathbf{COM}(CG_1 + OG_1) - \mathbf{COM}(CG_3 + OG_3)$ define the orientation of the glycerol molecule (carbons in blue; oxygens in red; hydrogens not shown). Then $\psi$ is the angle that vector $\mathbf{a}$ forms with the z-axis. (b) Definition of the angle $\phi$ . Let $\mathbf{p} = \mathbf{COM}(CG_2) - \mathbf{COM}(CG_1 + CG_3)$ define a vector in the plane of the glycerol molecule. Then $\phi$ is the projection on the membrane plane of the angle that vector $\mathbf{p}$ forms with the vector passing through the COM of two consecutive channels. . . . .	80
4.8	(a) Comparison of the PMF calculated using the FR method (red line) with one reported from unidirectional pulls [5] (black line), for a glycerol molecule passing through a GlpF channel. The PMF presents a small well followed by a large barrier ( $\sim 16 k_B T$ ) in the selectivity filter region. (b) $\phi$ angle variation along the channel during the F and R pulls. In the selectivity filter, the glycerol can pass only under the angles $\phi \in \{10^\circ, 100^\circ, 190^\circ, 280^\circ\}$ . . . . .	83
4.9	(a) The mean dissipative work $\overline{W}_d$ and (b) its corresponding diffusion coefficient $D$ . The solid lines correspond to position dependent $\overline{W}_d(z)$ and its corresponding $D(z)$ (the horizontal solid line in the bottom panel is the average of $D(z)$ ). The dashed line reveals a constant diffusion coefficient $D \sim 8.6 \text{ \AA}^2/\text{ns}$ , corresponding to a linear fit of $\overline{W}_d(z)$ in the narrow region of the channel. The diffusion coefficient of glycerol in bulk (for $z < -30 \text{ \AA}$ ) was found to be $D_{\text{bulk}} \gtrsim 95 \text{ \AA}^2/\text{ns}$ . . . . .	85
5.1	Structure of a subunit of LH2 from <i>Rs. molischianum</i> (MOLI), from <i>Rps. acidophila</i> (ACI) and from an unknown species (X). Each subunit consists of one $\alpha$ polypeptide (orange ribbon), one $\beta$ -polypeptide (magenta ribbon), three BChls (green; phytol chains truncated), and one carotenoid (yellow). The N-terminal domains are on top, the C-terminal domains on bottom. The surface of the subunit is superimposed onto the simplified representations. . . . .	92
5.2	The primary structure of the light-harvesting polypeptides of LH2 from MOLI, ACI and X, aligned at the conserved His residue. Transmembrane helical domains are highlighted in yellow. . . . .	92
5.3	Definition of the reaction coordinates $\theta$ and $R$ illustrated for two subunits of LH2 from MOLI. (left) Top view from the N-terminal or cytoplasmic side in space-filling representation with transmembrane helices highlighted in ribbon representation. (right) Top view of two neighboring LH2 subunits' $\alpha$ and $\beta$ polypeptides with definition of the reaction coordinates. . . . .	96

5.4	Histograms of the reaction coordinates $\theta$ (left) and $R$ (right) corresponding to equilibrium MD simulations as follows: MOLI dimer, 2 ns run (thick solid line), ACI dimer, 2 ns run (thick dashed line), MOLI LH2 ring, 3 ns (thin solid line) and ACI LH2 ring, 3 ns (thin dashed line). . . . .	100
5.5	Two dimensional density plot of the volume overlap of the transmembrane protein regions of MOLI (left) and ACI (right) dimers along the reaction coordinates $\theta$ and $R$ recorded in the SMD simulations described in the text; the trajectories corresponding to the four different sets of simulations are indicated by numbered arrows. The volumes are relative to the value corresponding to the equilibrium reaction coordinates $\theta_0$ and $R_0$ (see text), marked by the small circle. . . . .	101
5.6	Calculated potentials of mean force $U(\theta R)$ for both MOLI (solid lines) and ACI (dashed lines) dimers. The thick (thin) curves correspond to $R = 18 \text{ \AA}$ ( $R = 25 \text{ \AA}$ ). . . . .	106
5.7	Evolution of the angle reaction coordinate $\theta$ during energy minimization of MOLI (top), ACI (center) and X dimers (bottom), respectively. The curves for each dimer correspond to initial $\theta_0$ angle values of $36^\circ$ (red line), $40^\circ$ (black line) and $45^\circ$ (blue line). . . . .	113
5.8	Distribution histogram (left panels) and time-evolution (right panels) of the angle reaction coordinate $\theta$ for MOLI (a,b,c), ACI (d,e,f) and X (g,h,i) dimers corresponding to free MD simulations starting from conformations characterized by $\theta_0 = 36^\circ$ (a,d,g), $\theta_0 = 40^\circ$ (b,e,h) and $\theta_0 = 45^\circ$ (c,f,i). For each case the listed mean value $\langle\theta\rangle$ was calculated using the last 7 ns of the corresponding MD trajectory. . . . .	115
5.9	Time dependence of the RMSD of the dimer transmembrane $C_\alpha$ atoms with respect to the crystal structure in case of MOLI (top) and ACI (center), and to the build structure in case of the X dimer (bottom). . . . .	116
5.10	Potential of mean force $U(\theta)$ calculated using the FR method described in the text for MOLI (dashed line), ACI (dash-dotted line) and X (solid line) dimers. . . . .	118
5.11	Distribution function $P(\theta)$ of the relative orientation angle $\theta$ between adjacent subunits (thin lines) corresponding to the eight subunit MOLI (top panel) and nine subunit ACI (bottom panel) LH2 rings. The distribution histograms for were determined from 3 ns long equilibrium MD trajectories. The average angle distributions for all subunit pairs are also shown (solid lines). . . . .	119

# List of Abbreviations

1D	one-dimensional
ACI	<i>Rhodopseudomonas acidophila</i>
AFM	atomic force microscopy
AQP	aquaporin
AQPZ	aquaporin Z
Asp	aspartate
ATP	adenosine triphosphate
B800	bacteriochlorophyll-a molecule that absorbs at 800 nm
B850	bacteriochlorophyll-a molecule that absorbs at 850 nm
BChl	bacteriochlorophyll
CA	cumulant approximation
CD	circular dichroism
CIS	configuration interaction singles
COM	center of mass
CT	C-terminus
cv-SMD	constant velocity steered molecular dynamics
DOS	density of states
EMD	equilibrium molecular dynamics
F	forward
FWHM	full width at half maximum
GlpF	glycerol uptake facilitator

HF	Hartree-Fock
HGP	harmonic guiding potential
His	histidine
HOMO	highest occupied molecular orbital
JE	Jarzynski equality
Leu	leucine
LH	light harvesting
LH1	light harvesting complex 1
LH2	light harvesting antenna complex 2
LUMO	lowest unoccupied molecular orbital
MD	molecular dynamics
MFPT	mean first passage time
MLE	maximum likelihood estimator
MOLI	<i>Rhodospirillum rubrum</i>
NMR	nuclear magnetic resonance
NT	N-terminus
OD	linear absorption
PDB	protein data bank
PME	particle mesh Ewald
PMF	potential of mean force
PMFPNP	potential of mean force calculation based on Poisson-Nernst-Planck theory
POPC	1-Palmitoyl, 2-Oleoyl-sn-glycero 3-PhosphoCholine
POPE	1-Palmitoyl-2-Oleoyl-sn-glycero-3-PhosphoEthanolamine
PPC	pigment-protein complex
Q <sub>y</sub>	first optically active excited state
QC	quantum chemistry
R	reverse

<i>Rb.</i>	<i>Rhodobacter</i>
RC	reaction coordinate
RC-LH1	reaction center-light harvesting complex 1
RMSD	root mean square deviation
RNA	ribonucleic acid
<i>Rps.</i>	<i>Rhodopseudomonas</i>
<i>Rs.</i>	<i>Rhodospirillum</i>
SF	selectivity filter
SMD	steered molecular dynamics
SSA	stiff-spring approximation
TDM	transition dipole moment
TFT	transient fluctuation theorem
TM	trans-membrane
TMH	trans-membrane helical
US	umbrella sampling
WHAM	weighted histogram analysis method
ZINDO/S	Zerner's semiempirical intermediate neglect of differential overlap method parametrized for spectroscopy

# Abstract

Studies of structure-function relationships in biomolecular systems require to follow nanometer-size systems on time scales spanning from pico- to micro-seconds, while maintaining atomic scale spatial resolution in all-atom molecular dynamics (MD) simulations. In this work we propose new methods to investigate the following, intrinsically multiscale problems: (i) theoretical prediction of optical and spectral properties of pigment-protein complexes, (ii) reconstruction of potential of mean force and its corresponding diffusion coefficient from non-equilibrium molecular dynamics simulations, (iii) transport of potassium ion through the Gramicidin A channel and of glycerol through the GlpF channel, and (iv) prediction of the species-dependent oligomerization state of the light harvesting antenna complexes. The main novelty of these methods is that they rely only on the high resolution atomic structure of the biomolecular system. Therefore, they have not only explanatory, but predictive power as well.



# **1 Introduction**

Theoretical studies of structure-function relationships in biomolecular systems became recently feasible due to recent developments of cluster and grid computing technologies. These studies are notoriously difficult because they require following nanometer-size systems on time scales spanning from pico- to micro-seconds, while maintaining atomic scale spatial resolution in all-atom molecular dynamics (MD) simulations. Covalent bond vibrations in MD simulations call for femtosecond time steps. Thus, it is not surprising that the scientific community has developed great interest in finding new and practical methods to investigate these intrinsically multiscale problems. Although the systems studied here are rather different (e.g. photosynthetic light harvesting systems, protein channels), the unifying theme of the thesis is the multiscale modeling approach. The main novelty of the developed methods is that they rely only on the high resolution atomic structure of the biomolecular system. Hence, the obtained results have not only explanatory, but also predictive power.

The thesis covers six inter-related projects that are organized in four self-contained chapters, as follows. In Chp. 2 we propose a novel and general approach for calculating optical and spectral properties of pigment-protein complexes (PPCs). The method, based on a combination of molecular dynamics (MD) simulations, quantum chemistry (QC) calculations, and statistical mechanical modeling, is demonstrated by calculating the linear absorption (OD) and circular dichroism (CD) spectra of bacteriochlorophyll (BChl) aggregates in LH2 light harvesting antenna complexes from the purple bacterium *Rhodospirillum rubrum*. The results are found to be in good agreement with the experimental data. Since the method requires no fitting parameters, in principle, it can be applied to predict optical and spectral properties of any PPC with known high resolution atomic structure.

In Chp. 3, we describe a new method of investigation of biological processes that occur on time scales exceeding by several orders of magnitude the few tens of nanoseconds attainable by all-atoms MD simulations, e.g. molecular and ion transport through protein channels. In such cases, a simplified approach is to describe the transported molecule as

an overdamped Brownian particle that diffuses along the channel axis in the presence of an effective potential of mean force (PMF), which describes the interaction of the particle with the rest of the atoms in the system. A PMF is the Landau free energy profile along a reaction coordinate, e.g. channel axis. The reaction coordinate can be sampled efficiently only by steered MD (SMD) simulations, where the system is guided along the reaction coordinate. The proposed method, referred to as the “FR method”, allows one to calculate simultaneously the PMF and the corresponding diffusion coefficient (from the mean dissipative work) along the reaction coordinate, by employing a small number of fast non-equilibrium pulls in both forward and time reversed directions. The viability of the FR method is first tested on a small, but non-trivial system, the deca-alanine. The computed PMF is found to be in good agreement with previous results obtained by using either the Jarzynski’s equality or the umbrella sampling method. However, compared to these PMF calculation methods this new approach is about one order of magnitude faster.

In Chp. 4 we apply the FR method to study the transport of the potassium ion through the Gramicidin A ion channel and of the glycerol through the GlpF aquaglyceroporin channel. Gramicidin A is helical antibiotic that dimerizes in the lipid membrane into a cation selective channel. In spite of its structural simplicity, the energetics of the ion transport through the channel is not trivial. Most of the difficulty comes from the sensitivity to errors due to finite-size effect and from poor description of polarization effects by the existing molecular dynamics force-fields. The PMF calculated using the FR method reproduces the PMFs reported by umbrella sampling method [1–3], which have a large central barrier and two wells at the entrances in the channel. In order to account for the polarization effects, the charge of the potassium ion is reduced to half. The recalculated PMF profile yields barrier and well sizes very similar to the ones that would reproduce the experimental conductance measurements [4].

The *E. coli* glycerol uptake facilitator (GlpF) channel selectively transports both water and glycerol molecules, excluding charged solutes, including protons. The high resolution

structure of GlpF has provided much insight into the underlying microscopic mechanism of molecular transport and selectivity through GlpF. In particular, MD simulations have revealed the underlying microscopic mechanism of single file water and glycerol transport through GlpF. The potential of mean force (PMF) of glycerol in GlpF was also calculated by means of non-equilibrium SMD simulations by employing the non-equilibrium work relationship due to Jarzynski's equality [5]. The reported PMF was asymmetric, exhibiting a prominent potential well at the periplasmic side and several potential barriers, with the largest at the selectivity filter. The precise form of this PMF came under scrutiny once it was recognized that in practice Jarzynski's equality is applicable only sufficiently close to equilibrium, while the SMD simulations employed in PMF calculations usually drive the system far from equilibrium. Both the PMF and the corresponding diffusion coefficient of a glycerol molecule through GlpF are determined by employing the FR method. They are used to calculate the mean first passage time of glycerol through GlpF and the permeability of the channel to glycerol. The obtained PMF is compared to the one reported in [5].

In Chp. 5 we describe different approaches used to predict the oligomerization state of light harvesting antenna complexes from three different species of photosynthetic bacteria, from the sole knowledge of the high resolution crystal structure of a single subunit. These complexes are rings formed by the self-assembly of a well defined, species-dependent number of identical protein-chromophore subunits. In order to probe the assembly process of two subunits, two sets of calculations with different philosophies are performed. In one set of calculations, the subunits are constrained at different angles, allowing the systems to equilibrate under these constraints. Using umbrella sampling, information about the free energy profile as a function of angle is extracted. Pulling two subunits apart through SMD simulations gives complementary information into key events of the binding/unbinding process and a rough order of interaction strengths. The results suggest that the preferred angle between two subunits plays an important role in guiding the assembly of light harvesting complexes into a particular oligomerization state. For each of the species, the

preferred angle of a dimer is extracted from three sets of calculations, performed on dimer systems built at three different initial angle values: energy minimizations, free MD simulations, and PMF calculations along the angle reaction coordinate, using the FR method. Results indicate that all the species will most likely structure into a 9-ring light harvesting system.

## **2 Theoretical prediction of optical and spectral properties of pigment-protein complexes**

## 2.1 Introduction

Pigment-protein complexes (PPCs) play an important role in photosynthetically active biological systems and have been the subject of numerous experimental and theoretical studies [7]. In a PPC the photoactive pigment molecules are held in well defined spatial configuration and orientation by a scaffold of proteins. The spectral and optical properties of PPCs are determined by (i) the chemical nature of the pigment, (ii) the electronic interactions between the pigment molecules, and (iii) the interactions between pigment molecules and their environment (e.g., protein, lipid, solvent). Since in biological systems PPCs exist and function at physiological temperature, their electronic and optical properties are strongly affected by thermal fluctuations, which represent the main source of dynamic disorder in these systems.

Unfortunately, even in the simplest theoretical models of PPCs the simultaneous treatment of the electronic coupling between the pigments and the effect of thermal disorder can be done only approximately [8–11]. This chapter presents the formulation and implementation of an efficient method for calculating spectral and optical properties, e.g., linear absorption (OD) and circular dichroism (CD) spectra of PPCs at finite temperature, by using only atomic structure information. To demonstrate its usefulness, the proposed method is applied to calculate the OD and CD spectra at room temperature of the aggregate of bacteriochlorophyll-a (BChl-a) molecules in the light harvesting antenna complex (LH2) from the purple bacterium *Rhodospirillum (Rs.) molischianum*. Following their crystal structure determination, LH2 complexes from *Rs. molischianum* [12] and *Rhodopseudomonas (Rps.) acidophila* [13] have been extensively studied both experimentally [14–21] and theoretically [14–16, 22–29]. In *Rs. molischianum* the LH2 is an octamer of  $\alpha\beta$ -heterodimers, arranged in a ring-like structure [24, 30] (see Fig. 2.1). Each protomer consists of an  $\alpha$ - and a  $\beta$ -apoprotein that non-covalently bind one BChl-a molecule that

---

This chapter is based on published article [6].

absorbs at 800 nm (referred to as B800), two BChl-a molecules that absorb at 850 nm (referred to as B850) and at least one carotenoid that absorbs around 500 nm. The total of 16 B850 and 8 B800 BChls form two circular aggregates, both oriented parallel to the surface of the membrane. The excitonic coupling between the B800s is negligible because of their large spatial separation ( $\sim 22$  Å). Therefore, the optically active  $Q_y$  excited electronic states of the B800s are almost degenerate. On the other hand, the tightly packed B850s (with an average Mg–Mg distance of  $\sim 9.2$  Å within the  $\alpha\beta$ –heterodimer and  $\sim 8.9$  Å between the neighboring protomers) are strongly coupled and the corresponding  $Q_y$  excited states form an excitonic band in which the states that carry most of the oscillator strength are clustered about  $\sim 850$  nm (1.46 eV). Another important difference between the two BChl rings is that while the B800s are surrounded by mostly hydrophilic protein residues, the binding pocket of the B850s is predominantly hydrophobic [12]. Thus, although both B800s and B850s are chemically identical BChl-a molecules, their specific spatial arrangement and the nature of their immediate protein surrounding alter differently their spectral and optical properties. For example, it is quite surprising that the two peaks, due to the B800 and B850 BChls, in the experimental OD spectrum of LH2 from *Rs. molischianum* at room temperature [25, 31], have comparable widths although, as mentioned above, the B800 levels are almost degenerate while the B850 levels form a  $\sim 0.2$  eV wide excitonic band.

Clearly, novel methods for calculating optical spectra of PPCs by using computer simulations based entirely on the atomic structure of the system would not only provide a better understanding and interpretation of the existing experimental results but would also help in predicting and designing new experiments. The standard procedure to simulate the experimental spectra of LH2 systems (and PPCs in general) consists of two steps [9–11]. First, the excitation energy spectrum is determined based on the static crystal structure of the system and, second, the corresponding stick spectrum is “dressed up” with simulated Gaussian (in case of static disorder) and/or Lorentzian (in case of dynamic disorder) line widths char-



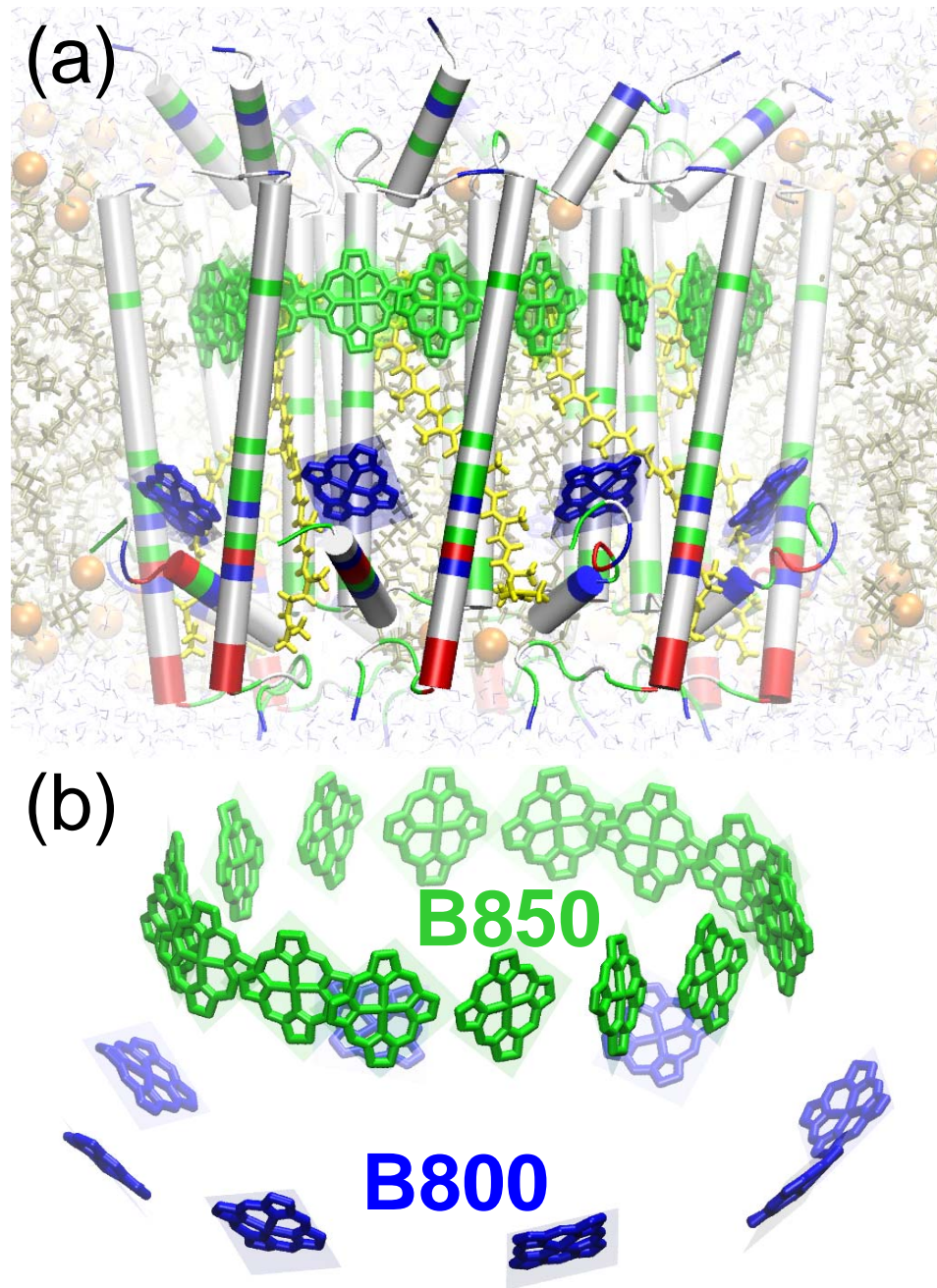


Figure 2.1: (a) Side view of an octameric LH2 complex from *Rs. molischianum* embedded in a fully solvated POPC lipid bilayer. The transmembrane helices of the apoprotein subunits are shown as cylinders (cartoon representation) and are colored by residue type; dark (light) colors represent hydrophilic (hydrophobic) residues. For clarity only the BChl macrocycles are shown and the front half of the lipids are not shown. The clearly visible B800 (B850) ring is surrounded mostly by polar and charged (nonpolar) protein residues. (b) Tilted side view of the quantum system formed by the optically active B800 and B850 macrocycles that form rings oriented parallel to the surface of the membrane.

acterized by empirically (and often self consistently) determined parameters. Alternatively, the spectral broadening of the stick spectrum can also be described through the coupling of the electronic excitations (modeled as two- or multilevel-systems) to a stochastic heat bath characterized by a model spectral density with empirical parameters. While either approach may yield excellent agreement between the simulated and experimental spectra, the empirical nature of the model parameters restricts their predictive power.

The proposed method for calculating optical spectra is based on a combination of all atom molecular dynamics (MD) simulations, quantum chemistry (QC) calculations and quantum many-body theory. The conformational dynamics of the LH2 ring embedded into its natural environment (a fully solvated lipid bilayer) are followed by means of classical MD simulations. Next, for each BChl, modeled as a quantum two level system, the  $Q_y$  excitation energy gap and transition dipole moment time series are determined along a properly chosen segment of the MD trajectory by means of QC calculations. Finally, the OD and CD spectra are determined as weighted sums of the Fourier transform of the quantum dipole-dipole correlation function (i.e., the absorption *lineshape function*) which, within the cumulant approximation, can be calculated from the sole knowledge of the energy gap time series. Formally, this method can also be regarded as a two step procedure. First, a stick spectrum is generated from the average values of the energy gap time series and, second, spectral broadening is applied through the corresponding lineshape function weighted by the mean transition dipole (rotational) strength in the case of OD (CD) spectrum. Since both peak position and broadening of the optical spectrum are obtained from the same energy gap time series determined from combined MD/QC calculations, the proposed method requires no empirical fitting parameters, making it ideal for predicting optical spectra of PPCs with known structure. Similar MD/QC methods were used previously by [32] for calculating the OD spectrum of BChl-a in methanol, and by [22] to determine the OD and CD spectra of B850s in LH2 from *Rs. molischianum*. The relationship between these studies and the present one will be established below.

## 2.2 Optical Transitions at Finite Temperature

In order to calculate the linear optical absorption of a PPC, we assume that the electronic properties of individual pigment molecules can be described in terms of a two-level system, formed by the ground state and the lowest excited singlet state (e.g., the  $Q_y$  state in the case of BChl-a) involved in the optical absorption process. Neglecting for the moment the direct interaction between the pigments, these two states for the  $n^{th}$  pigment ( $n = 1, \dots, N$ ) are denoted as  $|0\rangle \equiv |0_n\rangle$  and  $|n\rangle \equiv |1_n\rangle$ , respectively. Once the interaction between the pigment and its environment (composed of protein matrix, lipid membrane and solvent molecules) is taken into account, these two levels turn into, still well separated, energy bands  $|0; \lambda_0\rangle = |0\rangle|\lambda_0\rangle$  and  $|n; \lambda_n\rangle = |n\rangle|\lambda_n\rangle$ , where the quantum numbers  $\lambda_0$  and  $\lambda_n$  specify the state of the  $n^{th}$  pigment on the ground- and excited-state potential energy surface, respectively. Because the exact quantum mechanical treatment of the eigenstates  $|0; \lambda_0\rangle$ ,  $|n; \lambda_n\rangle$  and of the corresponding energy eigenvalues  $\mathcal{E}_{0, \lambda_0}$ ,  $\mathcal{E}_{n, \lambda_n}$  is not feasible, usually the quantum numbers  $\lambda_0$  and  $\lambda_n$  are associated with the vibronic states of the PPC, that can be treated within the harmonic approximation. A different approach, in which the dynamics of the nuclear degrees of freedom of the PPC are described by means of all-atom MD simulations, and the energy gap time series  $\Delta E_n(t) = \mathcal{E}_n(t) - \mathcal{E}_0(t)$  is calculated at each MD time step by QC calculations, is presented here. The main assumption of this approach is that the obtained energy gap time series  $\Delta E_n(t)$  can be used to calculate approximately equilibrium quantities (such as energy gap density of states and time autocorrelation functions) of the original system without the knowledge of the exact energy gap spectrum  $\Delta \mathcal{E}_{n, \lambda_n, \lambda_0} = \mathcal{E}_{n, \lambda_n} - \mathcal{E}_{0, \lambda_0}$ .

In the absence of the excitonic coupling between the pigment molecules, the Hamiltonian of the system can be written as  $\mathcal{H} = H_0 + H$ , where

$$H_0 = \sum_{\lambda_0} |0; \lambda_0\rangle \mathcal{E}_{0, \lambda_0} \langle 0; \lambda_0|, \quad (2.1a)$$

and

$$H = \sum_n H_n = \sum_{\lambda_n} |n; \lambda_n\rangle \mathcal{E}_{n, \lambda_n} \langle n; \lambda_n|. \quad (2.1b)$$

The dipole moment operator through which the incident light field couples to the  $n^{\text{th}}$  pigment is given by

$$\hat{\boldsymbol{\mu}}_n = \sum_{\lambda_n, \lambda_0} \mathbf{d}_{n, \lambda_n, \lambda_0} |n; \lambda_n\rangle \langle 0; \lambda_0|, \quad (2.2a)$$

where the transition dipole moment (TDM) matrix element  $\mathbf{d}_{n, \lambda_n, \lambda_0}$  within the Condon approximation [10] can be written

$$\mathbf{d}_{n, \lambda_n, \lambda_0} \approx \mathbf{d}_n \langle \lambda_n | \lambda_0 \rangle. \quad (2.2b)$$

Here  $\mathbf{d}_n = \langle 1 | \hat{\boldsymbol{\mu}}_n | 0 \rangle$  is the real TDM vector, whose time series can be determined from the same MD/QC calculations as  $\Delta E_n(t)$ . While  $\langle 1 | 0 \rangle = 0$ , in general the Franck-Condon factors  $\langle \lambda_n | \lambda_0 \rangle$  are finite [10].

### 2.2.1 Linear Absorption and Lineshape Function

Since the size of the PPC is much smaller than the wavelength of the light field, the latter can be approximated as homogeneous throughout the system and, according to standard linear response theory, the corresponding OD spectrum is proportional to the dipole-dipole correlation function

$$I(\omega) \propto \omega \sum_{n,m} \text{Re} \left[ \int_0^\infty dt e^{i\omega t} \langle \hat{\mu}_{m,i}^\dagger(0) \hat{\mu}_{n,i}(t) \rangle \right], \quad (2.3)$$

where  $\hat{\mu}_{n,i}(t) = e^{-iHt} \hat{\mu}_{n,i}(0) e^{iH_0 t}$  is the  $i \in \{x, y, z\}$  component of the time dependent electric dipole operator, and  $\langle \dots \rangle = \text{Tr} \{ Z_0^{-1} \exp(-\beta H_0) \dots \}$  with  $\beta = 1/k_B T$  the usual temperature factor and  $Z_0$  the corresponding partition function. To simplify notation, throughout this paper we use units in which  $\hbar = 1$ , and apply the convention of implicit summation over

repeated vector indices. By employing Eqs. (2.1)-(2.2), after some algebra, the quantum dipole correlation function in Eq. (2.3) can be expressed as

$$\langle \hat{\mu}_{m,j}^\dagger(0) \hat{\mu}_{n,i}(t) \rangle = d_{n,i} d_{m,j} \delta_{nm} \langle e^{iH_0 t} e^{-iH_n t} \rangle, \quad (2.4)$$

where  $\delta_{nm}$  is the Kronecker delta. By inserting Eq. (2.4) into Eq. (2.3) one obtains the sought OD spectrum of an aggregate of noninteracting pigments in their native environment

$$I(\omega) \propto \omega \sum_n d_n^2 A_n(\omega), \quad (2.5a)$$

where the *lineshape function* is defined as

$$A_n(\omega) = \text{Re} \int_0^\infty dt e^{i\omega t} \langle e^{iH_0 t} e^{-iH_n t} \rangle. \quad (2.5b)$$

The main difficulty in calculating the quantum time correlation function in Eq. (2.5b) is due to the fact that the Hamiltonians  $H_0$  and  $H_n$  do not commute. If they would, then the lineshape function could be expressed in terms of the energy gap density of states (DOS). Indeed, in this case  $\langle e^{iH_0 t} e^{-iH_n t} \rangle \approx \langle \exp(-i\Delta H_n t) \rangle$ , with  $\Delta H_n = H_n - H_0$ , and by calculating the time integral in Eq. (2.5b) would follow

$$A_n(\omega) \approx \pi \mathcal{N}(\omega), \quad (2.6a)$$

$$\mathcal{N}(\omega) \equiv \langle \delta(\omega - \Delta H_n) \rangle \approx \langle \delta(\omega - \Delta E_n(t)) \rangle, \quad (2.6b)$$

where the density of states  $\mathcal{N}(\omega)$  is approximated by the binned histogram of the energy gap fluctuations  $\Delta E_n(t)$ , obtained from combined MD/QC calculations. The Fourier transform of the exact spectral representation of the correlation function

$$\langle e^{-iH_0 t} e^{iH t} \rangle = \sum_{\lambda_0, \lambda_n} \rho_{\lambda_0} |\langle \lambda_0 | \lambda_n \rangle|^2 e^{-i(\mathcal{E}_{n, \lambda_n} - \mathcal{E}_{0, \lambda_0})t}, \quad (2.7a)$$

where  $\rho_{\lambda_0} = Z_0^{-1} \exp(-\beta \mathcal{E}_{0,\lambda_0})$  is the statistical matrix of the electronic ground state, yields

$$A(\omega) = 2\pi \sum_{\lambda_0, \lambda_n} \rho_{\lambda_0} |\langle \lambda_0 | \lambda_n \rangle|^2 \delta(\omega - \Delta \mathcal{E}_{n,\lambda_n,\lambda_0}), \quad (2.7b)$$

which can be regarded as a Franck-Condon weighted and thermally averaged density of state [10]. By setting the Franck-Condon factors  $\langle \lambda_0 | \lambda_n \rangle$  equal to unity in Eq. (2.7b) one obtains Eqs. (2.6). Since it is not possible to determine all these factors, it is often convenient to use Eqs. (2.6) as approximation of  $A_n(\omega)$  for calculating the OD spectrum, even though they usually overestimate its broadening.

A systematic way of calculating the correlation function in Eq. (2.5b) is the cumulant expansion method. Employing the second order cumulant approximation, often used in optical spectra calculations [9], we obtain

$$\begin{aligned} \langle e^{iH_0 t} e^{-iH_n t} \rangle &= \left\langle T \exp \left[ -i \int_0^t d\tau \Delta H_n(\tau) \right] \right\rangle \\ &\approx \exp \left[ -i \langle \Delta H_n \rangle t - \int_0^t d\tau (t - \tau) \mathcal{C}_n(\tau) \right], \end{aligned} \quad (2.8)$$

where  $T$  is the time ordering operator,  $\Delta H_n(t) = e^{iH_0 t} \Delta H_n e^{-iH_0 t}$ ,  $\mathcal{C}_n(t) = \langle \delta H_n(t) \delta H_n(0) \rangle$ , and  $\delta H_n(t) = \Delta H_n(t) - \langle \Delta H_n \rangle$ . To make progress, the quantum statistical averages in Eq. (2.8) will be approximated with classical ones involving the energy gap time series  $\Delta E_n(t)$ , i.e.,

$$\langle \Delta H_n \rangle \approx \langle \Delta E_n(t) \rangle \equiv \omega_n, \quad (2.9a)$$

$$\text{Re}[\mathcal{C}_n(t)] \approx C_n(t) \equiv \langle \delta E_n(t) \delta E_n(0) \rangle, \quad (2.9b)$$

where  $\delta E_n(t) = \Delta E_n(t) - \langle \Delta E_n \rangle$ . While approximating a quantum time correlation function by identifying its real part with the corresponding classical correlation function as in Eq. (2.9b) is widely used [32–34], other approximation schemes have also been considered in the literature [35]. Next, by invoking the *fluctuation dissipation theorem*  $\tilde{\mathcal{C}}_n(-\omega) =$

$\exp(-\beta\omega)\tilde{\mathcal{C}}_n(\omega)$ , where  $\tilde{\mathcal{C}}_n(\omega) = \int_{-\infty}^{\infty} dt \mathcal{C}_n(t) \exp(i\omega t)$  is the Fourier transform of  $\mathcal{C}_n(t)$ , the quantum correlation function in terms of the real *spectral density*

$$J_n(\omega) = \frac{1}{2} \left[ \tilde{\mathcal{C}}_n(\omega) - \tilde{\mathcal{C}}_n(-\omega) \right] = \frac{1}{2} \left( 1 - e^{-\beta\omega} \right) \tilde{\mathcal{C}}_n(\omega) \quad (2.10)$$

can be written as

$$\mathcal{C}_n(t) = \mathcal{C}'_n(t) - i\mathcal{C}''_n(t) = \int_0^{\infty} \frac{d\omega}{\pi} J_n(\omega) [\coth(\beta\omega/2) \cos \omega t - i \sin \omega t] \quad (2.11)$$

By identifying the real part of Eq. (2.11) with Eq. (2.9b) one can determine both the spectral density and the imaginary part of the quantum correlation function, i.e.,

$$J_n(\omega) = 2 \tanh(\beta\omega/2) \int_0^{\infty} dt C_n(t) \cos \omega t, \quad (2.12)$$

and

$$\mathcal{C}''_n(t) = \int_0^{\infty} \frac{d\omega}{\pi} J_n(\omega) \sin \omega t. \quad (2.13)$$

Thus, the lineshape function within the second cumulant approximation is

$$A_n(\omega) \equiv \bar{A}_n(\omega - \omega_n) = \int_0^{\infty} dt e^{-\phi_n(t)} \cos[(\omega - \omega_n)t + \varphi_n(t)], \quad (2.14a)$$

where the broadening and frequency shift functions are given by

$$\phi_n(t) = \int_0^{\infty} d\tau (t - \tau) C_n(\tau), \quad (2.14b)$$

and

$$\varphi_n(t) = \int_0^{\infty} d\omega J_n(\omega) \frac{\omega t - \sin \omega t}{\omega^2}. \quad (2.14c)$$

A straightforward extension of the above method for calculating the lineshape function and the OD spectrum of  $N$  excitonically coupled pigment molecules would require the

determination of the energies  $\mathcal{E}_{J,\lambda_J}$  and TDMs  $\mathbf{d}_J$  corresponding to the excitonic states  $|J;\lambda_J\rangle$ ,  $J = 1, \dots, N$ . Unfortunately, the required QC calculations (by considering all  $N$  pigments as a single quantum system) are still prohibitively expensive computationally. Therefore, we employed an *effective Hamiltonian* approximation for determining the time series  $\Delta E_J(t) = \mathcal{E}_J(t) - \mathcal{E}_0(t)$  and  $\mathbf{d}_J(t)$  from  $\Delta E_n(t)$  and  $\mathbf{d}_n(t)$  of the individual pigments. Assuming that these are coupled through the usual point dipole-dipole interaction

$$V_{nm} = \frac{1}{4\pi\epsilon_0\epsilon_r} \left[ \frac{\mathbf{d}_n \mathbf{d}_m}{r_{nm}^3} - 3 \frac{(\mathbf{d}_n \cdot \mathbf{r}_{nm})(\mathbf{d}_m \cdot \mathbf{r}_{nm})}{r_{nm}^5} \right], \quad (2.15)$$

where  $\epsilon_r$  is the relative dielectric permittivity of the medium,  $\mathbf{r}_n$  is the position vector of pigment  $n$ , and  $\mathbf{r}_{nm} = \mathbf{r}_m - \mathbf{r}_n$ , the eigenvalue equation one needs to solve at every MD timestep is

$$\sum_m [(\Delta E_n \delta_{nm} + V_{nm}) - \Delta E_J \delta_{nm}] c_m^{(J)} = 0. \quad (2.16)$$

In term of the coefficients  $c_n^{(J)} = \langle J|n\rangle$  the excitonic TDMs are

$$\mathbf{d}_J = \sum_n \langle J|n\rangle \mathbf{d}_n. \quad (2.17)$$

Next, by rewriting the Hamiltonian (2.1b) in diagonal form (i.e., in terms of noninteracting excitons)  $H = \sum_J H_J = \sum_{J,\lambda_J} |J;\lambda_J\rangle \mathcal{E}_{J,\lambda_J} \langle J;\lambda_J|$ , after some algebra one arrives at the equations

$$\left\langle \hat{\mu}_{m,j}^\dagger(0) \hat{\mu}_{n,i}(t) \right\rangle = \sum_J \langle J|n\rangle d_{n,i} d_{m,j} \langle m|J\rangle \left\langle e^{iH_0 t} e^{-iH_J t} \right\rangle, \quad (2.18a)$$

and

$$\sum_{n,m} \left\langle \hat{\mu}_{m,j}^\dagger(0) \hat{\mu}_{n,i}(t) \right\rangle = \sum_J d_{J,i} d_{J,j} \left\langle e^{iH_0 t} e^{-iH_J t} \right\rangle. \quad (2.18b)$$

Inserting Eq. (2.18b) into Eq. (2.3) one obtains the desired OD spectrum of the excitonic system

$$I(\omega) \propto \omega \sum_J d_J^2 A_J(\omega), \quad (2.19)$$



where

$$A_J(\omega) = \text{Re} \int_0^\infty dt e^{i\omega t} \langle e^{iH_0 t} e^{-iH_J t} \rangle. \quad (2.20)$$

We note that by replacing the site index  $n$  with the excitonic index  $J$  most of the above results for noninteracting pigments remain formally valid for the corresponding excitonic system as well. For example, similar expressions to (2.6) and (2.14) can be easily derived for estimating  $A_J(\omega)$ .

### 2.2.2 Circular Dichroism

By definition, the CD spectrum  $I_{CD}(\omega)$  is the difference between  $I_L(\omega)$  and  $I_R(\omega)$ , the OD spectra for left and right circularly polarized light, respectively. Unlike in the case of the OD spectrum, the calculation of  $I_{CD}(\omega)$  even within the leading order approximation requires taking into account the spatial variation of the light field across the PPC as well as the excitonic coupling between the pigment molecules, regardless how small this may be. The sensitivity of the CD spectrum to geometrical and local details of the PPC makes it a quantity difficult to predict by theoretical modeling. The CD spectrum is given by [11]

$$I_{CD}(\omega) = \frac{1}{4} [I_L(\omega) - I_R(\omega)] \propto \omega \text{Re} \int_0^\infty dt e^{i\omega t} \times \sum_{n,m} \frac{\pi}{\lambda} \epsilon_{ijk}(\mathbf{r}_n)_k \langle \hat{\mu}_{m,i}^\dagger(0) \hat{\mu}_{n,i}(t) \rangle \quad (2.21)$$

where  $\lambda$  is the wavelength of the incident light and  $\epsilon_{ijk}$  is the unit antisymmetric tensor of rank 3. Inserting Eq. (2.18a) into (2.21) and making use of Eq. (2.20), we obtain

$$I_{CD}(\omega) \propto \omega \sum_J R_J A_J(\omega), \quad (2.22a)$$

where

$$R_J = \frac{\pi}{\lambda} \sum_{n,m} \langle J|n \rangle [\mathbf{r}_n \cdot (\mathbf{d}_n \times \mathbf{d}_m)] \langle m|J \rangle \quad (2.22b)$$

is the so-called *rotational strength* of the excitonic state  $J$ . Note, that in the absence of the excitonic coupling all  $R_J = 0$  (because for a given  $J$  only one coefficient  $\langle J|n \rangle$  is nonzero) and the CD spectrum vanishes. The rotational strength plays the same role for the CD spectrum as the TDM strength for the OD spectrum. Specifically,  $R_J$  gives the coupling between the TDM of the excitonic state  $J$  and the orbital magnetic moment of the other excitons. The coupling to the local magnetic moment is assumed to be small (Cotton effect) and usually is discarded [11, 20].

## 2.3 Simulations of Optical Spectra

In order to apply the results derived in section 2.2 for calculating the OD and CD spectra of the B800 and B850 BChls in a single LH2 ring for *Rs. molischianum* the time series of the  $Q_y$  energy gap  $\Delta E_n(\ell\Delta t)$  and TDM  $\mathbf{d}_n(\ell\Delta t)$ ,  $\ell = 0, 1, \dots, N_t$ , for all individual BChls need to be determined. After building the computational model for the LH2 ring, we use all atom MD simulations to follow the dynamics of the nuclear degrees of freedom, by recording snapshots of the atomic coordinates at times  $t_\ell = \ell\Delta t$ . Then, QC calculations are used to compute  $\Delta E_n$  and  $\mathbf{d}_n$  for each of these snapshots.

### 2.3.1 Building the LH2 Rings

A perfect 8-fold LH2 ring was constructed starting from the crystal structure (pdb code 1LGH) of *Rs. molischianum* [12]. After adding the missing hydrogens, the protein system was embedded in a fully solvated POPC<sup>†</sup> lipid bilayer of hexagonal shape. Finally, a total of 16 Cl<sup>-</sup> counterions were properly added to ensure charge neutrality of the entire system of 87,055 atoms. In order to reduce the finite-size effects, the hexagonal unit cell (with side length  $\sim 60\text{\AA}$ , lipid bilayer thickness  $\sim 42\text{\AA}$  and two water layers of combined thickness  $\sim 35\text{\AA}$ ) was replicated in space by using periodic boundary conditions.

---

<sup>†</sup> Abbreviation for 1-Palmitoyl, 2-Oleoyl-sn-glycero 3-PhosphoCholine

### 2.3.2 Molecular Dynamics Simulations

The CHARMM27 force field parameters for proteins [36, 37] and lipids [38] were used. Water molecules were modeled as TIP3P [39]. The force field parameters for BChls and lycopenes were the ones used in [22]. After energy minimization, the system was subjected to a 2 ns long equilibration in the NPT ensemble [40] at normal temperature ( $T = 300$  K) and pressure ( $p = 1$  atm), using periodic boundary conditions and treating the full long-range electrostatic interactions by the particle mesh Ewald (PME) method [41]. All MD simulations were performed with the program NAMD 2.5 [42], with a performance of  $\sim 8.5$  days/ns on 24 CPUs of an AMD 1800+ Beowulf cluster. During equilibration an integration time step of 2 fs was employed by using the SHAKE constraint (bond lengths are set to their equilibrium values) on all hydrogen atoms [43]. After the 2 ns equilibration, a 1 ps production run with 1 fs integration step was carried out with atomic coordinates saved every other timestep, resulting in  $N_t = 500$  MD snapshots with  $\Delta t = 2$  fs time separation. These configuration snapshots were used as input for the QC calculations described below.

### 2.3.3 Quantum Chemistry Calculations

The time series of the  $Q_y$  transition energies  $\Delta E_n$  and dipole moments  $\mathbf{d}_n$  of individual BChls can be determined only approximately from the configuration snapshots obtained from MD simulations. The level of approximation used is determined by: (i) the actual definition of the optically active *quantum system*, i.e., the part of the system that is responsible for light absorption and needs to be treated quantum mechanically; (ii) the actual choice of the QC method used in the calculations; and (iii) the particular way in which the effect of the (classical) environment on the quantum system is taken into account in the QC calculations.

Because the optical properties of BChls are determined by the cyclic conjugated  $\pi$ -

electron system of the macrocycle, the quantum system was restricted to a truncated structure of the BChl-a containing 49 atoms in the porphyrin plane. The truncation consisted in removing the phytyl tail and in replacing the terminal  $\text{CH}_3$  and  $\text{CH}_2\text{CH}_3$  groups on the macrocycle with H atoms in order to satisfy valence requirements. Similar truncation schemes have been employed previously [32, 44], but the number of atoms retained in the optically active macrocycle was 44 and 84, respectively. According to the crystal structure [12], the B800 and B850 BChls in LH2 from *Rs. molischianum* differ only in the length of their phytyl chain, having a total of 107 and 140 atoms, respectively. The removal of the phytyl tail reduces dramatically both the size of the quantum system and the corresponding QC computational time. Furthermore, for the truncated BChls the non trivial task of automatic identification of the  $\text{Q}_y$  excited state in the case of a large number of such computations becomes easier and more precise. Although in general the different truncation schemes yield excitation energy time series with somewhat different (shifted) mean values, the corresponding energy fluctuations, which play the chief role in calculating the optical absorption properties of PPC at room temperature in their native environment, are less sensitive to the actual size of the truncated pigment.

The  $\text{Q}_y$  excitations of the truncated BChls were calculated by using Zerner’s semiempirical intermediate neglect of differential overlap method parametrized for spectroscopy (ZINDO/S) within the single-point configuration interaction singles (CIS) approximation [45, 46]. Because it is much faster and more accurate than most of the computationally affordable *ab initio* QC methods (e.g., the Hartree-Fock (HF) CIS method with the minimal STO-3G\* basis set), ZINDO/S CIS has been extensively used in the literature to compute low lying optically allowed excited states of pigment molecules [25, 26, 47, 48]. In cases similar to this study, where thousands of QC calculations are required, the proper balancing between speed and accuracy is absolutely essential. To further increase the computational speed, the active orbital space for the CIS calculations was restricted to the ten highest occupied (HOMO) and the ten lowest unoccupied (LUMO) molecular orbitals. According

to previous studies [32], as well as our own testings, the choice of a larger active space has negligible effect on the computed  $Q_y$  states. The ZINDO/S calculations were carried out with the QC program packages HyperChem 7.5 [49] and GAUSSIAN 98 [50]. Each calculation determined the lowest four excited states. Only in a small fraction ( $< 5\%$ ) of cases was the identification of the  $Q_y$  excited state (characterized by the largest oscillator strength and corresponding to transitions  $\text{HOMO} \rightarrow \text{LUMO}$  and  $\text{HOMO}-1 \rightarrow \text{LUMO}+1$ ) problematic, requiring careful inspection. We have found that even in such cases the  $Q_y$  state had the largest projection of the TDM along the  $y$ -axis, determined by the  $NB$  and  $ND$  nitrogen atoms.

The effect of the environment on the quantum system was taken into account through the electric field created by the partial point charges of the environment atoms, including those BChl atoms that were removed during the truncation process. Thus, the dynamics of the nuclear degrees of freedom (described by MD simulations) have a two-fold effect on the fluctuations of the  $Q_y$  state, namely they lead to: (1) conformational fluctuation of the (truncated) BChls, and (2) a fluctuating electric field created by the thermal motion of the corresponding atomic partial charges. In order to assess the relative importance of these two effects, the time series  $\Delta E_n(t)$  were calculated both in the presence and in the absence of the point charges. Since the ZINDO/S implementation in GAUSSIAN 98 does not work in the presence of external point charges, these calculations were done with Hyperchem. The ZINDO/S calculations without point charges were carried out with both QC programs and yielded essentially the same result. In [32] it has been found that at least a 2.2 ps long MD trajectory was needed for proper evaluation of optical observables related to their MD/QC calculations. However, our test calculations showed no significant difference between the energy gap autocorrelation functions calculated from a 2 ps and a 1 ps long energy gap time series. Therefore, to reduce the computational time we have opted for the latter. For each case, we have performed a total of 12,000 (500 snapshots  $\times$  24 BChls) ZINDO/S calculations. On a workstation with dual 3GHz Xeon EM64T CPU

it took  $\sim 2.3$  min/CPU for each calculation with point charges, and only  $\sim 0.7$  min/CPU without point charges. Thus, on a cluster of five such workstations, all 24,000 ZINDO/S runs were completed in  $\sim 1.9 + 0.6 = 2.5$  days.

## 2.4 Energy Gap Density of States

Figure 2.2 shows the  $Q_y$  energy gap DOS,  $\mathcal{N}(\omega)$ , of the individual B800 [top (a) panel] and B850 [bottom (b) panel] BChls calculated, according to Eqs. (2.6), as normalized binned histograms of the time series  $\Delta E_{B800} \equiv \Delta E_n(t_\ell)$  with  $n = 17, \dots, 24$ , and  $\Delta E_{B850} \equiv \Delta E_n(t_\ell)$  with  $n = 1, \dots, 16$ , respectively. In order to eliminate the noise due to finite sampling, the graphs have been smoothened out by a running average procedure. The same smoothing out procedure has been applied to all subsequent lineshape and spectra calculations. In the absence of the point charge distribution of the environment  $\mathcal{N}(\omega)$  for B800 and B850 (dashed lines) are almost identical, having peak position at 1.51 eV (817 nm) and 1.515 eV (818 nm), and full width at half maximum (FWHM) 51 meV and 59 meV, respectively. The fact that the peak position practically coincides with the mean energy gap is indicative that the DOS is symmetric with respect to its maximum. It should be noted that essentially the same mean energy gap of 1.5 eV was obtained in similar MD/QC calculations (i) in this study (data not shown), in the case of a truncated BChl-a in vacuum but artificially coupled to a Langevin heat bath at room temperature, and (ii) by Mercel et al. [32] for a BChl-a solvated in methanol, also at room temperature. Although in case (ii) the width of the DOS is somewhat broader (FWHM  $\approx 65$  eV) than in case (i), for which FWHM  $\approx 58$  meV, based on these results one can safely conclude that the thermal motion of the nuclei in individual BChls lead to  $Q_y$  energy gap fluctuations that are insensitive to the actual nature of the environment. Since in the LH2 from *Rs. molischianum* the surrounding of the B800s is polar, while that of the B850s is not, one expects that once the point charges of the environment are taken into account in the QC calculations,  $\mathcal{N}(\omega)$  should change

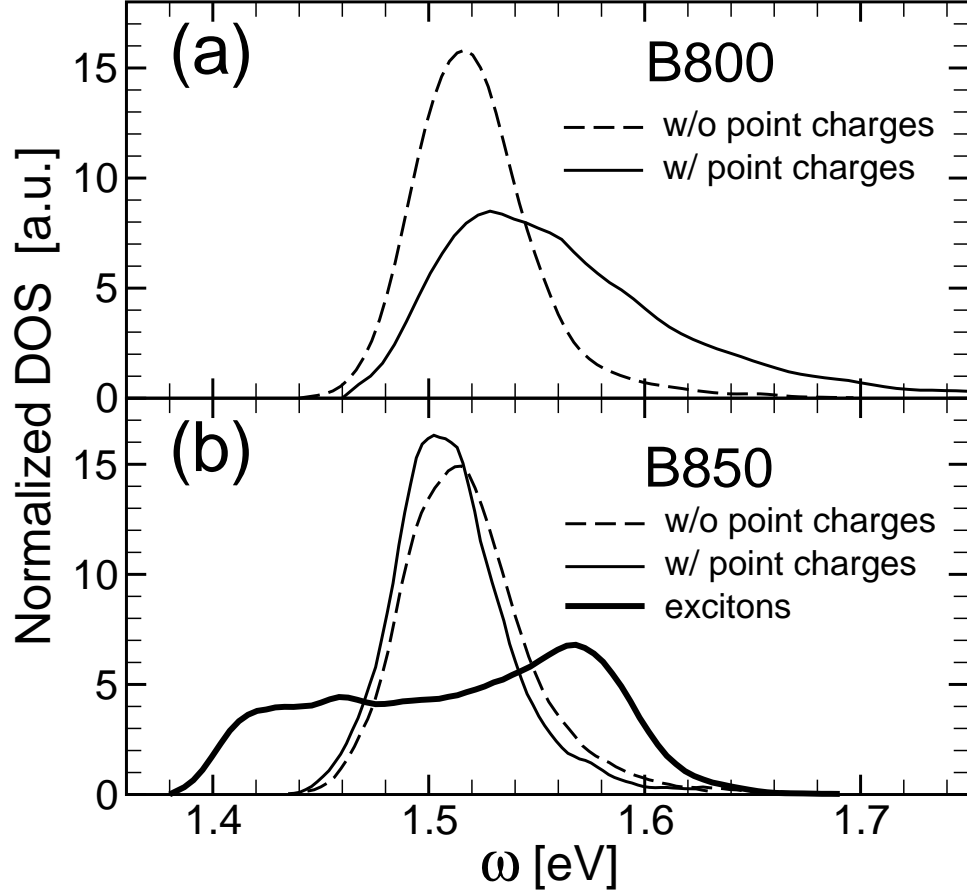


Figure 2.2: Normalized DOS,  $\mathcal{N}(\omega)$ , for (a) B800, and (b) B850 BChls in LH2 of *Rs. molischianum* computed as binned histograms of the corresponding  $Q_y$  excitation energy time series obtained from combined MD/QC simulations. Whether the charge fluctuations of the BChls' environment are included (solid lines) or not (dashed line) makes an important difference in  $\mathcal{N}(\omega)$  only for B800. In (b) the DOS of the B850 excitons is shown as a thick solid line.

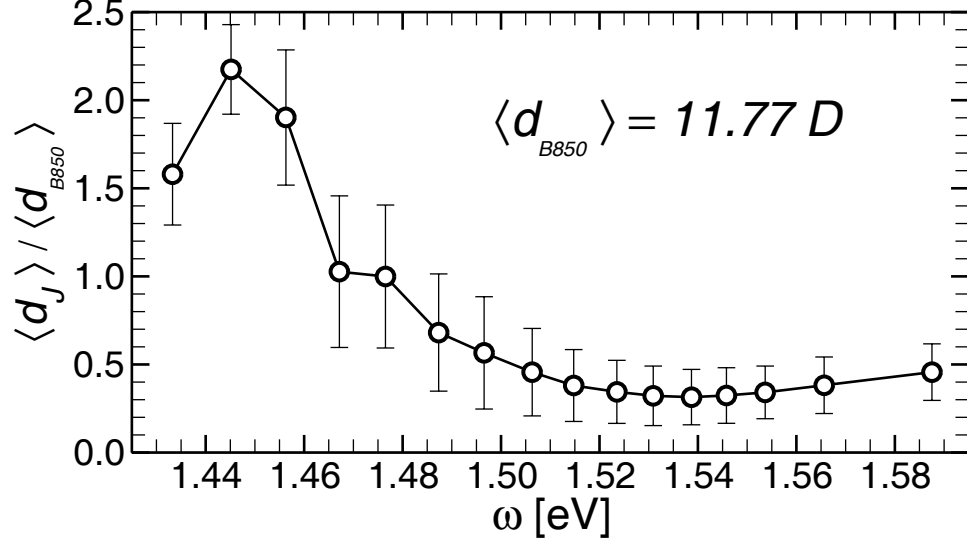


Figure 2.3: Average transition dipole moments  $\langle d_J \rangle$  corresponding to the  $J = 1, \dots, 16$  B850 excitonic states. Both  $\langle d_J \rangle$  and the corresponding error bars are expressed relative to the mean dipole moment of individual B850s.

dramatically only in the case of B800. Indeed, as shown in Fig. 2.2b (solid line), in the presence of point charges the peak of  $\mathcal{N}_{B850}(\omega)$  is only slightly red shifted to 1.502 eV (825 nm) and essentially without any change in shape with FWHM  $\approx 53$  meV. By contrast, the DOS for B800 in the presence of point charges (Fig. 2.2a) has qualitatively changed. The induced higher energy fluctuations not only spoil the symmetry of  $\mathcal{N}_{B800}(\omega)$ , but also increase dramatically its broadening, characterized by FWHM  $\approx 100$  meV. Thus, in spite of a small blueshift to 1.528 eV (811 nm) of the peak of  $\mathcal{N}_{B800}(\omega)$  the mean value of the energy gap  $\langle \Delta E_{B800} \rangle = 1.556$  eV (797 nm) is increased considerably, matching rather well the experimental value of 800 nm.

The time series of the excitonic energies  $\Delta E_J(t_\ell)$ ,  $J = 1, \dots, 16$ , of the B850 BChls were determined by solving for each MD snapshot, within the point-dipole approximation, the eigenvalue equation (2.16). In calculating the matrix elements (2.15),  $\mathbf{r}_n$  was identified with the position vector of the Mg atom in the  $n$ -th BChl. Consistent with the Condon approximation, the magnitude of the computed B850 TDM time series exhibited a standard deviation of less than 4% about the average value  $\langle d_{B850} \rangle = 11.77 D$ . The latter is by a



factor of  $k = 1.87$  larger than the experimentally accepted 6.3 D value of the  $Q_y$  TDM of BChl-a [51]. Thus, to account for this overestimate of the TDM by the ZINDO/S CIS method, it was rescaled by the factor  $k^{-1}$ . The relative dielectric constant of the protein environment in Eq. (2.15) was set to a reasonable value  $\epsilon_r = 1.86$ . The mean value of the nearest neighbor dipolar coupling energies between B850s were  $27 \text{ meV} \approx 220 \text{ cm}^{-1}$  within a protomer and  $24 \text{ meV} \approx 196 \text{ cm}^{-1}$  between adjacent heterodimers. Just like in the case of individual BChls, the DOS corresponding to the B850 excitonic energies (Fig. 2.2b - thick line) was calculated as a binned histogram of  $\Delta E_J(t_\ell)$ . As expected, the excitonic DOS is not sensitive to whether the point charges of the environment are included or not in the B850 site energy calculations.

The mean excitonic TDMs, calculated from Eq. (2.17) and expressed in terms of  $\langle d_{B850} \rangle$ , are shown in Fig. 2.3. The error bars represent the standard deviation of the time series  $d_J(t_\ell)$ . In agreement with previous studies, most of the dipole strength is amassed into the lowest three excitonic states.

As discussed in Sec. 2.2.1, a rough estimate of the lineshape function can be obtained as the combined DOS of the B800 BChls and B850 excitons. In this approximation the OD spectrum reads

$$I_{DOS}(\omega) \propto \omega \left[ \sum_J d_J^2 \langle \delta(\omega - \Delta E_J) \rangle + \sum_{B800} d_{B800}^2 \langle \delta(\omega - \Delta E_{B800}) \rangle \right], \quad (2.23)$$

where the  $B800$  index in the last term means summation over all B800 BChls. Figure 2.4 shows the calculated  $I_{DOS}(\omega)$  blueshifted by 20 eV (solid line) in order to match the B850 peak position with the one in the experimental OD spectrum [25, 31] (dashed line). While the B850 band and the relative heights of the two peaks in  $I_{DOS}(\omega)$  match rather well the experimental data, the position and the broadening of the B800 peak do not. This result clearly shows that in general peak positions in optical spectra may be shifted from the corresponding peak positions in the excitation energy spectrum due to correlation effects

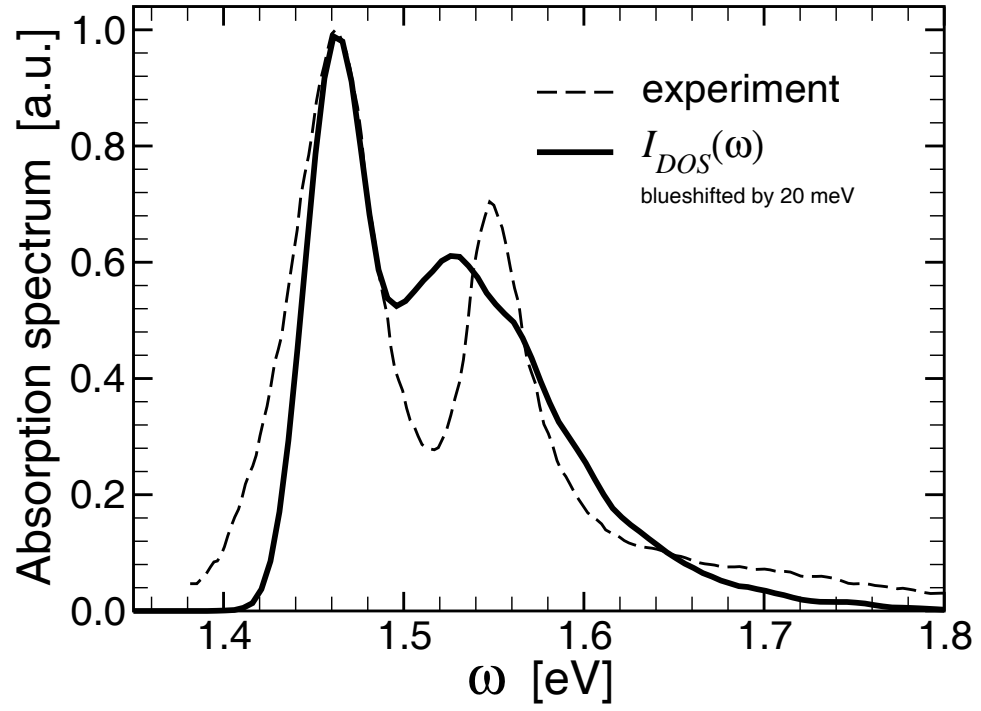


Figure 2.4: Absorption spectrum  $I_{DOS}(\omega)$  of LH2 for *Rs. molischianum* calculated as a combined DOS of B800 BChls and B850 excitons weighted by the corresponding dipole strengths (solid line).  $I_{DOS}(\omega)$  was blueshifted by 20 meV in order to overlay its B850 peak with the corresponding one in the experimental OD spectrum [31] (dashed line).

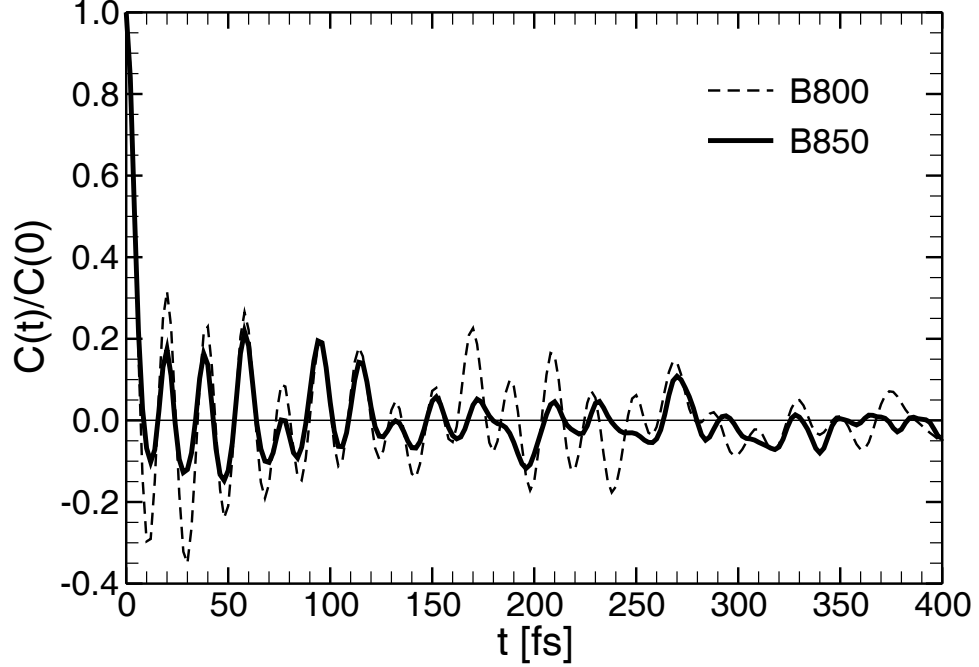


Figure 2.5: Normalized autocorrelation function  $C(t)/C(0)$  of the energy gap fluctuations  $\delta E(t) = E(t) - \langle E \rangle$  for individual B800 (dashed line) and B850 (solid line) BChls, calculated using Eq. (2.24). The mean square energy gap fluctuations are  $C_{B800}(0) = 3.16 \times 10^{-3} \text{ eV}^2$  and  $C_{B850}(0) = 8.68 \times 10^{-4} \text{ eV}^2$ .

between the ground and optically active excited states. The latter may also lead to different line broadening of the corresponding peaks. Thus, it appears that in principle, methods for simulating optical spectra in which the position of the peaks are identified with the computed excitation energies (stick spectrum) are not entirely correct and using instead more sophisticated methods that include quantum correlation effects should be preferred. Such method, based on the cumulant approximation of the lineshape function as described in Sec. 2.2, is used in the next section for calculating the OD spectrum of an LH2 ring from *Rs. molischianum*.

## 2.5 Linear Absorption Spectrum

The key quantity for calculating the lineshape functions of the individual B850 and B800 BChls is the (classical) autocorrelation function  $C_n(t) = \langle \delta E_n(t) \delta E_n(0) \rangle$  of the energy gap

fluctuation  $\delta E_n(t) = \Delta E_n(t) - \langle \Delta E_n \rangle$  determined from the combined MD/QC calculations. Because the time series  $\Delta E_n(t)$  were too short for a proper evaluation of the ensemble average in the individual  $C_n(t)$ , a single time correlation function  $C_{B800}(t)$  [ $C_{B850}(t)$ ] was determined by averaging over all B800 [B850] BChls according to the formula

$$C_\alpha(t_\ell) = \frac{1}{M} \sum_m \left[ \frac{1}{N_t - \ell} \sum_{k=1}^{N_t - \ell} \delta E_m(t_\ell + t_k) \delta E_m(t_k) \right], \quad (2.24)$$

where  $M = 8$ ,  $m = 17, \dots, 24$  for  $\alpha = B800$ ,

and  $M = 16$ ,  $m = 1, \dots, 16$  for  $\alpha = B850$ .

The normalized correlation functions  $C_\alpha(t)/C_\alpha(0)$ ,  $\alpha \in \{B800, B850\}$ , are plotted in Fig. 2.5.  $C_\alpha(0) = \langle \delta E^2 \rangle$  represents the variance of the energy gap fluctuations with  $C_{B800}(0) = 3.16 \times 10^{-3} \text{ eV}^2$  and  $C_{B850}(0) = 8.68 \times 10^{-4} \text{ eV}^2$ . The behavior of the two correlation functions is rather similar during the first 150 fs. Following a sharp decay to negative values in the first 9 fs, both functions exhibit an oscillatory component of approximately 18.5 fs period and uneven amplitudes which are larger for the B800. After  $\sim 150$  fs, the autocorrelation functions behave in a distinctive manner, both becoming negligibly small for  $t > 400$  fs.

The spectral densities  $J_\alpha(\omega)$  for B800 and B850, determined according to Eq. (2.12), are shown in Fig. 2.6. The prominent peak about  $\omega_p = 0.22 \text{ eV}$  is due to the fast initial decay of  $C_\alpha(t)$ . Being reported in previous studies [22, 32], by using both *ab initio* (HF/CIS with STO-3G\* basis set) and semi empirical QC methods, these spectral features appear to be intrinsic properties of BChl-a, most likely originating from a strong coupling of the pigment to an intramolecular C=O vibronic mode. Often, the environment in a PPC is modeled as an equivalent harmonic (phonon) heat bath for which the cumulant approximation is exact [9]. The corresponding phonon spectral density can be written as  $J(\omega) = \omega^2 \sum_\lambda g_\lambda^2 \delta(\omega - \omega_\lambda)$ , where  $g_\lambda$  is the coupling constant to the phonon mode  $\lambda$ . Thus, one can interpret the magnitude of the spectral functions in Fig. 2.6 as a measure of

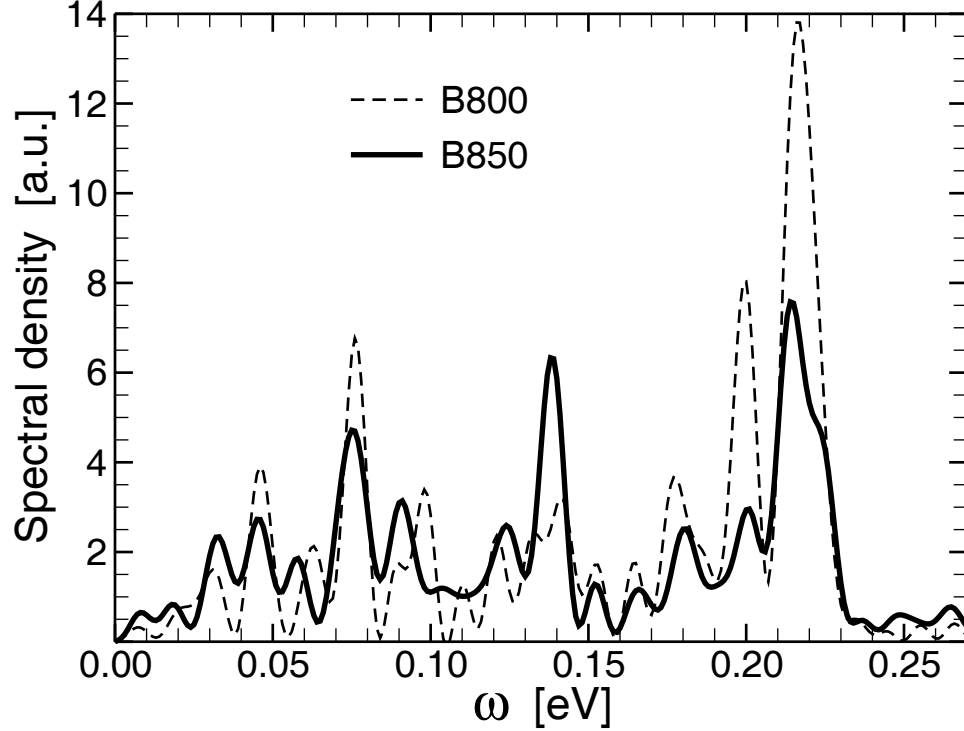


Figure 2.6: Spectral density function  $J(\omega)$  for B800 (dashed line) and B850 (solid line) obtained according to Eq. (2.12).

the coupling strength to phonons of that particular frequency. The complex structure of the spectral functions indicate that all inter and intra molecular vibronic modes with frequency below  $\omega_p$  will contribute to the lineshape function. Hence, attempts to use simplified model spectral functions appear to be unrealistic, even if these may lead to absorption spectra that match the experimental results.

The lineshape functions of individual B800 and B850, calculated from Eqs. (2.14), are plotted in Fig. 2.7. The origin of the frequency axis corresponds to the mean energy gaps  $\omega_{B800}$  and  $\omega_{B850}$ , respectively. The highly polarized surrounding of the B800 BChls in *Rs. molischianum* renders  $A_{B800}(\omega)$  twice as broad (FWHM  $\approx 26$  meV) as  $A_{B850}(\omega)$  (FWHM  $\approx 13$  meV). Also, the redshift of the peak of the former ( $\Delta\omega \approx 25$  meV) is more than three times larger than that of the latter ( $\Delta\omega \approx 7$  meV).

Since the available simulation data is not sufficient to properly estimate the excitonic lineshape functions  $A_J(\omega)$ , by neglecting the effect of *exchange narrowing* [11, 20], we ap-

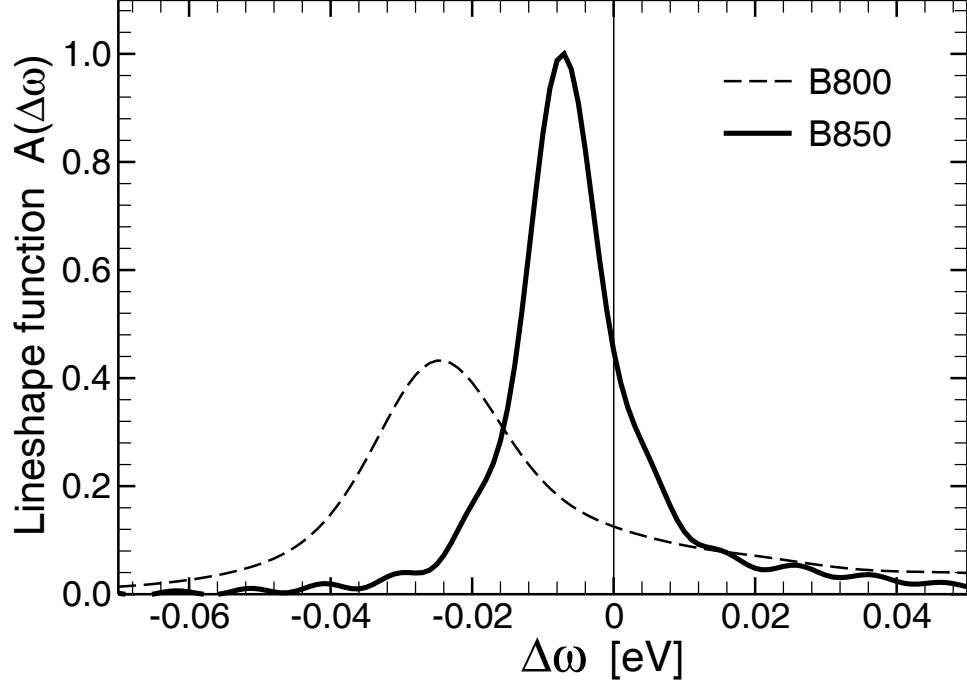


Figure 2.7: Lineshape functions  $\bar{A}_{B800}(\Delta\omega)$  (dashed line) and  $\bar{A}_{B850}(\Delta\omega)$  (solid line).

proximated these with  $A_{B850}(\omega)$ . Thus, the OD spectrum of the LH2 BChls was calculated by using the formula

$$I(\omega) \propto \omega \left[ \sum_J d_J^2 \bar{A}_{B850}(\omega - \omega_J) + 8d_{B800}^2 \bar{A}_{B800}(\omega - \omega_{B800}) \right], \quad (2.25)$$

where  $\omega_J = \langle \Delta E_J \rangle$ .

As shown in Fig. 2.8, after an overall blueshift of 20 meV,  $I(\omega)$  matches remarkably well the experimental OD spectrum, especially if we take into account that it was obtained from the sole knowledge of the high resolution crystal structure of LH2 from *Rs. molischianum*. The reason why both B800 and B850 peaks of  $I(\omega)$  are somewhat narrower than the experimental ones is most likely due to the fact that the effect of static disorder is ignored in the present study. Indeed, all calculations were based on a single LH2 ring, while the experimental data is averaged over a large number of such rings. While computationally expensive, in principle, the effect of static disorder could be taken into account

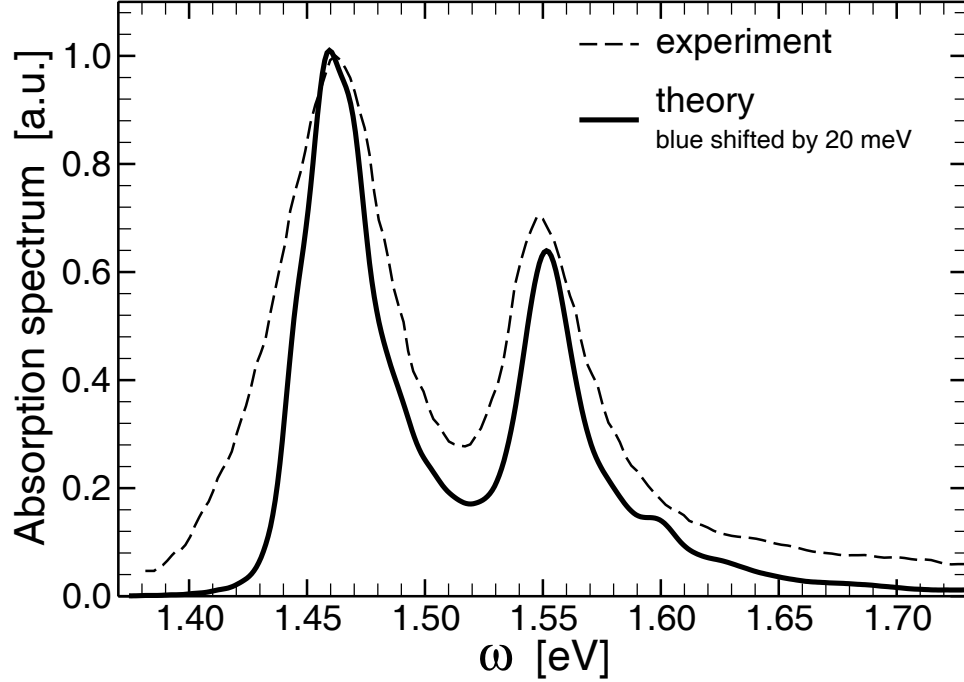


Figure 2.8: Computed (solid line) and experimental (dashed line) absorption spectra (in arbitrary units) of the BChl aggregate in *Rs. molischianum* LH2. The computed spectrum has been blue shifted by 20 meV for best match.

by repeating the above calculations for different initial configurations of the LH2 ring and then averaging the corresponding OD spectra.

To conclude this section, we relate our work to two previous combined MD/QC studies [22, 32]. In [32] it is argued that the *ab initio* QC method (HF/CIS with the STO-3G\* basis set) should be preferred to semi empirical methods for calculating optical spectra because it reproduces better their experimental results. The FWHM of their calculated semi empirical and *ab initio* absorption spectra of BChl-a in methanol are  $\sim 65$  meV and  $\sim 125$  meV, respectively. These values are similar to the ones we obtained for the same type of calculations for BChl-a in vacuum and in LH2 embedded in its native environment. Since, except [32], all experimental results on the  $Q_y$  absorption band of BChls we are aware of, have a FWHM of  $< 85$  meV at room temperature, we conclude that in fact the semi empirical ZINDO/S method should be preferable to the *ab initio* QC method. In general, the latter overestimates the broadening of the OD spectrum by a factor of 2 to 3. In [22],

the OD spectrum of individual B850s (i.e., without excitonic coupling) calculated with the *ab initio* method yielded the same FWHM of  $\sim 125$  meV as in [32]. Once the excitonic coupling was included within the framework of a polaron model, and it was assumed that the entire oscillator strength was carried by a single exciton level of a perfect B850 ring, the FWHM of the resulting OD spectrum was reduced to  $\sim 43$  meV as a result of exchange narrowing. Even though the obtained OD spectrum matched well the corresponding part of the experimental one, when applying the same method to the B800s, where there is no exchange narrowing, the polar environment further broadens the corresponding OD spectrum to a FWHM of  $\sim 250$  meV that is clearly unphysically large. The conclusion is again that the ZINDO/S CIS semi empirical method should be preferred for calculating optical spectra of PPCs.

## 2.6 Circular Dichroism

The CD spectrum of the LH2 BChls from *Rs. molischianum* was determined by following the theoretical approach described in Sec. 2.2 and by employing the same time series (obtained from the combined MD/QC calculations) used for calculating the OD spectrum.

First, the rotational strength of both B850 excitons and B800 BChls were determined by using Eq. (2.22b). In this equation, just like in the case of the point-dipole interaction matrix elements (2.15), the vector  $\mathbf{r}_n$  described the position of the Mg atom in the  $n^{th}$  BChl. As already clarified in Sec. 2.2.2, the calculation of the rotational strength of the B800 BChls requires solving the corresponding excitonic Hamiltonian (2.16) regardless of how small the dipole-dipole coupling is between these BChls. The calculation does not yield either noticeable corrections to the B800 excitation energies or admixture of the corresponding  $Q_y$  states, however, it leads to sizable mean rotational strengths as shown in Fig. 2.9 (filled circles). Similarly to the TDM strengths (Fig. 2.3), the largest (negative) mean rotational strengths are carried by the four lowest B850 excitonic states as shown in



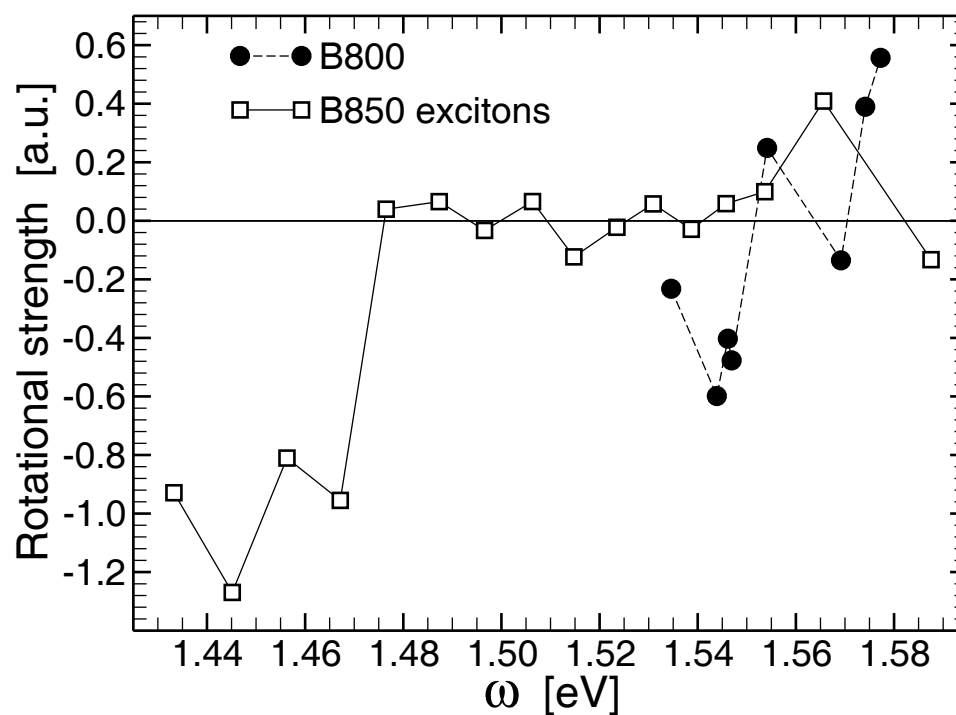


Figure 2.9: Mean rotational strength of the excitonically coupled B800 (circles) and B850 (rectangles) BChls as a function of the corresponding excitonic energies. The purpose of the thin lines is to guide the eye.

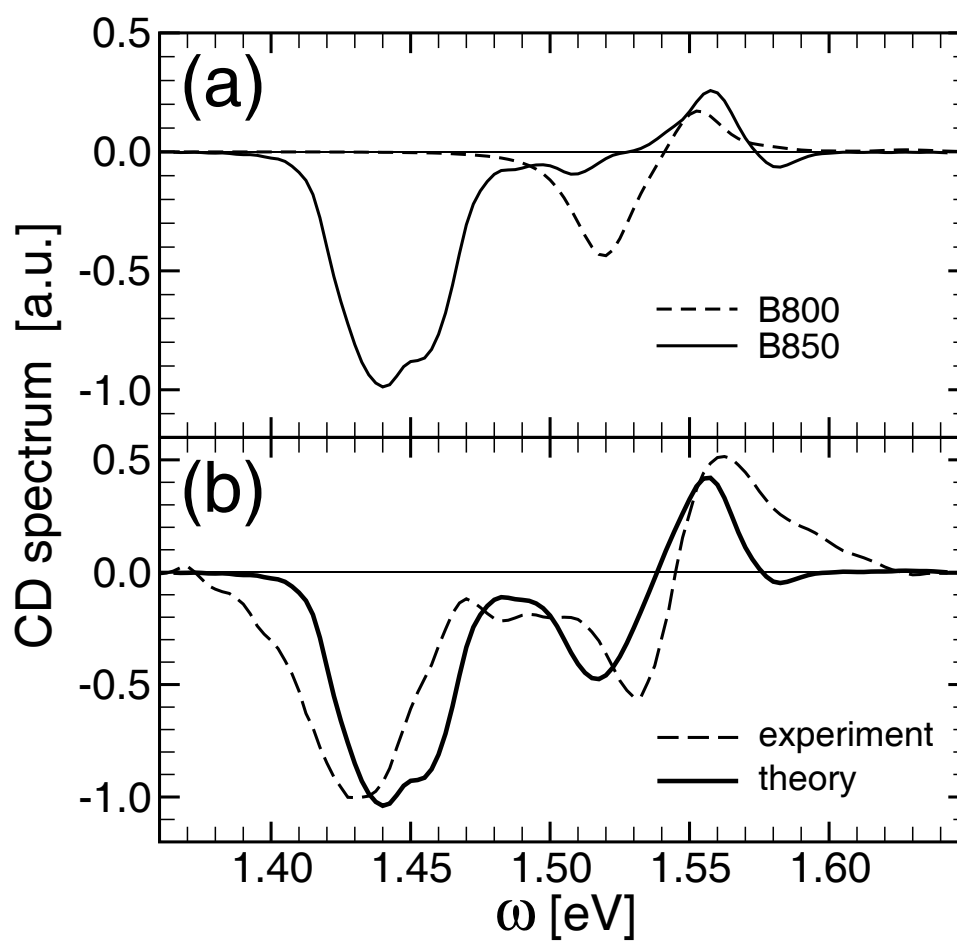


Figure 2.10: (a) CD spectrum contributions due to B800 (dashed line) and B850 (solid line) BChls. (b) Comparison between the computed (solid line) and experimental (dashed line) CD spectrum of the BChl aggregate in *Rs. molischianum* LH2.

Fig. 2.9 (open squares). The second highest excitonic state also has a sizable rotational strength and is responsible for enhancing the positive peak of the B800 contribution to the CD spectrum (Fig. 2.10a).

Second, the CD spectrum is calculated from Eq. (2.22a) where the summation index  $J$  runs over all B850 and B800 excitonic states and  $A_J(\omega) = \bar{A}_\alpha(\omega - \omega_J)$ , with  $\alpha \in \{B850, B800\}$ . Figure 2.10a shows the CD spectrum contribution by the B850 (solid line) and B800 (dashed line). Both contributions have the same qualitative structure with increasing energy: a pronounced negative peak followed by a smaller positive one. The B850 negative CD peak is about twice as large as the corresponding B800 peak. The total CD spectrum, given by the superposition of the B850 and B800 contributions, is shown in Fig. 2.10b (solid line) and matches fairly well the experimental spectrum [25] (dashed line). It should be emphasized that apart from an overall scaling factor the CD spectrum was calculated from the same MD/QC data as the OD spectrum by following the procedure described above.

## 2.7 Conclusions

By applying an approach that combines all atom MD simulations, efficient semi empirical QC calculations and quantum many-body theory, it has been shown that starting from the sole knowledge of the atomic structure of the LH2 ring from *Rs. molischianum*, the OD and CD spectra of this PPC can be predicted with reasonable accuracy at affordable computational costs.

The configuration snapshots, taken with femtosecond frequency during the MD simulation of the PPC in its native, fully solvated lipid-membrane environment at room temperature and normal pressure, provide the necessary input for the QC calculations of the optical excitation energies and transition dipole moments of the pigment molecules. The obtained time series are used to evaluate within the second cumulant approximation the op-

tical lineshape functions as the Fourier transform of the quantum dipole-dipole correlation function. The choice of the ZINDO/S CIS method for the QC calculations was motivated by the fact that it is almost two orders of magnitude faster and much more accurate than the most affordable *ab initio* method (HF/CIS with the STO-3G\* basis set). Compared to the former, the latter method overestimates by a factor of 2 to 3 both the excitation energies and the broadening of the energy spectrum. Just like in several previous studies [25, 26, 47, 48], we have found that the ZINDO/S method repeatedly yields results in good agreement with existing experimental data.

By investigating the excitation energy spectrum of the LH2 BChls both in the presence and in the absence of the atomic partial charges of their environment, this study has convincingly demonstrated that the large broadening of the B800 peak is due primarily to the electric field fluctuations created by the polar surrounding environment of the B800s. There is no such effect for the B850s, which lay in a nonpolar local environment. The broadening of the B850 peak is due to the excitonic coupling between these BChls. Since only the lowest three excitonic states carry most of the available dipole strength, in spite of the  $\sim 0.2$  eV wide excitonic band, the B850 absorption peak has a FWHM only slightly larger than the B800 one.

# **3 Calculating free energy profiles and diffusion coefficients from non-equilibrium molecular dynamics simulations**

## 3.1 Introduction

The study of the structure-function relationship of large biomolecules often requires following their dynamics on a meso- or even macro-scopic time scale while retaining its atomic scale spatial resolution. A typical example is molecular and ion transport through protein channels [54]. While structural details are needed at atomic resolution in order to determine the forces that guide the diffusion of the particles across the channel, the duration of the permeation process may exceed by several orders of magnitude the time scale of a few tens of nanoseconds currently attainable by all atom molecular dynamics (MD) simulations [55]. In such cases, a simplified alternative approach is to model the transported molecule in the channel as an overdamped Brownian particle that diffuses along the axis of the channel in the presence of an effective *potential of mean force* (PMF) that describes its interaction with the rest of the atoms in the system [56]. A PMF is the Landau free energy profile along a *reaction coordinate* (RC), or *order parameter* [55], and it can be determined from the equilibrium statistical distribution function of the system by integrating out all degrees of freedom except the RC [57, 58]. In principle, both the effective diffusion coefficient and the PMF, quantities that enter the Langevin equation of motion (or, equivalently, the corresponding Fokker-Planck equation [59]) which determines the dynamics of the transported molecule, can be determined from MD simulations. In practice, however, the calculation of PMFs is rather difficult and computationally expensive [57, 60].

Since even the longest equilibrium MD (EMD) trajectories can sample only a small region of the RC domain of interest, the one situated in the vicinity of the PMF minimum, simple EMD simulations are not suitable for PMF calculations. The traditional method for calculating PMFs using biased EMD simulations is *umbrella sampling* (US) [57, 61, 62]. However, US may become computationally unaffordable when too many sampling windows are required. This may happen when the amplitude of the equilibrium fluctuations

---

This chapter is based on published article [52] and manuscript [53].

of the RC is very small compared to the size of the RC interval in which the PMF is sought.

In such cases the RC can be sampled efficiently by employing *steered molecular dynamics* [63] (SMD) in which the system is guided, according to a predefined protocol, along the RC by using, e.g., a *harmonic guiding potential* (HGP). By choosing a sufficiently large value for the elastic constant of the HGP, i.e., within the *stiff-spring approximation* [5, 64] (SSA), the distance between the target and actual value of the RC at a given time can be kept below a desired value. In general, for a large system ( $\sim 10^5$  atoms) computationally one can afford only a limited number (typically  $\sim 10$ ) of such *nonequilibrium* SMD pulls, and the real challenge is to find a way to reconstruct the PMF (at least semiquantitatively) along the RC using this limited amount of data. In principle, the equilibrium PMF can be reconstructed from the *Jarzynski equality* (JE) that relates the equilibrium free energy difference  $\Delta F$  between two states to the average of the external work  $W$  done along all nonequilibrium paths that connect those states and are subject to the preestablished RC variation protocol [65, 66]. In terms of the *dissipative work*  $W_d = W - \Delta F$ , JE can be written as  $\langle \exp(-\beta W_d) \rangle = 1$ , where  $\beta = 1/k_B T$ ,  $k_B$  is the Boltzmann constant and  $T$  is the temperature of the heat bath (environment). In the case of *reversible works* the system is at equilibrium at all times. Therefore  $W_d = 0$  and JE is trivially satisfied. In general SMD pulls are nonequilibrium with  $W_d > 0$  along most of the trajectories. However, the validity of JE depends crucially on a small fraction of trajectories with  $W_d < 0$ , that transiently violate the *second law of thermodynamics*. Since such trajectories (whose number decreases exponentially with  $\overline{W}_d$ ) are very unlikely to occur among a few fast SMD pulls, it is clear that the sought PMF cannot be determined by the direct application of JE, except when the pulling paths are close to equilibrium (i.e., when  $\overline{W}_d < k_B T$ ). Under near-equilibrium conditions, the validity of JE has been confirmed in an RNA stretching experiment [67, 68]. Even though many papers were dedicated to the applications of JE [69–79] and other *fluctuation theorems* [80, 81], there are surprisingly few studies which use SMD simulations combined with the JE to calculate PMFs for large biomolecules [5, 82].

This chapter presents a simple and efficient method for calculating PMFs, referred to as the *FR method*, that is compared to other popular methods in literature. The proposed FR method calculates simultaneously both the PMF  $U(z)$  and the corresponding diffusion coefficient  $D(z)$  along a RC  $z$  for a classical many particle system by employing a small number of fast nonequilibrium SMD pulls in both forward (F) and time reverse (R) directions, without invoking JE. In fact, as already mentioned, for such limited number of processes JE fails to hold. The essence of the FR method, can be summarized as follows: Several fast F and R SMD pulls are carried out within the SSA. The latter guaranties that (i) the RC follows closely its target value determined by the pulling protocol, (ii) the change in PMF ( $\Delta U$ ) is well approximated by the corresponding change in the free energy ( $\Delta F$ ) of the system biased by the HGP, and (iii) the work distribution function  $P_{F/R}(W)$  along F/R paths is Gaussian. A few F and R SMD trajectories are sufficient to sample  $P_{F/R}(W)$  about its maximum (see Fig. 3.1) and, therefore, determine approximately the mean F/R work  $\bar{W}_{F/R}$ . However, the same data is insufficient for even a rough estimate of the variance  $\sigma_W^2 = \overline{W^2} - \bar{W}^2$ , i.e., of the actual width of  $P_{F/R}(W)$ . From Crooks' *transient fluctuation theorem* [83] (TFT) [see Eq. (3.16)], which is more general than JE, follows that if  $P_F(W)$  is Gaussian then  $P_R(W)$  is also a Gaussian with the same variance  $\sigma_W^2 = 2k_B T \bar{W}_d$ , and peak position  $\bar{W}_R = \bar{W}_F - 2\Delta F$ . Thus, (i) the PMF is given by  $\Delta U = \Delta F = (\bar{W}_F - \bar{W}_R)/2$ , and (ii) the mean dissipative work is the same for both F and R paths, given by  $\bar{W}_d = (\bar{W}_F + \bar{W}_R)/2$ . From  $\bar{W}_d$  the position dependent diffusion coefficient is  $D = k_B T v / (d\bar{W}_d/dz)$ .

Thus, the reason why previous studies failed to reconstruct the PMF from *unidirectional* SMD pulls far from equilibrium by using JE is because such approach requires the complete sampling of the corresponding work distribution function, which is impossible to obtain from a limited number of pulls. While the mean work can be easily estimated, obtaining the PMF and the mean dissipative work requires either the knowledge of the precise variance of the F work distribution function (e.g., when the F SMD paths are close to equilibrium and  $\bar{W}_d$  is small) or additional information that may come from a set of R



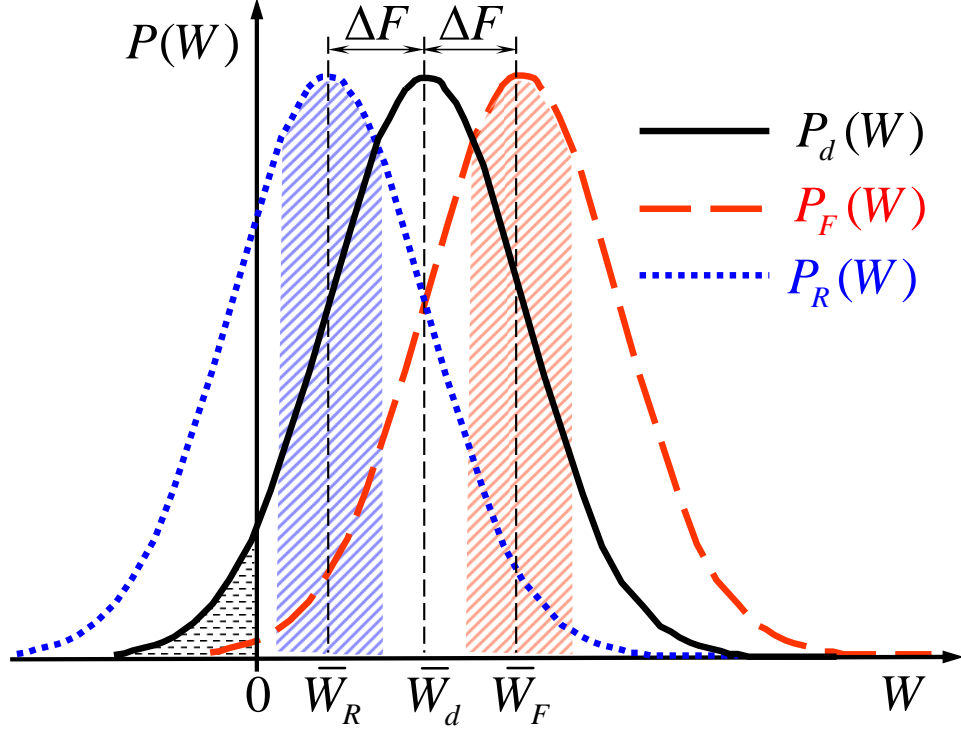


Figure 3.1: Gaussian forward (long-dashed), reverse (dotted) and dissipative (solid) work distribution functions within the *stiff spring approximation*. The shaded region in  $P_{F/R}(W)$  is the one sampled in F/R SMD pulls. The tail region of  $P_d(W)$  corresponding to negative dissipative work is also highlighted.

SMD pulls, as outlined above. The solution to this problem offered by the FR method is surprisingly simple, but its validity depends crucially on Crooks' TFT and the Gaussian nature of  $P_{F/R}(W)$  guaranteed by SSA. In particular, the conclusion that  $\bar{W}_d$  is the same for both F and R SMD paths is highly non trivial.

In the following, we consider a classical many particle system: deca-alanine peptide in vacuum, artificially coupled to a heat bath, used to test the efficiency and viability of the FR method. The system is described by the Hamiltonian  $H_0(\Gamma)$ , where  $\Gamma \equiv \{\mathbf{r}, \mathbf{p}\}$  represents the phase space coordinates of all the atoms in the system. It is assumed that the conditions for which JE and TFT hold are met, i.e., the dynamics are Markovian and preserve the equilibrium ensemble, and the energy of the system is finite [83]. These conditions are met in MD simulations in both NVT and NPT ensembles [64].

## 3.2 Reaction Coordinates and Potential of Mean Force

In general, any PMF calculation starts with the identification of a properly chosen RC whose change in time describes the evolution of the state of the system [57, 58]. In the case of our system, a proper RC is the end-to-end distance, denoted by  $z$ .

By definition, the PMF  $U(z)$  is determined from the equilibrium distribution function of the system by integrating out all degrees of freedom except the RC  $z$ , i.e., [57]

$$e^{-\beta U(z)} \equiv p_0(z) = \int d\Gamma \frac{e^{-\beta H_0(\Gamma)}}{Z_0} \delta[z - \tilde{z}(\Gamma)] , \quad (3.1)$$

where  $p_0(z)$  is the equilibrium distribution function of the reaction coordinate,  $Z_0$  is the partition function and  $\delta(z)$  is the Dirac-delta function whose filtering property guarantees that the integrand in Eq. (3.1) is nonzero only when the RC has the desired value, i.e., when  $\tilde{z}(\Gamma) = z$ . Hereafter, the convention that  $z$  [or  $z(t)$ ] represents the target value of the RC, while  $\tilde{z} \equiv \tilde{z}(\Gamma)$  represents the actual value of the RC is used. Also, unless otherwise stated, the energy is measured in units of  $k_B T$ , e.g., in Eq. (3.1) one needs to set  $\beta = 1$ .

In principle, the equilibrium distribution function  $p_0(z)$  can be easily computed from EMD simulations, since it is proportional to the logarithm of the binned histogram of the RC sampled along the MD trajectory. Thus, the PMF is given by

$$U(z) = -\log[p_0(z)] . \quad (3.2)$$

In terms of the  $U(z)$  the equilibrium average of any function  $f(\tilde{z})$  of the RC is

$$\begin{aligned} \langle f(\tilde{z}) \rangle_0 &= \int d\Gamma \frac{e^{-H_0(\Gamma)}}{Z_0} f(\tilde{z}) \int dz \delta[z - \tilde{z}(\Gamma)] \\ &= \int dz e^{-U(z)} f(z) = \int dz p_0(z) f(z) . \end{aligned} \quad (3.3)$$

As already mentioned, in practice, even the longest EMD trajectories sample only the re-

gion of the reaction coordinate domain within the vicinity of the PMF minimum) and, therefore, the direct application of Eq. (3.1) is impractical.

### 3.3 Harmonic Guiding Potential

In order to properly sample energetically more difficult to reach regions, one needs to *guide* or *steer* the system towards those regions by employing, e.g., a harmonic guiding potential (HGP)

$$V_z(\tilde{z}) \equiv V(\tilde{z}(\Gamma)|z) = \frac{k}{2}[\tilde{z}(\Gamma) - z]^2, \quad (3.4)$$

where  $k \equiv k_z$  is the stiffness (elastic constant) of the HGP. The Hamiltonian of the new biased system becomes  $H_z = H_0 + V_z(\tilde{z})$ . As a result, atom “ $j$ ” in the selection that defines the RC will experience an additional force

$$\mathbf{F}_j = -\frac{\partial V_z}{\partial \mathbf{r}_j} = -k[\tilde{z}(\Gamma) - z] \frac{\partial \tilde{z}(\Gamma)}{\partial \mathbf{r}_j}. \quad (3.5)$$

Thus, the HGP (3.4) will force the system to evolve in the configuration space in such a way that at all times  $\tilde{z}$  stays confined in the vicinity of  $z$ .

The free energy difference  $\delta F_z = F_z - F_0$  between the equilibrium states of the systems described by the Hamiltonians  $H_z$  and  $H_0$  can then be written as a Gaussian convolution of  $\exp[-U(z)]$ . Indeed,

$$\begin{aligned} e^{-\delta F_z} &= \int d\Gamma \frac{e^{-\beta H_0(\Gamma)}}{Z_0} e^{-V_z(\tilde{z}(\Gamma))} \\ &= \left\langle e^{-V_z(\tilde{z})} \right\rangle_0 \\ &= \int dz' e^{-U(z')} e^{-V_z(z')} = \int dz' e^{-U(z')} e^{-\frac{k}{2}(z-z')^2}. \end{aligned} \quad (3.6)$$

### 3.4 Stiff Spring Approximation

The sought PMF,  $U(z)$ , can be obtained from Eq. (3.6) by Gaussian deconvolution of the free energy factor  $\exp(-\delta F_z)$ . However, it is more convenient to resort to the large  $k$  or *stiff-spring approximation* [5, 64, 84] (SSA). Assuming that we seek to determine  $U(z)$  with a spatial resolution  $\delta z$ , by choosing the spring constant such that  $k \gg 2/(\delta z)^2$  one can easily see that in Eq. (3.6) the main contribution to the last integral comes from the region  $|z - z'| \ll \delta z$ , and therefore one can write

$$e^{-\delta F_z} \approx e^{-U(z)} \int dz' e^{-\frac{k}{2}(z-z')^2} = \sqrt{\frac{2\pi}{k}} e^{-U(z)}. \quad (3.7)$$

The logarithm of both sides in Eq. (3.7) yields  $\delta F_z = F_z - F_0 = U(z) + \text{const.}$  and, therefore,

$$\Delta U = U(z) - U(z_0) \approx \Delta F = F_z - F_{z_0}. \quad (3.8)$$

Thus, within the SSA the PMF of the unbiased system is well approximated by the free energy difference of the system biased by the HGP. To make sure that the distance between the target  $z(t)$  and actual  $\tilde{z}$  values of the RC on average stays smaller than the desired  $\delta z$ , one needs to choose the spring constant according to

$$k > \max \left\{ \frac{2\alpha}{(\delta z)^2}, \frac{2U_{\max}}{(\delta z)^2} \right\}, \quad (3.9)$$

where  $U_{\max}$  is the highest PMF barrier one wants to explore, and  $\alpha \gg 1$ .

### 3.5 PMF Using Umbrella Sampling and WHAM

In umbrella sampling[57, 61, 62, 85], the range of RC values of interest is divided into  $N_w$  *sampling windows* centered about conveniently chosen values  $z_i$ ,  $i = 1, \dots, N_w$ . Next, the reaction coordinate is sampled in each window separately by preparing identical replicas

of the system and applying the harmonic guiding potential  $V_{z_i}(\tilde{z})$ . As a result, the biased distribution functions can be readily obtained by direct sampling of the reaction coordinate for the biased system, i.e.,  $p_i(z) = (Z_0/Z_i) e^{-V_i(z)} p_0(z)$ , where, for brevity, the index  $z_i$  has been replaced by  $i$ . By inverting this equation, the equilibrium distribution in each window can be expressed in terms of the biased distribution of the reaction coordinate. The standard method for efficiently stitching together the biased  $p_i(z)$ 's in order to obtain the equilibrium  $p_0(z)$ , and therefore the sought PMF, is the so called *weighted histogram analysis method* or WHAM [85], according to which

$$p_0(z) = \frac{\sum_{i=1}^{N_w} \mathcal{N}_i p_i(z)}{\sum_{i=1}^{N_w} \mathcal{N}_i e^{-V_i(z)} / \langle e^{-V_i} \rangle}, \quad (3.10a)$$

$$\langle e^{-V_i} \rangle = \int dz p_0(z) e^{-V_i(z)}, \quad (3.10b)$$

with  $\mathcal{N}_i$  the number of data points used to construct  $p_i(z)$ . The above non-linear coupled WHAM equations, that need to be solved iteratively, minimize the errors in determining  $p_0(z)$ . When applicable, US combined with WHAM is perhaps the best choice for calculating PMFs. In practice, however, one often encounters situations in which the minimum number of US windows required to properly cover the range of RC values of interest is excessively large and the application of the method may become computationally unattainable. Molecular and ion transport through channel proteins is a good example. Potassium transport through gramicidin A channel, and glycerol transport through the GlpF channel are discussed in detail in Chp. 4.

## 3.6 The Transient Fluctuation Theorem and Jarzynski's Equality

In steered molecular dynamics (SMD) simulations[63], where initially the system is in an equilibrium state characterized by  $z(0)$ , the target value of the RC  $z(t)$  is varied in time according to a given protocol. For example, in constant velocity SMD (cv-SMD)  $z(t) = z(0) + vt$ ,  $0 \leq t \leq \tau$ , where  $v$  is the constant pulling speed equal to the ratio of the total pulling distance to the desired simulation time  $\tau$ . The SMD pulling paths of the system when  $t$  increases from 0 to  $\tau$  are referred to as forward (F) paths. The time reverse (R) pulling paths are obtained by starting the system from an equilibrium state corresponding to  $z(\tau)$  and reversing the sign of  $t$  in  $z(t)$  for F paths:  $z_R(t) = z_F(\tau - t) = z(\tau) - vt$ ,  $0 \leq t \leq \tau$ . The choice of a sufficiently large spring constant (see Sec 3.4) in this time dependent HGP [Eq. (3.4)] guarantees that the instantaneous RC,  $\tilde{z}(t)$ , follows closely the target value  $z(t)$  during the pulling process. Thus, cv-SMD is a fast sampling method of the RC by driving the system out of equilibrium. The fastest pulls drive the system out of equilibrium the most. The work done during a cv-SMD simulation is given by

$$W_t \equiv W_z = \int_{z_0}^{z(t)} dz [\partial V_z(\tilde{z}) / \partial z] = k \int_{z_0}^{z(t)} dz (z - \tilde{z}) . \quad (3.11)$$

Crooks has shown that under rather general conditions, listed at the beginning of this section, the following nonequilibrium fluctuation theorem holds[83]

$$\langle f(W) e^{-W_{dF}} \rangle_F = \langle f(-W) \rangle_R , \quad (3.12a)$$

or

$$\langle f(W) \rangle_F = \langle f(-W) e^{-W_{dR}} \rangle_R . \quad (3.12b)$$

Here  $f(W)$  is an arbitrary function of the work  $W$ , and

$$\langle \dots \rangle_{F/R} = \int dW P_{F/R}(W) \dots , \quad (3.13)$$

represents the average over forward/reverse paths or, equivalently, the average with respect to the forward/reverse work distribution functions  $P_{F/R}(W)$ . The dissipative work in a F/R process is given by

$$W_{dF/R} = W_{F/R} \mp \Delta F , \quad (3.14)$$

with  $\Delta F = F_{z(\tau)} - F_{z(0)}$ . The Jarzynski equality (JE) follows immediately from Eqs. (3.12) by setting  $f(W) = 1$ , and it can be written in any of the following forms:

$$\langle \exp(-W_{dF}) \rangle_F = \langle \exp(-W_{dR}) \rangle_R = 1 , \quad (3.15a)$$

$$\langle \exp(-W) \rangle_F = e^{-\Delta F} , \quad \langle \exp(-W) \rangle_R = e^{\Delta F} . \quad (3.15b)$$

Another important equality, that connects the F and R work distribution functions, can be derived from Eqs. (3.12) by setting  $f(W') = \delta(W - W')$  and carrying out the integral with respect to  $W'$ . The result is Crooks' transient fluctuation theorem (TFT)

$$\frac{P_F(W)}{P_R(-W)} = e^{W_{dF}} . \quad (3.16)$$

This equation is used to derive the new results in Sec. 3.8.

### 3.7 PMF from Unidirectional SMD Pulls

A very popular alternative for calculating PMFs is based on the application of the JE from repeated unidirectional nonequilibrium SMD simulations [5, 64, 73–75, 82, 84, 86–

88]. Within the SSA the sought PMF can be readily obtained from Eqs. (3.8) and (3.15b)

$$\Delta U(z) \approx \Delta F = -\log \langle \exp(-W_z) \rangle_F . \quad (3.17)$$

Here the index  $F$  indicates that the average is taken over the ensemble of forward pulling paths. As already mentioned, the average of the exponential in Eq. (3.17) cannot be estimated reliably even for a reasonably large number of SMD pulls, unless the pulling speed is sufficiently small so that the system is close to equilibrium along the pulling paths. This is due to the fact that the overlap between  $\exp(-W)$  and the sampled part of  $P_F(W)$  is in general exponentially small. Nevertheless, there exist two approaches that in principle may give fairly good estimates of Eq. (3.17), provided that the system is not too far from equilibrium during pulls.

The first method is the cumulant approximation (CA) [5, 64, 84], according to which

$$\Delta U(z) = -\log \langle \exp(-W_z) \rangle \approx \overline{W}_z - \sigma_z^2/2 , \quad (3.18a)$$

$$\sigma_z^2 = \overline{W_z^2} - \overline{W}_z^2 , \quad (3.18b)$$

where for simplicity we have dropped the index “F” and  $\sigma_z^2$  is the variance (2<sup>nd</sup> cumulant) of the work. It has been shown that within SSA the work distribution function  $P_F(W)$  is Gaussian, and therefore that in this case the cumulant approximation (3.18) in fact is exact. However the reason why in practice Eq. (3.18) is valid only close to equilibrium is because SMD pulling paths can sample only a narrow region about the peak of the Gaussian  $P_F(W)$ . This allows for a fairly accurate determination of the mean work  $\langle W_z \rangle$  but, in general, seriously underestimates the variance  $\sigma_z^2$ .

The second method for evaluating the average in Eq. (3.17) is a weighted histogram approach suggested by Hummer and Szabo [74, 75], and indirectly by Crooks [83]. The



nonequilibrium fluctuation theorem due to Crooks can also be written as

$$\langle f[z(t)] \exp(-W_d) \rangle_F = \langle f[z_R(0)] \rangle_R = \langle f[z(t)] \rangle_{eq} \quad (3.19)$$

where  $z_R(t)$  represents the time evolution of the control parameter during reverse pulls,  $f[z]$  is an arbitrary function and the index “eq” means the equilibrium average corresponding to the biased system with Hamiltonian  $H_{z_0}$ . By inserting  $f[z] = \delta(z - \tilde{z})$  into Eq. (3.19) one obtains

$$\begin{aligned} \langle \delta(z - \tilde{z}) e^{-W_{z'}} \rangle_F &= \frac{Z_0}{Z_{z_0}} \left\langle \delta(z - \tilde{z}) e^{-V_{z'}(\tilde{z})} \right\rangle_0 \\ &= \frac{e^{-V_{z'}(z)}}{\langle e^{-V_{z_0}} \rangle_0} e^{-U(z)} , \end{aligned} \quad (3.20)$$

Since the equilibrium average  $\langle \exp(-V_{z_0}) \rangle_0$  corresponding to the unbiased system contributes only an additive constant to the PMF, from Eq. (3.20) one obtains the following result

$$U(z) = -\log \langle \delta(z - \tilde{z}) \exp(-\Delta W_{z'}) \rangle , \quad (3.21a)$$

where

$$\begin{aligned} \Delta W_{z'} &= W_{z'} - V_{z'}(\tilde{z}) \\ &= k \int_0^t d\tau \dot{z}'(\tau) [z'(\tau) - \tilde{z}(\tau)] - \frac{k}{2} [z'(\tau) - \tilde{z}(\tau)]^2 . \end{aligned} \quad (3.21b)$$

Thus,  $U(z)$  can be calculated from the work time series obtained in repeated cv-SMD simulations by constructing a weighted histogram of the RC according to Eqs. (3.21). This method resembles to the US and WHAM and is preferable to the cumulant approximation method whenever we have a large number of pulling paths. However, in the case of large systems when only a limited number of trajectories can be sampled this method is inapplicable because of insufficient data.

### 3.8 PMF from Forward and Reverse SMD Pulls

This section presents a new method we propose for calculating PMFs from few fast SMD pulls along the RC in both F and R directions, hereafter referred to as the *FR method* [52]. It is assumed that the pulls are done with a sufficiently stiff spring such that the SSA holds (Sec. 3.4). In this case, the F work distribution  $P_F(W)$  is Gaussian, and according to Crooks' TFT (3.16) it follows that the R work distribution  $P_R(W)$  is also Gaussian. Thus one can write

$$P_{F/R}(W) = \left(2\pi\sigma_{F/R}^2\right)^{\frac{1}{2}} \exp\left[-\frac{(W - \overline{W}_{F/R})^2}{2\sigma_{F/R}^2}\right] \quad (3.22)$$

where  $\overline{W}_{F/R}$  and  $\sigma_{F/R}^2$  are the mean work and variance corresponding to the F and R pulling directions, respectively. The mean dissipative work in the two distinct pulling directions is

$$\overline{W}_{dF/R} = \int dW (W \pm \Delta F) P_{F/R}(W) = \overline{W}_{F/R} \mp \Delta F. \quad (3.23)$$

Inserting (3.22) into (3.16) and taking into account that  $W_{dF} = W - \Delta F$ , after little algebra it follows that TFT can hold only if

$$\sigma^2 \equiv \sigma_F^2 = \sigma_R^2 = \overline{W}_F + \overline{W}_R \quad (3.24a)$$

and

$$\Delta F = (\overline{W}_F - \overline{W}_R)/2. \quad (3.24b)$$

Finally, inserting Eq. (3.24a) into (3.23), one finds that the mean dissipative work is the same in both F and R pulling directions, i.e.,

$$\overline{W}_d \equiv \overline{W}_{dF} = \overline{W}_{dR} = (\overline{W}_F + \overline{W}_R)/2. \quad (3.24c)$$

Equations (3.24) are the key formulas of the FR method for calculating PMFs from fast F

and R SMD pulls. Assuming that a few ( $\sim 10$ ) such SMD pulls can sample reasonably well the work about the peak position  $\bar{W}_{F/R}$  of  $P_{F/R}(W)$ , as indicated by the shaded regions in Fig. 3.1, then Eqs. (3.24) yield essentially with the same degree of accuracy both the desired PMF,  $\Delta U \approx \Delta F$ , and the mean dissipative work,  $\bar{W}_d$ . This feature makes the proposed method superior to the currently used approaches described in the previous sections. In fact, these other methods can only determine the mean total work  $\bar{W}_F$  with some statistical correction either through the cumulant approximation or a weighted histogram method. Furthermore, since it is reasonable to assume that  $\bar{W}_d$  is proportional to the pulling speed  $v$ , one can readily determine the position dependent friction coefficient  $\gamma(z)$  from the slope of the mean dissipative work  $\gamma(z) = (d\bar{W}_d(z)/dz)/v$ . Then, the corresponding diffusion coefficient is given by the Einstein relation (in  $k_B T$  energy units)

$$D(z) = \gamma(z)^{-1} = v (d\bar{W}_d(z)/dz)^{-1} . \quad (3.25)$$

Now that both  $U(z)$  and  $D(z)$  are determined, the equation of motion of the RC on a meso (or macro) time scale is given by the Langevin equation corresponding to an overdamped Brownian particle[56]

$$\gamma(z)\dot{z} = -dU(z)/dz + \xi(t) , \quad (3.26a)$$

or equivalently, the corresponding Fokker-Planck equation for the probability distribution function  $p(z, t)$  of the RC

$$\partial_t p(z, t) = -\partial_z j(z, t) = \partial_z D(z) \partial_z p(z, t) + \partial_z U'(z) p(z, t) , \quad (3.26b)$$

where  $\xi(t)$  is the Langevin force (modeled as a Gaussian white noise) and  $j(z, t)$  is the probability current density.

It is extremely important to emphasize that far from equilibrium the variance  $\sigma_W^2 \equiv \sigma_z^2$  of the F/R work calculated from SMD pulls data [cf. (3.18)] is in general much smaller

than the variance  $\sigma^2$  of the actual work distribution function, and therefore it cannot be used to estimate even approximately the mean dissipative work, unless an exponentially large number of SMD trajectories are collected and used for this purpose.

Also note that  $P_F(W)$  and  $P_R(W)$  are identical Gaussians centered about  $\bar{W}_F$  and  $\bar{W}_R$ , respectively. One can also define a distribution function for the dissipative work through  $P_d(W) = P_F(W + \Delta F) = P_R(W - \Delta F)$ , which is centered about  $\bar{W}_d$  (Fig. 3.1). This allows to calculate the fraction of the SMD trajectories that violates the second law, i.e., for which  $W_d < 0$ ; these trajectories are crucial in establishing the validity of the JE. Thus,

$$\begin{aligned} \langle e^{-W_d} \rangle |_{W_d < 0} &= \int_{-\infty}^0 dW P_d(W) e^{-W} \\ &= \frac{1}{2} \text{erfc}(\bar{W}_d^{1/2}) \sim \frac{\exp(-\bar{W}_d)}{\bar{W}_d^{1/2}}, \end{aligned} \quad (3.27)$$

which clearly indicates that for  $\bar{W}_d > 1$  (i.e.,  $\bar{W}_d > k_B T$  in SI units) the number of such trajectories is exponentially small, and finding any of them in SMD simulations of large biomolecules is rather unlikely.

### 3.9 Generalized Acceptance Ratio Method

The idea of combining results from both F and R simulations is not new, dating back to the original Bennett’s *acceptance ratio method* [89]. However, in previous such studies [90–95] the focus was mainly on determining the free energy difference between two states and to estimate the corresponding error, unlike in the FR method in which the PMF, the mean dissipative work and the corresponding diffusion coefficient are determined simultaneously from specially designed F and R pulls with Gaussian distributed work. For example, starting from the nonequilibrium fluctuation theorem (3.12) and following the general philosophy of the Bennett acceptance ratio method, Crooks has shown [83] that the

best estimate (i.e., with smallest error) of the free energy [see Eqs. (3.12)]

$$e^{-\Delta F} = \langle f(W) \rangle_F / \langle f(-W) e^{-W} \rangle_R \quad (3.28)$$

is obtained by choosing the  $f(W) = 1/[1 + n_F/n_R \exp(W - \Delta F)]$ , where  $n_{F/R}$  represent the number of F/R paths sampled. Essentially the same result was derived by Pande and collaborators [94] by applying the *maximum likelihood estimator* (MLE) method to Crooks' TFT (3.16). Thus, the best estimate of the free energy difference  $\Delta F$  between two equilibrium states corresponding to the RCs  $z(0)$  and  $z(t)$  is given by the solution of the following transcendental equation

$$\sum_{i=1}^{n_F} \frac{1}{1 + n_F/n_R \exp(W_{Fi} - \Delta F)} - \sum_{i=1}^{n_R} \frac{1}{1 + n_R/n_F \exp(-W_{Ri} - \Delta F)} = 0. \quad (3.29)$$

To calculate the PMF  $U(z)$  along the RC,  $z$ , by using the above MLE method, first, one needs to divide the domain of interest  $\{z_{min}, z_{max}\}$  into  $N$  intervals determined by the division points  $z_i$ ,  $i = 0, \dots, N$ . Then the system needs to be steered into these points via SMD, and in each of them it needs to be equilibrated. Then, depending on the available computational resources, a well defined number of F and R cv-SMD pulls should be carried out between adjacent division points, each time starting from a different equilibrium configuration. Finally, solving Eq. (3.29) within the SSA, one determines the change  $\Delta U_i = U_i - U_{i-1}$  along each segment  $(z_{i-1}, z_i)$ . Although the above methods that combine F and R SMD pulls can determine only the free energy difference between initially equilibrated states, in practice was found that in many cases Eqs. (3.24b)-(3.24c) give good results even between the division points  $z_i$ . This means that  $N$  does not need to be a large number, so the computational overhead due to the intermediate equilibrations can be significantly reduced.

## 3.10 Helix-to-Coil Transition of Deca-Alanine

The efficiency of the “FR method” was demonstrated by calculating the PMF and the diffusion coefficient of single-file water molecules in single-walled carbon nanotubes [52]. The obtained results were found to be in very good agreement with the results from other PMF calculation methods, e.g., umbrella sampling [57, 61, 62].

The viability of the “FR method” is further tested by applying it to calculate the PMF of deca-alanine as a function of its end-to-end distance when it undergoes the helix-to-coil transition. The computed PMF and diffusion coefficient can be used to estimate important characteristics of the studied systems, e.g., the mean folding time of the stretched deca-alanine.

The computational model of deca-alanine and the performed non-equilibrium steered MD (SMD) simulations are presented in the following subsection. The results of the PMF calculations are described in Sec. 3.10.2.

### 3.10.1 System Modeling and SMD simulations

Deca-alanine is a small oligopeptide composed of ten alanine residues (Fig. 3.2). The equilibrium conformation of deca-alanine, in the absence of solvent and coupled to an artificial heat bath at room temperature, is an  $\alpha$ -helix. The system can be stretched to an extended (coil) conformation by applying an external force that pulls its ends apart. Once the stretched system is released it will refold spontaneously into its native  $\alpha$ -helical conformation. Thus, this can be regarded as a simple protein unfolding and refolding problem that can be comfortably studied via SMD simulations due to the relatively small (104 atoms) system size. It is natural to define the reaction coordinate as the distance  $R$  between the first ( $CA_1$ ) and the last ( $CA_{10}$ )  $C_\alpha$  atoms. The PMF  $U(R)$  that describes the energetics of the folding/unfolding process was calculated in two steps: (i) a small number (in general 10) F and R pulling trajectories were generated using SMD simulations, and (ii) these trajectories

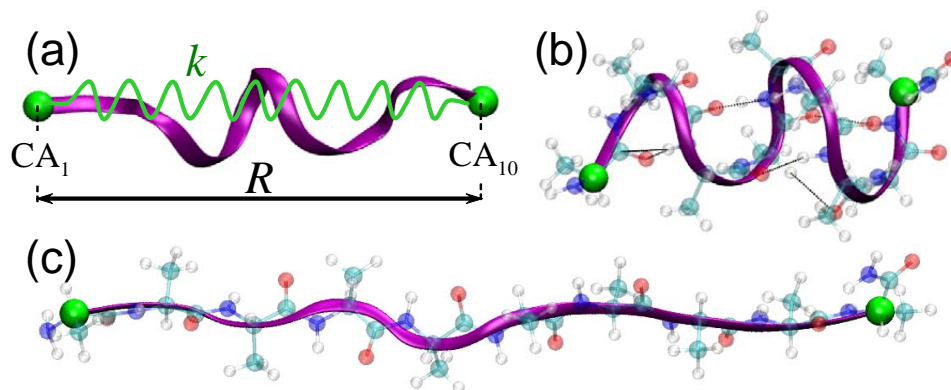


Figure 3.2: (a) Cartoon representation of deca-alanine. The reaction coordinate  $R$  is defined as the distance between the first ( $CA_1$ ) and last ( $CA_{10}$ )  $C_\alpha$  atoms, i.e., the end-to-end distance of the peptide. The spring, with elastic constant  $k$ , connecting  $CA_1$  and  $CA_{10}$  corresponds to an elastic guiding potential  $V(R;t) = (k/2)[R - R_0(t)]^2$  that can be used to cycle deca-alanine between the (b) folded and (c) unfolded (completely stretched) conformations. In (b) and (c) the backbone (sidechain) atoms are shown in cartoon (CPK) representation. The hydrogen bonds that stabilize the  $\alpha$ -helix are also shown in configuration (b).

were used in the PMF calculation methods described in this section, i.e., the “FR method” [Eqs. (3.24)], the JE method [Eq. (3.18)] and the MLE method [Eq. (3.29)]. The SMD harmonic guiding potential (3.4) corresponded to an ideal spring of tunable undeformed length  $R(t)$  inserted between  $CA_1$  and  $CA_{10}$  (see Fig. 3.2a).

The molecular model of deca-alanine was built by employing the molecular modeling software VMD [96]. All simulations were performed with NAMD 2.5 [42] and the CHARMM27 force field for proteins [36, 37]. A cutoff of 12 Å (switching function starting at 10 Å) for van der Waals interactions were used. An integration time step of 2 fs was employed by using the SHAKE constraint on all hydrogen atoms [43]. The temperature was kept constant at 300 K by coupling the system to a Langevin heat bath. The system was subjected to several equilibrium MD and non-equilibrium SMD simulations.

The reaction coordinate  $R \in (13, 33)$  Å was divided into ten equidistant intervals (windows) delimited by the points  $R_i = 13 + 2i$ ,  $i = 0, \dots, 10$ . Next, a pool of equilibrium states were generated for each  $R_i$  from 4 ns long equilibrium MD trajectories. These states were

used as starting configurations for the SMD F and R pulls on each of the ten intervals. The spring constant in these equilibrium MD simulations was  $k = 50 \text{ kcal/mol/\AA}^2$ .

The equilibrium length of the folded deca-alanine was determined by two long free MD simulations starting from a compressed  $R = 13 \text{ \AA}$  and the completely stretched  $R = 33 \text{ \AA}$  configurations of deca-alanine. Both simulations led to the same equilibrium length  $R = 14.5 \text{ \AA}$ .

In order to calculate the PMF  $U(R)$  a total of six sets of F and R SMD simulations were carried out. Ten simulation windows were used in each of the first three sets of SMD runs, but three different pulling velocities:  $v_1 = 1 \text{ \AA/ps}$ ,  $v_2 = 10^{-1} \text{ \AA/ps}$ , and  $v_0 = 10^{-4} \text{ \AA/ps}$  were employed. The sets corresponding to  $v_{1,2}$  consisted of 10 F and 10 R SMD trajectories. For the quasi-equilibrium pulling speed  $v_0$  only one F and R runs were performed. In the last three sets of SMD simulations a single simulation window was used to cover the entire range of the RC, with the same three velocities as in the previous SMD runs. For all six sets of SMD simulations, the stiff-spring constant was set to  $k = 500 \text{ kcal/mol/\AA}^2$ .

To construct the forward and reverse work distribution functions on the segment  $R \in (17, 21) \text{ \AA}$ , 2000 pulls in both F and R directions were performed. In order to generate a sufficient number of starting equilibrium configurations it was necessary to extend the equilibration runs at both  $R = 17 \text{ \AA}$  and  $R = 21 \text{ \AA}$  to 5 ns. All of these simulation used a pulling velocity of  $v = 1 \text{ \AA/ps}$  and a spring constant of  $k = 500 \text{ kcal/mol/\AA}^2$ .

Finally, to estimate the mean refolding time of the completely stretched deca-alanine 100 free MD simulations starting from an equilibrium configuration corresponding to  $R = 33 \text{ \AA}$  were performed. When deca-alanine reached its folded, equilibrium length of  $14.5 \text{ \AA}$  the simulation was stopped and the refolding time recorded.

### 3.10.2 Potential of Mean Force

The PMFs  $U(R)$  calculated using the FR method corresponding to the six different pulling protocols described in Sec. 3.10.1 are shown in Fig. 3.3. As expected, for the very small



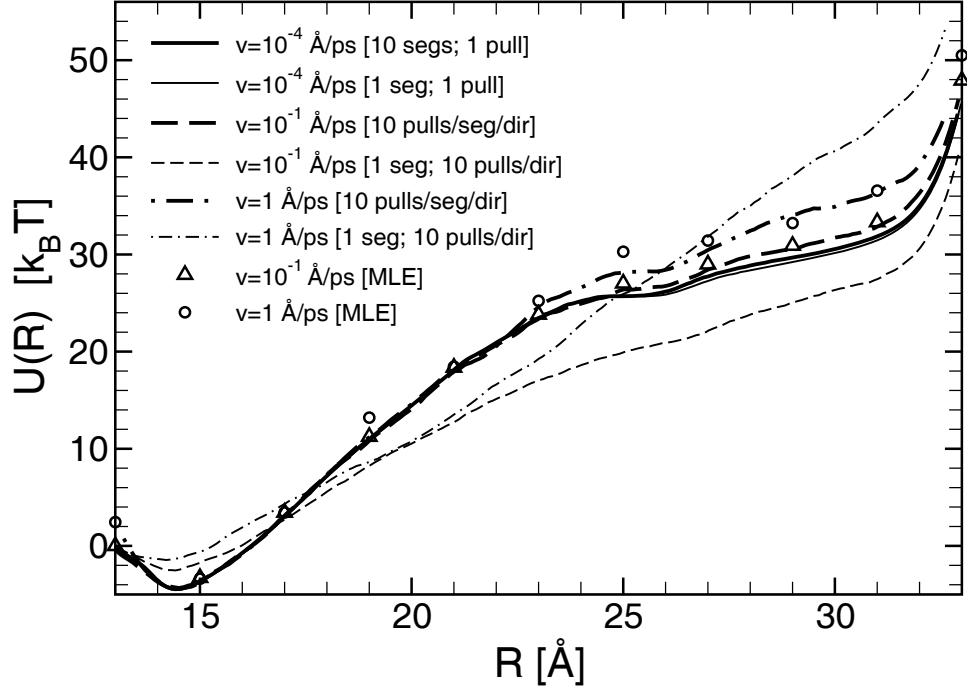


Figure 3.3: Potential of mean force (PMF) of deca-alanine as a function of the reaction coordinate  $R$ . The different curves were obtained with the FR method by employing different simulation and PMF calculation protocols described in the text.

pulling speed  $v_0$  the system is in quasi-equilibrium throughout the SMD runs leading to the same (true) PMF regardless of the number of simulation windows considered. However, while the dissipative work is negligible for both F and R processes, repetition of these simulations resulted in different PMFs for  $R > 24$  Å, and it will be discussed below (see also Fig. 3.6). Not surprisingly, in case of the very fast pulling speed  $v_1$ , the PMF for the single simulation window is rather poor along  $R$  except, as expected, at the end-points of the window. Indeed, the FR method allows one to calculate the PMF difference between two equilibrium states connected by fast F and R SMD processes that follow the same protocol. However, it is remarkable that using ten simulation windows, even at this large pulling speed the resulting PMF is rather close to the real one. For the still fast pulling speed  $v_2$  the situation is similar. While the single simulation window case lead to a rather poor PMF (though somewhat better than in the  $v_1$  case), the ten simulation windows case give a PMF that is almost indistinguishable from the true PMF. For comparison, the PMFs

calculated at  $R_i$ ,  $i = 1, \dots, 10$  using the MLE method for both  $v_1$  and  $v_2$  are also shown in Fig. 3.3. Based on these results one may conclude that the FR method gives very good PMF even for fast pulling speeds and only a few F and R trajectories, provided that a sufficient number of simulation windows are used.

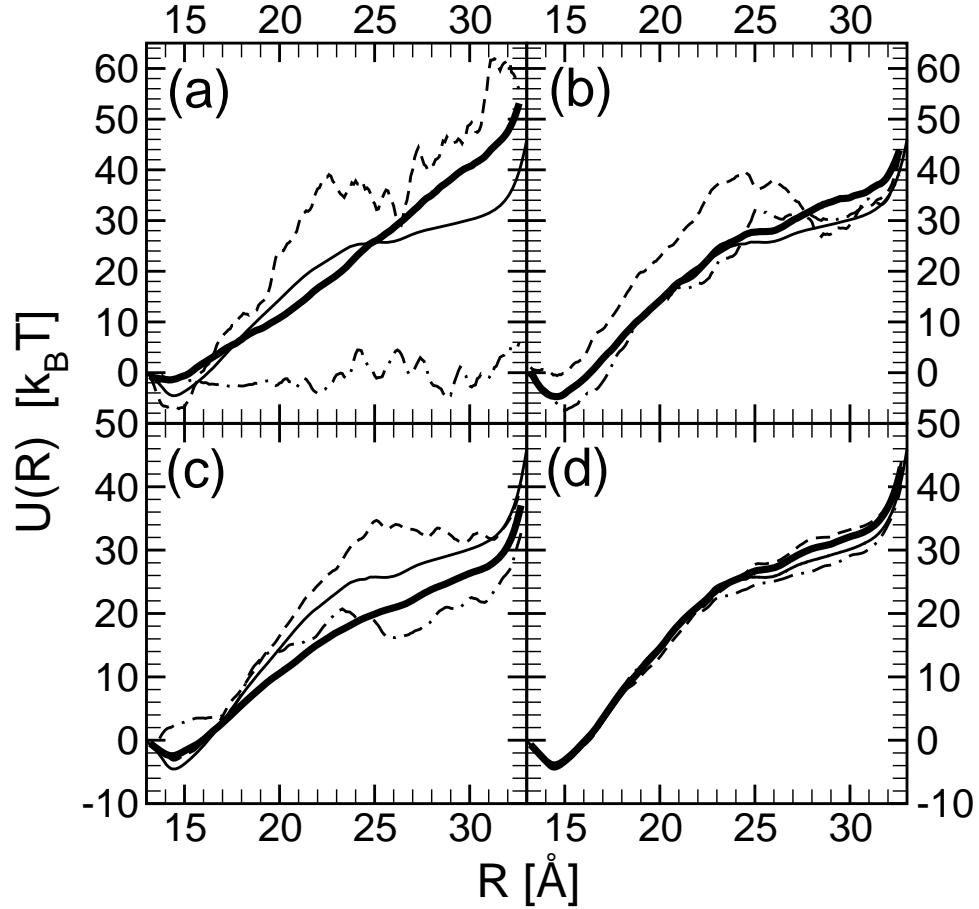


Figure 3.4: Comparison between the PMFs  $U(R)$  obtained using the FR method (thick solid line) and the cumulant approximation of the JE corresponding to the ten forward (dashed line) and reverse (dot-dashed line) SMD trajectories, respectively. The thin solid line corresponds to the exact PMF. The upper (lower) panels correspond to a uniform pulling speed of  $1 \text{ \AA/ns}$  ( $0.1 \text{ \AA/ns}$ ). The PMFs in the right panels were determined by dividing the  $20 \text{ \AA}$  pulling distance into ten equidistant segments (the system being equilibrated in each of the end points of the individual segments), while the PMFs in the left panels were determined by considering the entire pulling distance as a single segment.

A comparison between the PMFs  $U(R)$  obtained from the FR method and the cumulant approximation of the JE method (applied separately for the F and for the R SMD trajec-

ries) are shown in Fig. 3.4. The FR method provides better PMF in all cases, but especially in those with one simulation window (Fig. 3.4a and c) and for the very large pulling speed  $v_1$  (Fig. 3.4 a and b). For the ten simulation windows with pulling speed  $v_2$  (Fig. 3.4 d) the FR and JE methods are comparable though even in this case the JE F (R) method systematically over (under) estimates the PMF. An average of the JE PMFs for the F and R trajectories would lead a result very close to the FR one.

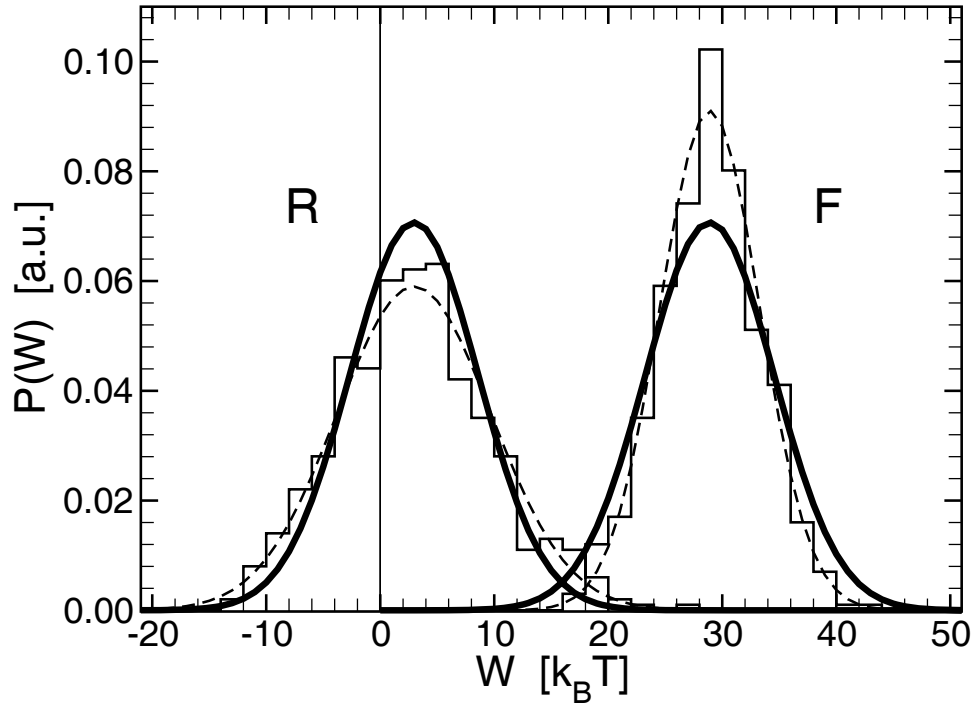


Figure 3.5: Histogram of the distribution functions (thin solid lines) of the forward ( $W_F$ ) and reverse ( $W_R$ ) works along the segment  $R \in (17, 21)$  Å. Although the histograms seem to be Gaussian (dashed lines) they are not identical as predicted by the FR method (see text for details).

An important prediction of the FR method is that, provided that the stiff-spring approximation holds, the F and R work distributions are identical Gaussians centered about the mean F and R work, and therefore shifted by  $2\Delta U$ . This prediction was tested by determining the work distribution histogram corresponding to 2000 F and a same number of R SMD trajectories corresponding to the RC segment  $R \in (17, 21)$  Å. The results are shown in Fig. 3.5. Although the histograms seem to be Gaussian (dashed lines) they are

not identical as predicted by the FR method. In a previous study [97] the clear deviation from Gaussian of the external work distribution in case of deca-alanine was pointed out and it was attributed to the non-Markovian nature of the underlying dynamics of the system. However, in this case the work distributions look Gaussian and the relatively small but clearly noticeable difference between them may be due either to the failure of the stiff-spring approximation or to incomplete sampling. After all, the end-to-end distance is a poor and insufficient reaction coordinate for describing the folding and unfolding processes of a polypeptide.

This last point becomes rather clear when the system is subjected to repeated folding (R) and unfolding (F) at the quasi-equilibrium speed  $v_0$ . As already mentioned, at this speed the system is at almost equilibrium throughout the SMD pulls and one expects that the PMF is given by the external work, i.e., the energy dissipation (which is a stochastic quantity) is negligible. While for  $R < 24 \text{ \AA}$  one obtains systematically the same result, for  $R \in (24, 33) \text{ \AA}$  one obtains different PMFs depending on the direction of pulling as shown in Fig. 3.6b. A careful inspection of these trajectories reveal that the folding and unfolding processes occur through different, globally differentiable pathways in the mentioned range of  $R$ . Thus, it appears that  $R$  is not sufficient to specify the metastable intermediate states of the system, and therefore, one needs extra order parameters, e.g., H-bond distribution. Indeed, the dynamics of the formation and rupture of the H-bonds during folding and unfolding, respectively, is rather different. As shown in the inset snapshots in Fig. 3.6b, the formation of the six H-bonds during the R process is much more homogeneous than their rupture during the corresponding F process. This observation is reinforced by the time dependence of the average H-bonds in deca-alanine shown in Fig. 3.6a. Thus, there are at least two distinctive pathways in the helix-to-coil transition of deca-alanine, both being explored during quasi-static pulls. During fast pulling, however, one of the pathways is preferred compared to the other.

Finally, as an application of the determined PMF and the diffusion coefficient, which

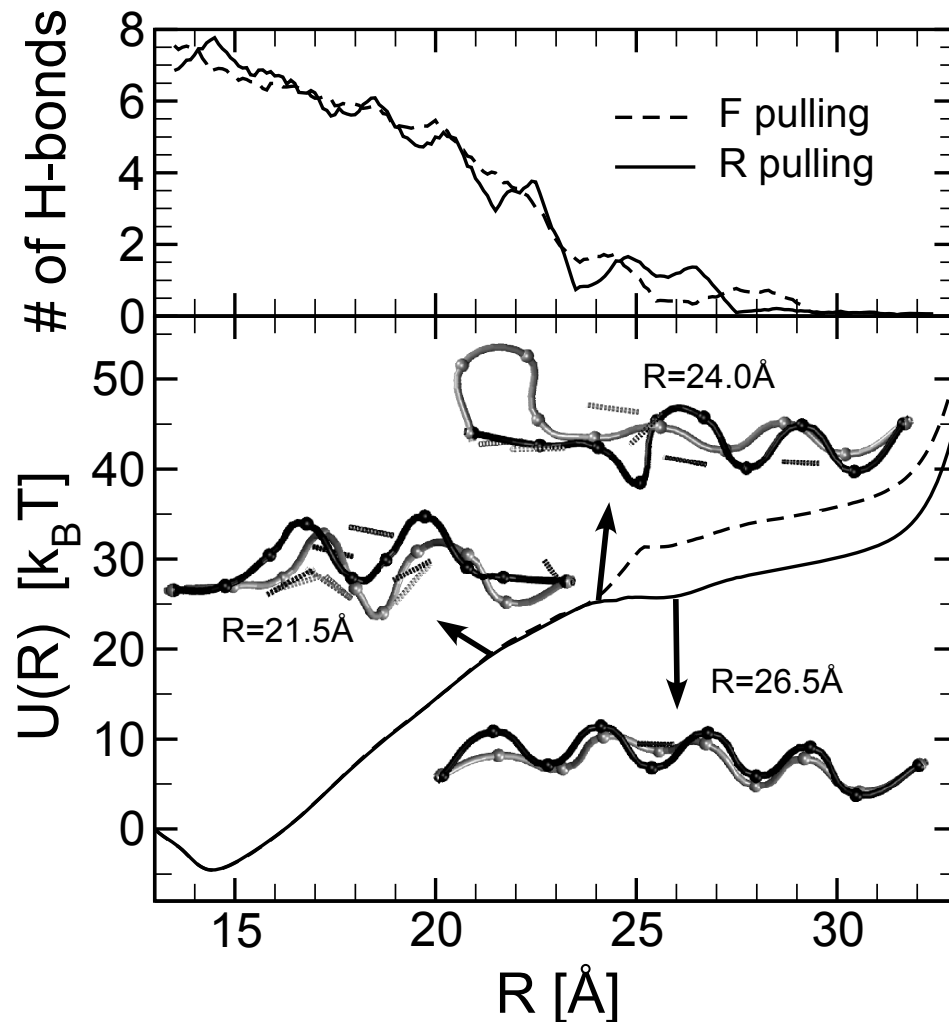


Figure 3.6: (a) Variation of the number of hydrogen bonds in deca-alanine during the quasi-equilibrium ( $v = 10^{-4}$  Å/ns) F and R pulls. (b) The PMF  $U(R)$  calculated as the external work done during the quasi-equilibrium F (dashed line) and R (solid line) pulls. The discrepancy between the two PMFs is most likely due to the difference on how the H-bonds are formed and destroyed during the forced folding and unfolding processes, respectively, as indicated in the inset snapshots of the peptide. Dark (light) color corresponds to the R (F) process.

was found to be approximately constant  $D \approx 0.27 \text{ \AA}^2/\text{ps}$ , the mean folding time (i.e., coil-to-helix transition) was calculated by employing Eq. (4.1). The theoretical result of  $\tau \approx 140 \text{ ps}$  compares rather well with the corresponding  $\approx 100 \text{ ps}$  obtained from the 100 free MD refolding simulations described in Sec. 3.10.1.

### 3.11 Conclusions

In this chapter we have described several methods for calculating PMFs, with emphasis on approaches that require nonequilibrium simulations of biological systems. Due to their large sizes, the sampling of the reaction coordinates would take prohibitively long time using EMD simulations and therefore determining the PMF would be impossible.

It was shown that by employing Crooks' TFT [83] within the stiff spring approximation the potential of mean force along a suitably chosen reaction coordinate can be determined (at least semiquantitatively) from combining a few fast forward and time reversed nonequilibrium processes started from an equilibrium configuration and subject to the same evolution protocol of the reaction coordinate. In the proposed FR method one determines simultaneously both the PMF ( $U$ ) and the mean dissipative work ( $\overline{W}_d$ ) without invoking JE. In fact, JE is not even satisfied for fast F or R pulls simply because processes with negative dissipative work (that transiently violate the second law and are exponentially small in number) are not sampled. The FR method is based on a key observation involving Crooks' TFT (which is more general than JE): whenever the F work distribution function  $P_F(W)$  is Gaussian (e.g., in the case of the stiff-spring approximation) then  $P_R(W)$  is also Gaussian. Furthermore,  $P_{F/R}(W)$  have the same width and are shifted by precisely twice the corresponding free energy difference between the equilibrium states connected by the F and R processes. Thus, both  $U$  and  $\overline{W}_d$  can be readily determined from the mean F and R works ( $\overline{W}_{F/R}$ ). The practical success of the FR method stems from the fact that the mean work  $\overline{W}_{F/R}$  can be measured rather accurately from only a few fast F/R pulls. This also explains

why previous methods, based on the direct application of JE, fail to work away from equilibrium, making them inefficient for practical applications. Indeed, the width of  $P_{F/R}(W)$ , which is proportional to  $\overline{W}_d$ , cannot be determined even approximately from a few unidirectional pulls, unless these are close to equilibrium and rendering  $P_{F/R}(W)$  sufficiently narrow.

To test its viability, the FR method was applied to determine the PMF of a non-trivial, but fairly small system, the deca-alanine. The computed PMF is found to be in good agreement with previous results obtained by using either the Jarzynski equality or the umbrella sampling method. However, compared to these PMF calculation methods this new approach is about one order of magnitude faster and, in addition, also provides the position dependent diffusion coefficient along the RC. By employing the computed PMF and diffusion coefficient in a suitable stochastic model important characteristics of the studied systems, e.g., the mean folding time of the stretched deca-alanine were estimated.

The FR method's feasibility is truly emphasized in Chp. 4, where it is used in the study the transport of potassium ion through Gramicidin A channel and of glycerol molecule through the glycerol uptake facilitator (GlpF) channel.

# **4 Ion and molecular transport through protein channels**



## 4.1 Introduction

Living cells interact with their extracellular environment through the cell membrane, which acts as a protective layer for preserving the internal integrity of the cell. However, the cells metabolism requires molecular and ion transport through this membrane, a function that is fulfilled by a wide variety of transmembrane transporter proteins. While active transporters consume external energy (usually originating from ATP hydrolysis) to conduct molecules along or against the free energy gradient, channel proteins are passive transporters that facilitate the diffusion of specific ions or molecules across the membrane down a free energy gradient.

In order to better understand their transport properties, e.g. to calculate the ion or molecular flux through a protein channel, one needs to model the motion of the transported ion or molecule as an overdamped Brownian particle that diffuses along the axis of the channel in the presence of the potential of mean force created by the protein and environment atoms. The FR method, described in the previous chapter, provides an efficient and sufficiently accurate way to determine simultaneously both the free energy profile and the corresponding position dependent diffusion coefficient of the transported particle along the axis of the channel protein. The transport of potassium ion through the gramicidin A channel and the transport of the glycerol molecule through the glycerol transport facilitator (GlpF) channel are investigated in this chapter.

Gramicidin A is an antibiotic that forms a cation selective channel in lipid bilayers. The 15 alternating Asp and Leu amino acids fold into an unusual barrel-like helix. Two such helices dimerize in a lipid membrane, forming a  $\sim 26$  Å long and 4 – 5 Å wide cylinder-like channel (Fig. 4.1). Being the smallest known ion channel, the gramicidin A is an important testing system for ion permeation models.

The *Escherichia coli* GlpF channel is an aquaporin channel [99] that is specialized in

---

This chapter is based on manuscripts [53] and [98].

transporting not only water, but also small linear sugar alcohol molecules [100, 101], like glycerol. However, the pore is completely impermeable to charged species, such as protons [102]. This remarkable property is critical for the conservation of the electrochemical potential across the membrane. The GlpF channel forms tetramers in the lipid membrane, with glycerol and water molecules present inside the channel [103]. Each monomeric pore has two half-membrane spanning repeats. Half of each repeat is  $\alpha$ -helical and the other half has a specific non-helical structure [103, 104]. The channel is 25 Å long and its diameter measures 3.5 Å at its narrowest region, called the selectivity filter.

## 4.2 Potassium Ion Transport through Gramicidin-A Channel

NMR studies have shown that each end of the channel has a cation binding site [105] that is occupied as the ion concentration is increased. The conductance is at maximum when the average ion population in the channel is one. The backbone carbonyls inside the pore are oriented such that the electronegative oxygen atoms face inward. These oxygens attract the cations, hence the cation (and not anion) selectivity of the channel [106–108].

In spite of the structural simplicity of the gramicidin A, the energetics of the ion transport through the channel is far from trivial. Most of the difficulty arises from the sensitivity to errors due to finite-size effects and from the poor description of the polarization effects by the existing force-fields. Polarization and orientation of the  $\sim 6$  single-file water molecules inside the channel [109] (see Fig. 4.1b) is not a simple linear response to the field of an ion. The linear arrangement of waters seems to play an important role in stabilizing the ion within the channel [110].

Previous potential of mean force (PMF) calculations for potassium ion ( $K^+$ ) through gramicidin A yielded a large central barrier and conductances orders of magnitude below those measured. The observed conductance and concentration dependence can only be

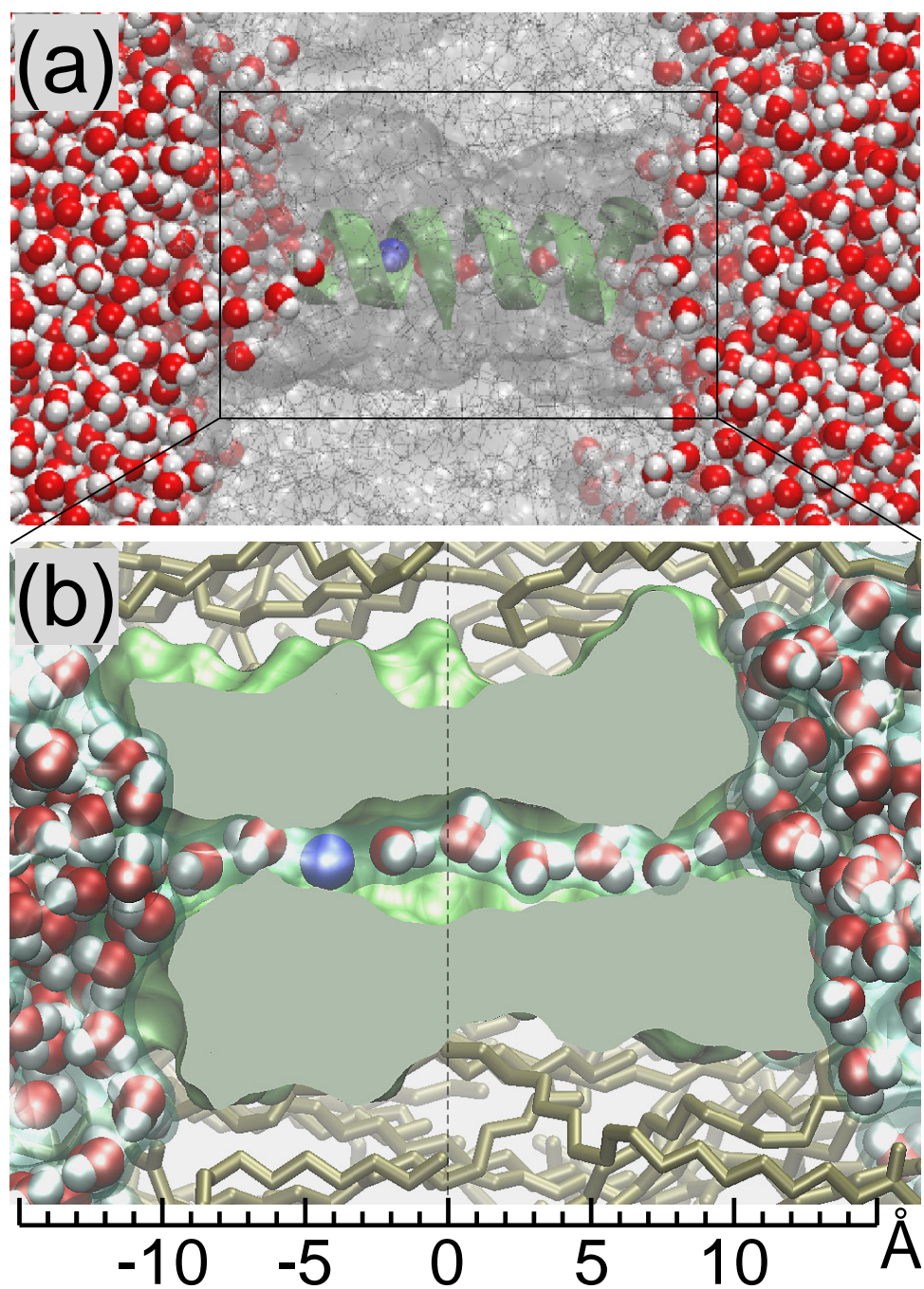


Figure 4.1: (a) Gramicidin A channel ( $\alpha$ -helix dimer colored in green) in POPE lipid bilayer (grey), solvated in water (van der Waals representation). Potassium ions (blue), and the oxygen (red) and hydrogen (white) atoms of water molecules are represented as spheres; (b) Cross-section of the gramicidin A channel. The  $K^+$  ion (blue) and the water molecules move single file inside the pore.

reproduced by  $8 k_B T$  energy wells at the channel ends and  $5 k_B T$  barrier in the middle of the pore [4]. Although PMF calculation methods that try to compensate for finite-size and polarization effects have improved significantly, they continue to yield results that do not match the experimental ones. Most of these methods employ molecular dynamics (MD) simulations with umbrella sampling (US) [1–3] and combined MD simulations with continuum electrostatics theory [111].

Next, the FR method is applied to calculate the PMF of the  $K^+$  ion along the axis of the gramicidin A.

### 4.2.1 Modeling of Gramicidin-A Channel

The computer model of gramicidin A was constructed from its high resolution NMR structure (Protein Data Bank code 1JNO [112]). After adding the missing hydrogens, the structure was energy minimized. Using the VMD [96] plugin *Membrane* the system was immersed into a previously pre-equilibrated patch of POPE<sup>†</sup> lipid bilayer. The membrane patch was a  $\sim 72$  Å sided square. The lipids within 0.55 Å of the protein were removed. The membrane-protein complex was then solvated in water, using the VMD plugin *Solvate*. The final system contained, besides the protein, 155 lipid molecules and 5,700 water molecules (in total 36,727 atoms). An initial 10,000 steps of restrained energy minimization was followed by 0.5 ns of restrained pre-equilibration. The harmonic restraints were applied on the backbone atoms of the protein (spring constant  $k_{bb} = 20$  kcal/mol/Å<sup>2</sup> per atom) and on the phosphorus (P) atoms of the lipid headgroups (spring constant  $k_P \approx 10$  kcal/mol/Å<sup>2</sup> per atom). Thus, the lipids were allowed to arrange themselves in the membrane plane (xy-plane), without changing the thickness of the lipid bilayer ( $\sim 41$  Å). By releasing the P atoms, the lipids were relaxed for 1 ns along the z-axis as well.

Next, a  $K^+$  ion was added at the entrance of the channel. Simultaneously, a  $Cl^-$  counterion was added in the solution, to maintain charge neutrality of the system. A second sys-

---

<sup>†</sup>Abbreviation for 1-Palmitoyl-2-Oleoyl-sn-glycero-3-PhosphoEthanolamine

tem, containing 20 extra pairs of  $K^+$  and  $Cl^-$  (corresponding to 200 mM electrolyte) was created. The added ions were randomly placed using the VMD plugin *Autoionize*. Each ionized system was energy minimized for 10,000 steps and equilibrated for 0.5 ns with the potassium ion placed in three different positions along the channel axis (parallel to the z-axis), namely at  $z \in \{-15, 0, 15\}$  Å. The origin of the z-axis corresponded to the middle of the channel (in between the two helices). In order to prevent the pore from being dragged during the SMD pulls of the  $K^+$  ion, two types of restraints were imposed: (i) backbone atoms restrained to their equilibrium positions (referred to as *fully restrained*); (ii) backbone atoms restrained only along the z-axis and with their center of mass (COM) loosely restrained ( $k_{COM} \approx 1.8$  kcal/mol/Å<sup>2</sup> per backbone atom) (referred to as *z-restrained*).

#### 4.2.2 SMD Simulations

The steered molecular dynamics (SMD) simulations, needed in the PMF calculations using the FR method, were performed on three systems: (S1) backbone of the channel fully restrained with only one pair of  $K^+Cl^-$  ions in the system, (S2) backbone of the channel fully restrained with 200 mM electrolyte concentration, and (S3) backbone of the channel z-restrained and electrolyte concentration 200 mM.

The used SMD pulling protocol was as follows: 10 forward (F) and 10 reverse (R) pulls performed along the channel axis  $z$  on two segments:  $z \in (-15, 0)$  Å and  $z \in (0, 15)$  Å, corresponding to the helical monomers. The velocity of the pulls was  $v = 15$  Å/ns, while the elastic constant of the pulling spring, attached to the  $K^+$ , was  $k = 20$  kcal/mol/Å<sup>2</sup>. The effective point charge of the potassium ion was  $+1.0$  e.

To show that the PMF is independent of the pulling protocol, two more PMFs were calculated when the pulls were performed (i) with a higher velocity  $v = 30$  Å/ns, and (ii) along the z-direction only (ion is free to move in the xy-plane). In order to compensate for the polarization effects, another free energy profile was computed from SMD simulations where the effective point charge of the potassium ion was considered  $+0.5$  e.

### 4.2.3 Potential of Mean Force

Potential of mean force profiles calculated on systems with fully restrained backbones yield very little dependence on the electrolyte concentration (see Fig. 4.2). The PMF barrier decreased by  $\sim 1.5 k_B T$  ( $\sim 3.5\%$ ) when concentration was increased by a factor of 20, from  $\sim 10$  mM (thin solid line) up to 200 mM (dotted line).

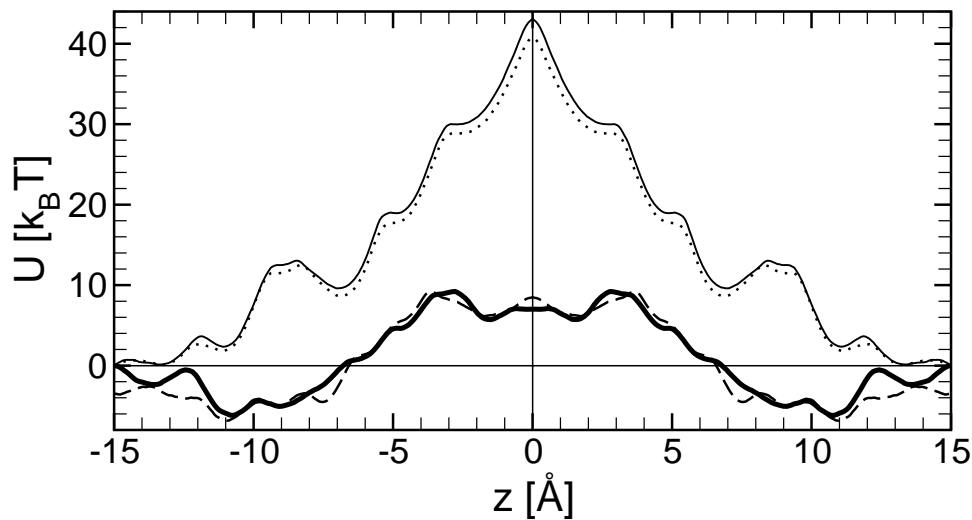


Figure 4.2: Comparison of PMFs obtained for systems with fully restrained protein backbones, but with different electrolyte concentrations:  $\sim 10$  mM ( $1 \text{ K}^+$  and  $1 \text{ Cl}^-$ ) (thin solid line) and 200 mM (21 pairs of ions) (dotted line); PMFs of systems with z-restrained backbones, 200 mM electrolyte, but with different pulling protocols: along the z-direction only (dashed line) and along the channel axis (thick solid line).

As described in the literature [2, 3, 113], the protein flexibility in the membrane plane plays indeed a crucial role in the potassium transport through the gramicidin A channel. The PMF barrier height of  $\sim 40 k_B T$ , obtained for the system that has the protein backbone fully restrained (dotted line), was lowered to only  $\sim 15 k_B T$  when the backbone restraints are imposed only along the z-axis (thick solid line). The large difference indicates that the gramicidin channel equilibrated with water inside is too narrow for the potassium ion to pass. The protein flexibility in the membrane plane allows the channel to locally increase its radius around the  $\text{K}^+$  (mostly due to the electrostatic interactions), assuring a much easier passage of the ion through the pore. In order to allow the potassium ion to sample

the membrane plane (xy-plane), SMD pulls were performed along the z-direction only. The PMF calculated from these pulls (dashed line) overlaps almost perfectly the one obtained from pulls along the channel axis (thick solid line). Because the  $K^+$  occasionally escaped between the two helices into the lipid bilayer during the SMD pulls along the z-direction, the latter PMF (thick solid line) was preferred and used for further comparisons.

The potential of mean force  $U(z)$  was calculated separately for the two segments (corresponding to the  $\alpha$ -helical monomers) using the FR method (see Sec. 3.8). Forward (F) and reverse (R) works are plotted in Fig. 4.3a and Fig. 4.3b, respectively. For better comparison of the F and R works, the direction of the latter was reversed ( $z$  increases with time).

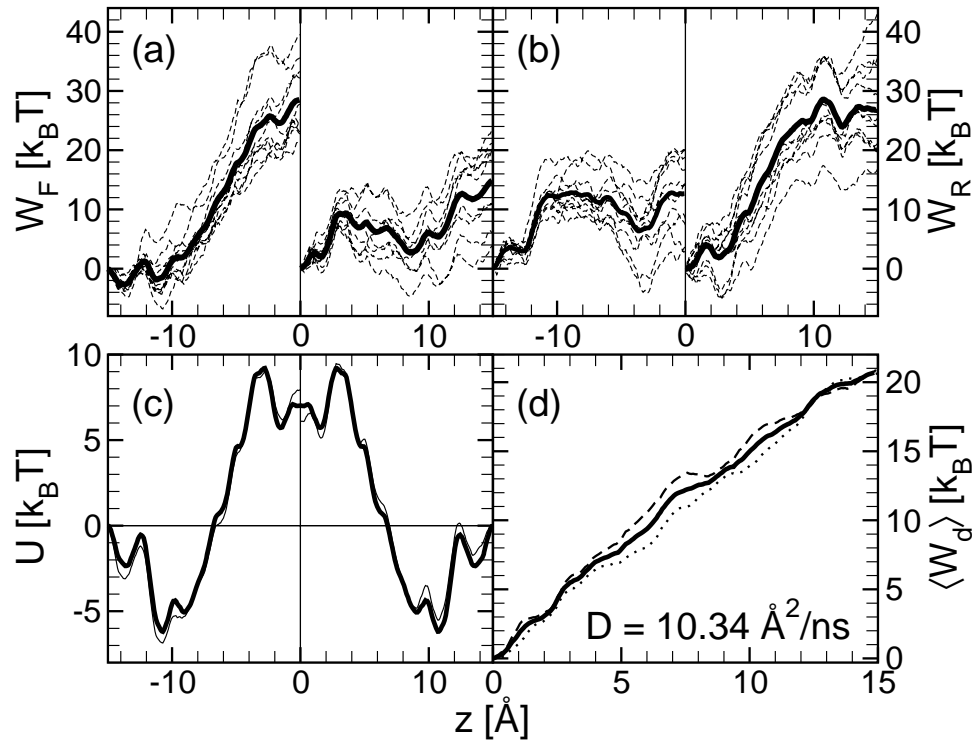


Figure 4.3: Results from the “FR method” when calculating the potential of mean force  $U(z)$  and the corresponding diffusion coefficient  $D(z)$  (velocity of pulls is  $v = 15 \text{ Å/ns}$ ): (a) individual (thin lines) and mean (thick line) work obtained from forward pulls; (b) individual (thin lines) and mean (thick line) work obtained from reverse pulls (explanations in the text); (c)  $U(z)$  (dashed line) and its symmetrized version (solid line); (d) mean dissipative work  $\bar{W}_d(z)$  for the two intervals,  $z \in (-15, 0) \text{ Å}$  (dotted line) and  $z \in (0, 15) \text{ Å}$  (dashed line), and their arithmetic mean (solid line).  $z = 0 \text{ Å}$  corresponds to the middle of the channel and  $z = 15 \text{ Å}$  to its extremities. The linear slope of  $\bar{W}_d(z)$  yields constant  $D = 10.34 \text{ Å}^2/\text{ns}$ .

Due to the symmetry of the channel with respect to its center, the PMF of the two segments (dashed lines in Fig. 4.3c) are nearly mirror-imaged. Therefore, the PMF of the entire channel was obtained by symmetrizing  $U(z)$  with respect to the center of the channel (i.e.  $z = 0 \text{ \AA}$ ) (solid line in Fig. 4.3c). The corresponding mean dissipative works  $\bar{W}_d(z)$  have similar profiles as well (dotted and dashed lines in Fig. 4.3d) ( $z = 0 \text{ \AA}$  corresponds to the center of the channel and  $z = 15 \text{ \AA}$  to the two ends). Their arithmetic average (solid line in Fig. 4.3d) varies almost linearly with  $z$  and, therefore, yields a constant diffusion coefficient  $D \sim 10.3 \text{ \AA}^2/\text{ns}$  [Eq. 3.25)].

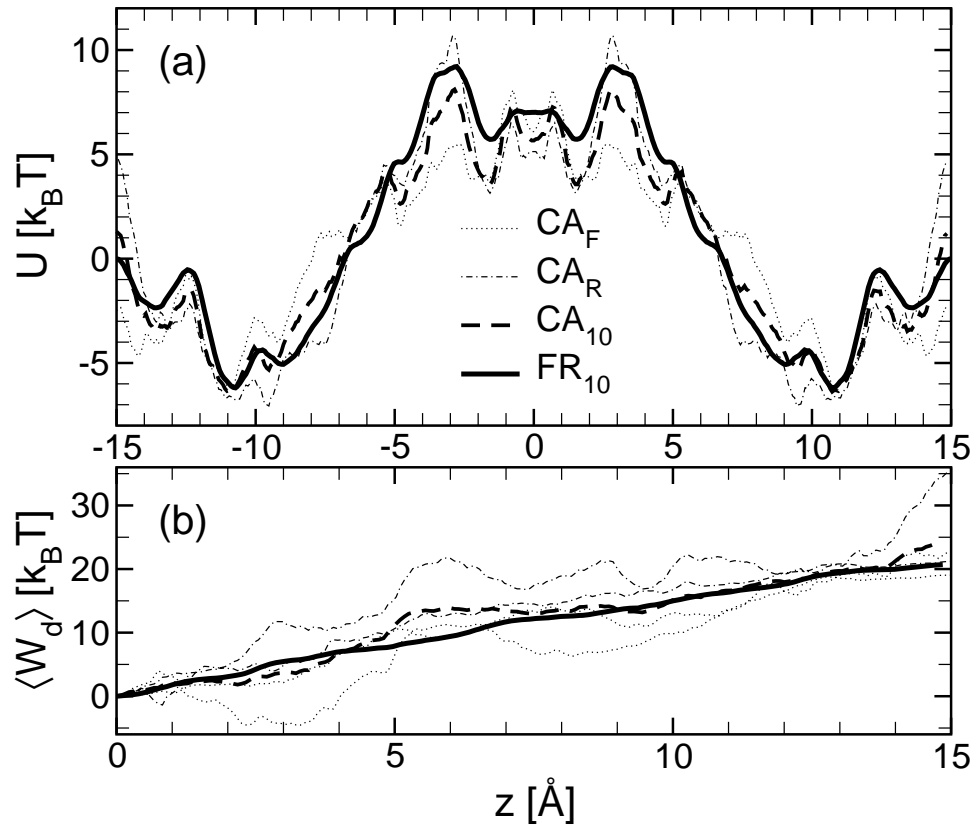


Figure 4.4: Comparison of (a) potential of mean force and (b) mean dissipative work, obtained by applying on 10 SMD pulls the “FR method” (thick solid line) and the Jarzynski equality with the cumulant approximation (CA) (thick dashed line). The CA was employed in both forward (dotted lines) and reverse (dash-dotted lines) directions.

The PMFs obtained by employing Jarzynski’s equality (JE) with the cumulant approximation (CA) separately for forward (F) and reverse (R) pulls are compared to  $U(z)$  cal-



culated using the FR method in Fig. 4.4a. The fact that, for a small set of pulls, JE is valid only in the close-to-equilibrium regime (slow pulls) is indicated by the discrepancies between the F (dotted line) and R (dash-dotted line) PMFs. The two symmetrized PMFs, calculated from 10 high velocity ( $v = 15 \text{ \AA/ns}$ ) pulls, should be one and the same. Instead, they are characterized by differences in barrier height ( $3.5 k_B T$ ) and channel entrance energy level ( $7 k_B T$ ). However, the average of the two PMFs (thick dashed line) is similar to  $U(z)$  calculated using the FR method.

In Fig. 4.4b is shown the comparison of the mean dissipative work ( $\overline{W}_d$ ) calculated using FR method and using JE with CA method. Although the channel is symmetric, the  $\overline{W}_d$  calculated from CA is smaller for forward pulls (dotted lines) than for the reverse ones (dash-dotted lines). The poor sampling of the works yields F and R work distributions with different variances. According to the CA, the  $\overline{W}_d$  is half of the variance of the (assumed Gaussian) work distribution [Eq. 3.18], hence the discrepancy in the forward and reverse  $\overline{W}_d$ . Similarly to the PMF, the average of the F and R  $\overline{W}_d$ , although not linear (thick dashed line), has the same magnitude as the mean dissipative work obtained from the FR method calculations (thick solid line).

The PMF  $U(z)$  calculated using the FR method (thick black line in Fig. 4.5) has two  $\sim 6 k_B T$  wells positioned at the entrances in the channel ( $z \sim \pm 10.8 \text{ \AA}$ ) and two  $\sim 15 k_B T$  barriers positioned near the ends of the monomeric helices that are close to the central part of the channel ( $z \sim \pm 3 \text{ \AA}$ ). A tiny barrier ( $\sim 1.4 k_B T$ ), corresponding to the small distance between the helices, is positioned in the middle ( $z \sim 0 \text{ \AA}$ ) of the  $\sim 3.5 k_B T$  well separating the two main barriers. The calculated profile of  $U(z)$  resembles the reported PMFs [1, 2], obtained by means of umbrella sampling (US) (see Sec. 3.5). Besides the slightly different positioning of the channel entrance wells, there are two important dissimilarities. First, the barrier height of the PMF computed using the FR method is only  $\sim 15 k_B T$  as compared to the  $\sim 20 k_B T$  reported with US. Second, the middle peak of the  $U(z)$  obtained from FR (US) method is  $\sim 2 k_B T$  below (above) the two side peaks.

The black curves in Fig. 4.5 compare  $U(z)$  calculated using the FR method for 10 pulls and  $v = 15 \text{ \AA/ns}$  (thick solid line) with PMFs obtained using only 5 pulls (solid thin line) and higher velocity ( $v = 30 \text{ \AA/ns}$ ) pulls (dashed line). Even having half the number of trajectories and twice faster pulls, the calculated PMF of a  $K^+$  ion inside the gramicidin A channel is essentially the same.

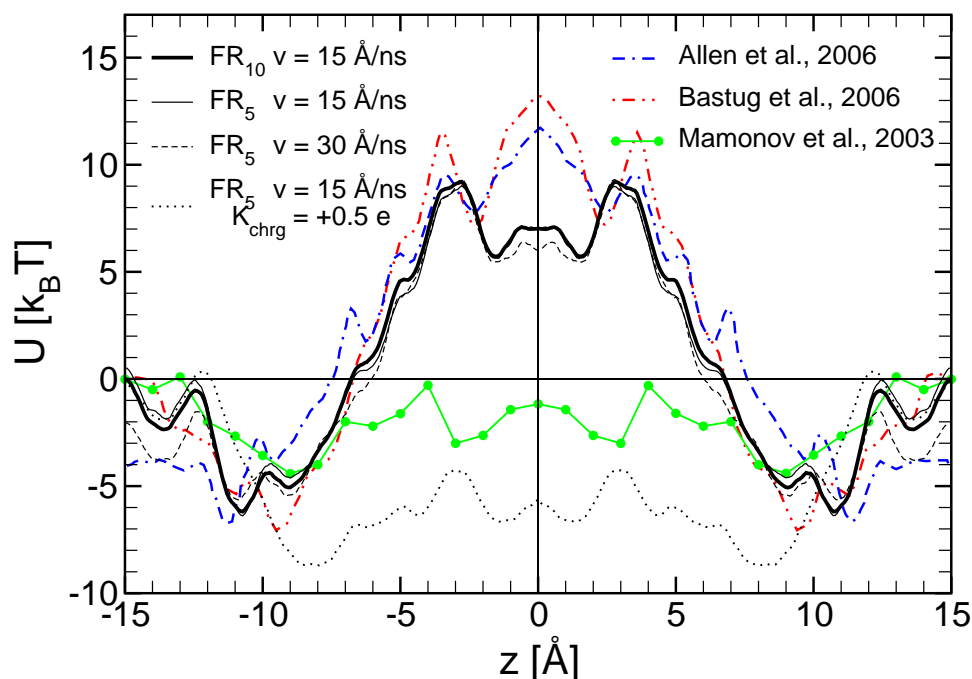


Figure 4.5: Potential of mean force (PMF) calculated using various pulling protocols for the FR method (black lines) compared to PMFs obtained from umbrella sampling (US) method ([2] (red line) and [1] (blue line)) and from Poisson-Nernst-Planck theory combined with MD simulations (PMFPNP) [111] (green bulleted line).

The FR method has nicely reproduced the results reported by US, but the well and barrier sizes are far from the ones needed to reproduce conduction measurements on the gramicidin-A channel, i.e.  $\sim 8 k_B T$  well depth and  $\sim 5 k_B T$  barrier height [4]. Theoretical studies have shown that the main problems in using MD simulations to calculate PMFs for ions through the gramicidin-A channel are the finite-size effects, induced polarization in the lipid hydrocarbons and, most importantly, the polarization of water [1, 114]. The AMBER94 and CHARMM27 (not polarizable) are the force-fields that reproduce best (not

very well though) the  $K^+$  ion's interaction with water and protein. But even after applying corrections for the mentioned issues, the PMF calculated using US has a barrier height that is around  $12 k_B T$  [114]. In order to take into account polarization effects caused by the  $K^+$ , the partial charge of the ion was reduced to half in the MD simulations described in Sec. 4.2.2. By recalculating  $U(z)$  using the FR method (black dotted line in Fig. 4.5), the well position is shifted by  $2.5 \text{ \AA}$  towards the center of the channel and its depth increased to  $8.5 k_B T$ . However, the most dramatic change is the reduction of the barrier height from  $\sim 15 k_B T$  to  $\sim 4.2 k_B T$ . Although the approach is not precise, the well and barrier sizes are very close to their estimated correspondents [4]. The continuum electrostatics approach gives very poor estimates for the  $K^+$  ion conduction through the gramicidin-A channel [4] as well. But a combined continuum electrostatics-MD simulations method, based on the Poisson-Nernst-Planck theory and developed to calculate PMFs (PMFPNP) [111, 115], yields a qualitatively correct PMF profile. Nonetheless, the well and barrier sizes are underestimated by a factor of two (green bulleted line in Fig. 4.5) [111].

#### 4.2.4 Conclusions

The FR method was applied to calculate the PMF of a potassium ion through the gramicidin A channel. The results yielded a central barrier of  $\sim 15 k_B T$  and two wells at the entrance in the channel of  $\sim 6 k_B T$ , and were consistent even for a small number and/or high velocity SMD pulls. The protein flexibility, due to restraints applied to the protein backbone only along the z-direction, has proven to play indeed a major role in the  $K^+$  transport by considerably lowering the  $\sim 40 k_B T$  barrier, obtained when the backbone atoms were fully restrained. The dissipative work inside the channel was found to be linear in  $z$ , yielding a constant diffusion coefficient  $D \sim 10.3 \text{ \AA}^2/\text{ns}$ .

The PMF calculated from the same pulls using Jarzynski's equality with the cumulant approximation yielded, as expected, inconsistent results for forward and reverse pulls on both monomeric  $\alpha$ -helices. However, the average of the forward and reverse PMFs was

very similar to the PMF calculated with the FR method.

The FR method has proven its viability and efficiency by reproducing the potential of mean force given by the umbrella sampling method in considerably less time ( $\sim 3$  days per PMF on 64 CPUs). However, the conduction of the channel cannot be reproduced with any of the PMF profiles, mainly due to the very large central barrier. The main issue in simulating PMFs for ions is the poor treatment of polarization effects by the currently existing force-fields. To account for it, the  $K^+$  ion charge was reduced to half. The calculated PMF exhibited barrier and well sizes very close to the values needed to reproduce the experimental data.

### 4.3 Glycerol Transport through GlpF Channel

The conducting properties of aquaporins (AQPs) are studied experimentally by means of calculating cell membrane permeabilities. The main challenge remains that the precise protein to lipid composition of liposomes [116] or other reconstituted systems [117, 118], i.e. the water channel density in the membrane, is required to find the channel permeability from the knowledge of the membrane permeability.

The water permeability of a channel can be determined by either all-atom free MD simulations or SMD simulations that maintain a hydrostatic pressure difference on the two sides of the membrane [119, 120]. Reproducing the experimental measurements though is still troublesome. For example, the computer models find the permeability of water through AQPZ to be smaller than through GlpF [121], whereas experiments suggest that it is the other way around.

Free energy profiles [5, 82] provide the underlying mechanism of molecular transport through the channel. For large pores, such as the AQPs, these PMFs are computationally too expensive for biased equilibrium methods (e.g. umbrella sampling). Using SMD simulations, on the other hand, is feasible only for relatively narrow pores, because of the

sampling problems. For glycerol transport through the GlpF aquaglyceroporin, there has been only one PMF reported so far [5], and it was computed from unidirectional SMD pulls, by employing Jarzynski's equality (JE). Hence, the few studies of the glycerol transport through GlpF [5, 91, 122] are based on it. The PMF came under scrutiny since it was shown that in practice JE is applicable only sufficiently close to equilibrium (see Sec. 4.2).

Next, the FR method is applied to calculate the PMF profile  $U(z)$  and the corresponding diffusion coefficient  $D(z)$  along the channel axis. The calculated PMF has a dependence on the axial orientation of the glycerol molecule as well.  $U(z)$  and  $D(z)$  are used as input in a stochastic model, in order to calculate the mean first passage time that a glycerol molecule needs to go through the GlpF channel and the permeability of the GlpF to glycerol.

### 4.3.1 Modeling of GlpF Channel

The high resolution (2.2 Å) X-ray crystal structure of monomeric glycerol-bound GlpF (organism *Escherichia coli*) was obtained from the Protein Data Bank (PDB code 1FX8 [103]). Crystallographic glycerol molecules were removed from the initial structure of the GlpF monomer and missing hydrogens were added to the system. Applying consecutive 90° rotations with respect to the z-axis (the channel axis), a tetrameric pore of GlpF channels was generated. A magnesium (Mg) ion was located at the center of the pore and was kept fixed during all the simulations. The obtained protein structure was energy minimized in vacuum for 30,000 steps.

In order to mimic the membrane environment of *E. coli*, the minimized system was embedded into a previously equilibrated patch of fully hydrated POPE lipid bilayer using VMD [96] *Membrane* plugin. Lipid molecules within 0.55 Å of protein were removed. The system was then solvated in water using *Solvate* plugin of VMD. Two 8 Å thick layers of water were added on the sides of the membrane. Water molecules within 0.8 Å of protein and lipid were deleted as well. To neutralize the total charge, 6 chloride ( $\text{Cl}^-$ ) ions were randomly placed in the water using VMD plugin *Autoionize*. The final system, shown in

Fig. 4.6, contained, besides the proteins and ions, 387 lipids and 18,180 water molecules (system total of 118,278 atoms). The approximate unit cell dimensions were  $136 \text{ \AA} \times 129 \text{ \AA} \times 75 \text{ \AA}$ .

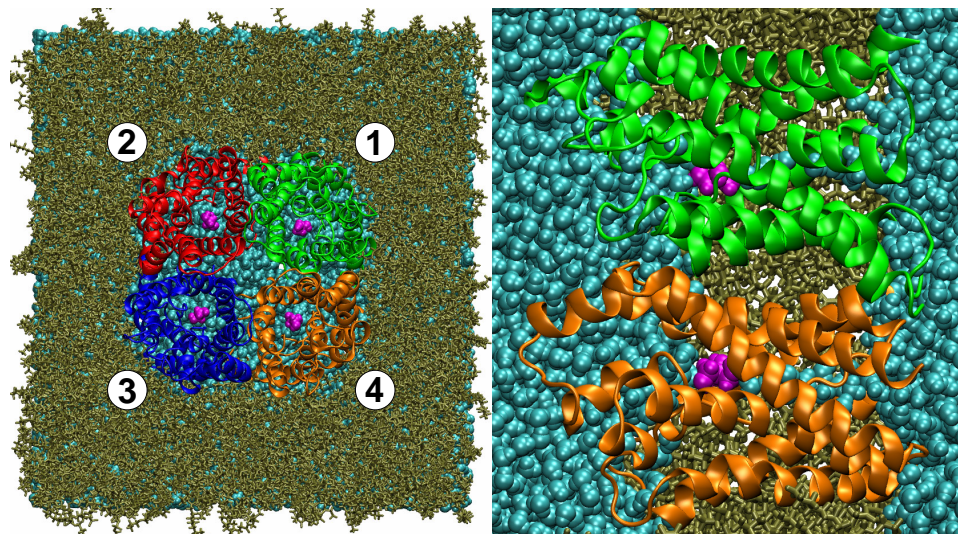


Figure 4.6: Cartoon representation of the simulated GlpF tetramer containing the glycerol molecules: the four GlpF channels in the middle are colored in green, red, blue and orange; the glycerol molecules inside the channels are colored in magenta; the lipids are shown in tan colored licorice representation; the water is represented in light blue spheres. The left panel is a top view of the system, while the panel on the right-hand side represents the side view of two of the channels.

All simulations were performed using the parallel molecular dynamics program NAMD 2.6 [42] with CHARMM27 forcefield for proteins and lipids [37, 123], and TIP3 model for water. A cutoff distance of  $12 \text{ \AA}$  (switching function starting at  $10 \text{ \AA}$ ) was assumed for van der Waals interactions. Periodic boundary conditions were used with flexible cell option enabled. Long-range electrostatic interactions were computed using Particle Mesh Ewald (PME) method. The density of the grid points for PME was around  $1/\text{\AA}^3$ . Unless otherwise mentioned, the following multiple timestepping scheme was used: 1 fs for bonding interactions, 2 fs for non-bonding interactions and 4 fs for electrostatic interactions. NPT ensemble was assumed, with temperature and pressure set for physiological conditions at 310 K and 1 atm., respectively. Langevin dynamics was utilized to ensure constant temperature in all simulations. The Langevin damping coefficient was set to  $5 \text{ ps}^{-1}$ . The

pressure was maintained constant using the Nosé-Hoover Langevin piston method with a decay period of 200 fs and a damping timescale of 100 fs.

For the energy minimization of the system, the following harmonic restrains were applied. The backbones were restrained to their initial positions with a spring constant  $k_{bb}$ . Using a spring constant  $k_P$ , the phosphorus (P) atoms were restrained only along the z-axis (perpendicular on the membrane plane) to the initial mean z-coordinates of all the P atoms within the same lipid layer. Each layer of the membrane contained about 200 phosphorus atoms. 30,000 energy minimization steps were performed using  $k_{bb} = 20$  kcal/mol/Å<sup>2</sup> per atom and  $k_P = 20$  kcal/mol/Å<sup>2</sup> per layer.

Next, using rigid hydrogen bonds, the system equilibration was carried out in three stages, having all the interactions calculated at every 2 fs. First, the entire system was subjected to 0.5 ns gradual heating of 1 K/ps up to 310 K followed by 30,000 energy minimization steps. All restrains were maintained, but the elastic constant of the P atoms was reset to  $k_P = 2000$  kcal/mol/Å<sup>2</sup> per each layer. The lipid environment was relaxed for 0.5 ns by releasing the restrains on the P atoms. Finally, a 5 ns equilibration of the system was accomplished having only the magnesium (Mg) atom fixed.

Finally, four glycerol molecules were added at the periplasmic entrance of each channel (see Fig. 4.6). The conformation and position of a glycerol molecule inside a GlpF channel was extracted from the same crystallographic structure file [103]. The water within 0.8 Å of the glycerols was removed to avoid any overlaps. While harmonically constraining the backbone atoms ( $k_{bb} = 20$  kcal/mol/Å<sup>2</sup> per atom) and each glycerol's heavy atoms center of mass (COM) ( $k_{gly} = 40$  kcal/mol/Å<sup>2</sup> per atom), the obtained system was energy minimized for 10,000 steps and equilibrated for 0.1 ns. In order to orient the glycerol molecules parallel to the z-axis, another 0.5 ns of restrained equilibration was performed. The restrains on the glycerol atoms were removed ( $k_{gly} = 0$ ), but another elastic spring was attached to each glycerol such that the angle  $\psi$ , between the axis of the molecule and the z-axis (see Fig. 4.7a), was maintained below 10°. The restrain on the  $\psi$  angle ( $k_\psi = 150$

kcal/mol/rad<sup>2</sup>) was enforced to assure the proper orientation of the glycerol when entering the channel.

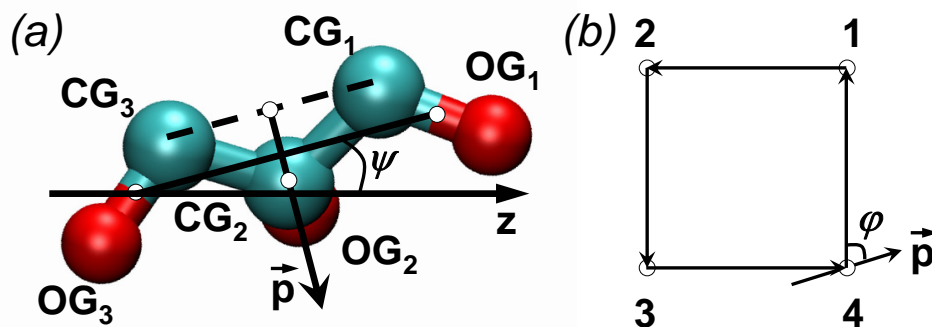


Figure 4.7: (a) Definition of the angle  $\psi$ . Let  $\mathbf{a} = \text{COM}(CG_1 + OG_1) - \text{COM}(CG_3 + OG_3)$  define the orientation of the glycerol molecule (carbons in blue; oxygens in red; hydrogens not shown). Then  $\psi$  is the angle that vector  $\mathbf{a}$  forms with the  $z$ -axis. (b) Definition of the angle  $\phi$ . Let  $\mathbf{p} = \text{COM}(CG_2) - \text{COM}(CG_1 + CG_3)$  define a vector in the plane of the glycerol molecule. Then  $\phi$  is the projection on the membrane plane of the angle that vector  $\mathbf{p}$  forms with the vector passing through the COM of two consecutive channels.

### 4.3.2 SMD Simulations

In order to calculate the PMF profile  $U(z)$  and its corresponding diffusion coefficient  $D(z)$  along the channel axis ( $z$  denotes the reaction coordinate) using the FR method, constant velocity steered molecular dynamics (cv-SMD) simulations were employed to simultaneously pull four glycerol molecules, one through each of the GlpF channels. Forward (F) and reverse (R) works were calculated for each of the pulls [see Eq. (3.11)] and then, using Eqs. (3.24b) and (3.24c), the potential of mean force  $U(z)$  and the mean dissipative work  $\overline{W}_d(z)$  were obtained.

Crooks' TFT allows one to calculate the free energy difference between two equilibrated states. However, the accuracy of the PMF profile in between these two states depends on how far from equilibrium the system is in each of the intermediate states. Therefore, each channel was split into small segments along the RC  $z$  as follows. The narrow part of the channel, i.e.  $z \in (-5, 14)$  Å, was sliced into four unequal segments:  $(-5, -2)$ ,



$(-2, 1)$ ,  $(1, 8)$  and  $(8, 14)$  Å. The periplasmic vestibule domain was divided into three segments:  $(-25, -19)$ ,  $(-19, -10)$  and  $(-10, -5)$  Å and the cytoplasmic one into another three intervals:  $(14, 20)$ ,  $(20, 26)$  and  $(26, 32)$  Å. The system was equilibrated for 0.5 ns in the states that separate these intervals ( $z_0 \in \{-25, -19, -10, -5, -2, 1, 8, 14, 20, 26, 32\}$  Å). Starting configurations, extracted from these equilibration simulations at 100 ps intervals, were used to run five sets of F and R cv-SMD pulls ( $v = 30$  Å/ns) on each of the 10 segments of the channel. The selectivity filter ( $z \in (-5, 1)$  Å) (SF) was sampled better (15 sets of pulls), in order to determine more accurately the barrier height of the channel, located in this region. In the narrow region of the channel the pulls were performed along the  $z$ -axis. Due to the large cross-section of the vestibule regions however, the sampling in the  $xy$ -plane is very poor at high speed pulls, and therefore, the glycerol molecule was pulled on an axis close to the geometric center of the channel. In the case of the segments with  $z \in (-19, -10)$  and  $z \in (14, 20)$  this axis was not parallel to the  $z$ -axis. Combining the PMF profiles computed for each of the 10 segments, the sought  $U(z)$  for the entire channel was obtained. The total mean dissipative work  $\overline{W_d}(z)$ , calculated following the same strategy, yielded the position dependent diffusion coefficient  $D(z)$  [see Eq. (3.25)].

The equilibrations and SMD pulls were performed by applying on the heavy atoms of the glycerol a stiff spring characterized by an elastic constant  $k_z = 100$  kcal/mol/Å<sup>2</sup>. The restrain on the orientation of the glycerol was maintained during all simulations ( $k_\psi = 150$  kcal/mol/rad<sup>2</sup>). The protein flexibility in the membrane plane plays a crucial role in ion transport [2, 3], as shown in Sec. 4.2. Tests indicated that this is true for the glycerol transport through the GlpF channel as well, mainly because the molecule can barely squeeze through the SF region. Therefore, the backbone of the protein was restrained along the  $z$ -axis only with  $k_{bb} = 20$  kcal/mol/Å<sup>2</sup> per atom. To avoid dragging of the channels during the SMD pulls, the COM of each channel was loosely restrained ( $k_{COM} = 500$  kcal/mol/Å<sup>2</sup> over 1071 backbone atoms). All simulations had the Mg atom fixed.

### 4.3.3 Potential of Mean Force and Diffusion Coefficient

As already mentioned, the glycerol molecule is about the size of the cross-section of the narrow part of the GlpF channel. Thus, the interior shape of the channel, especially in the SF region, is locally influenced by the glycerol. The crucial role of the protein flexibility in the membrane plane was proven by PMF calculations (results not shown) that yielded a barrier height in the SF region almost three times larger when the protein backbone atoms were constrained around their equilibrium values, as compared to the case (presented here) when the protein is allowed to fluctuate in the membrane plane.

The PMF of a glycerol molecule passing through a GlpF channel was successfully retrieved using the FR method (see Sec. 3.8), from forward (F) and reverse (R) SMD pulls on the 10 segments, as described in Sec. 4.3.2 (red line in Fig. 4.8a). The PMF exhibits two important features: a small well at the entrance in the narrow region of the GlpF channel (right before the SF) and a large barrier ( $\sim 16 k_B T$ ) at the SF. The free energy profile is compared to previously reported PMF [5] (black line), that was calculated from unidirectional SMD pulls by employing Jarzynski's equality and the cumulant expansion. The sharp features in the latter PMF are not retrieved with the FR method. For example, it seems very unlikely that the molecule would have a binding site in the SF region, where the main barrier is. Also, the obtained periplasmic well is much smaller than its reported correspondent.

Due to sampling issues in the vestibule domains, the PMF was calculated along axes close to the channel axis in those regions. It is unknown though if these axes are along the absolute minimum of the PMF in these regions. An attempt of sampling these regions with a relatively small number of pulls (up to 10-15) would lead to a free energy profile that does not even have the narrow region entrance well and has an overestimated barrier in the SF (results not shown). Because the molecule is pulled so fast towards the narrow region of the channel, it hits the walls of the channel, without "finding its way" into the very

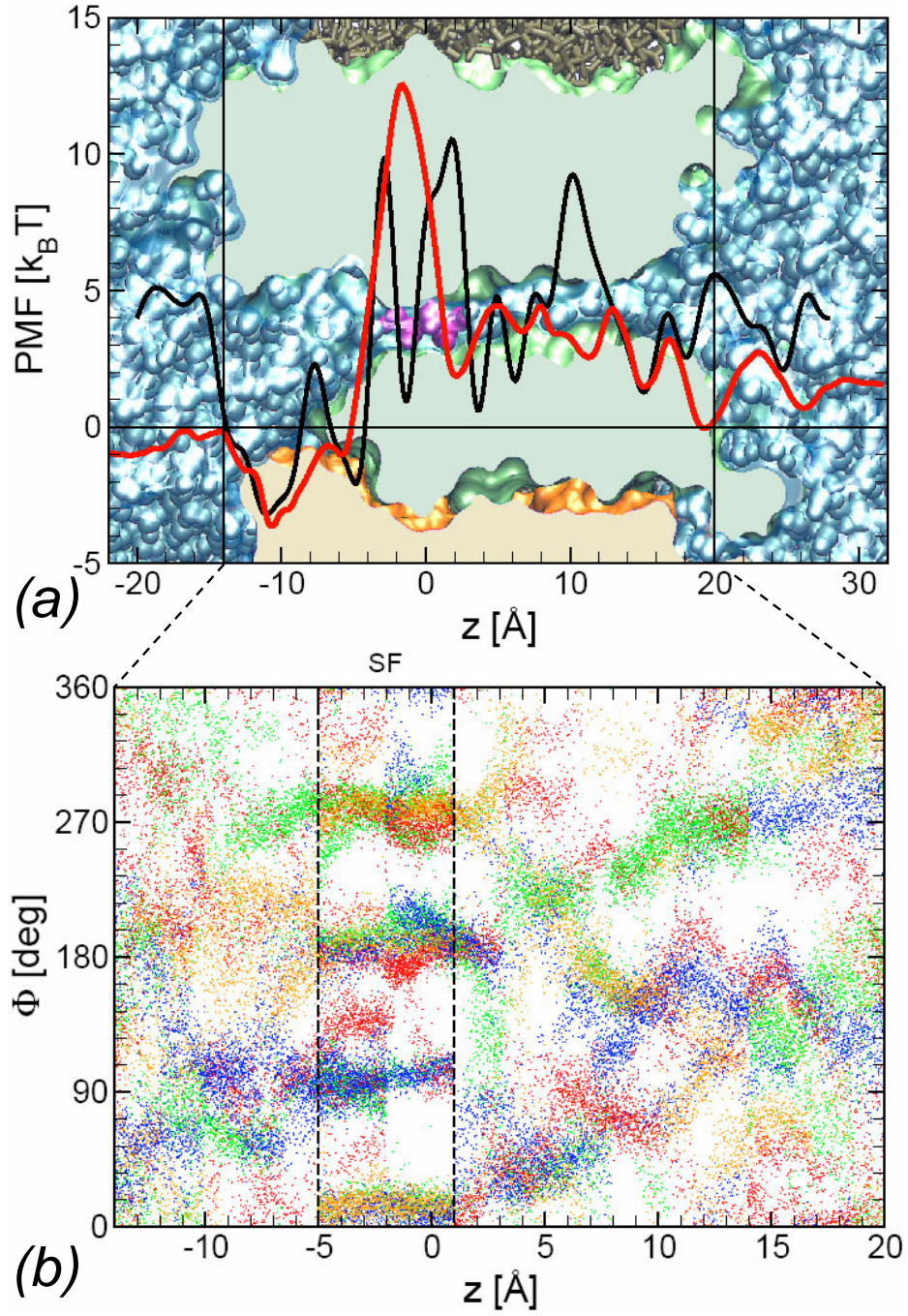


Figure 4.8: (a) Comparison of the PMF calculated using the FR method (red line) with one reported from unidirectional pulls [5] (black line), for a glycerol molecule passing through a GlpF channel. The PMF presents a small well followed by a large barrier ( $\sim 16 k_B T$ ) in the selectivity filter region. (b)  $\phi$  angle variation along the channel during the F and R pulls. In the selectivity filter, the glycerol can pass only under the angles  $\phi \in \{10^\circ, 100^\circ, 190^\circ, 280^\circ\}$ .

narrow SF. The increased friction with the channel walls overestimates the work needed to pull the molecule into the SF region, hence the larger barrier. On the other hand, guiding the molecule along an axis until the entrance in the narrow region eliminates the artificial increase in the work.

Assuming that all the channels are equivalent on large time scales, the sampling can be enhanced by calculating the PMF over all the channels, i.e. considering the F and R works from all the four channels (total of 20 works in each direction) into Eq. (3.24b). Each of the channels is found in a different conformation that does not change over a few nanoseconds (fluctuations in the protein conformation occur on much larger time periods). On this time scale, the lipids form an extremely viscous environment that almost freezes the tetramer's exterior conformation.

The presence of the glycerol molecule increases the radius of the SF [104, 124, 125] as compared to when only water is present. Basically, when the glycerol is passing through the SF region, the GlpF pore molds itself around the molecule. Moreover, the channel “allows” the glycerol to slip through this region under only four axial orientations. These orientations correspond to the following  $\varphi$  angle values:  $\varphi = \{10^\circ, 100^\circ, 190^\circ, 280^\circ\}$  (see Fig. 4.8b). The extensive sampling of the SF (15 pulls in both directions for each channel) confirmed that the restriction on the axial orientation of the glycerol molecule in this region is genuine. Since the axial orientation of the glycerol outside the SF is rather different in each channel, the  $\varphi$  angle is sampled in a satisfactory manner only when all four channels are taken into account.

The calculated PMF suggests that glycerol transport properties are determined by the barrier size in the SF. Hence the increase in the number of pulls in this region from 5 to 15. Due to the better sampling, the barrier height was more accurately determined, and found to decrease from  $19 k_B T$  (for 5 pulls) to  $16 k_B T$  (for 15 pulls) (see Fig. 4.8a). Moreover, the calculated total barrier height matches very well the  $16.1 k_B T$  Arrhenius activation energy of the glycerol in the GlpF channel [101].

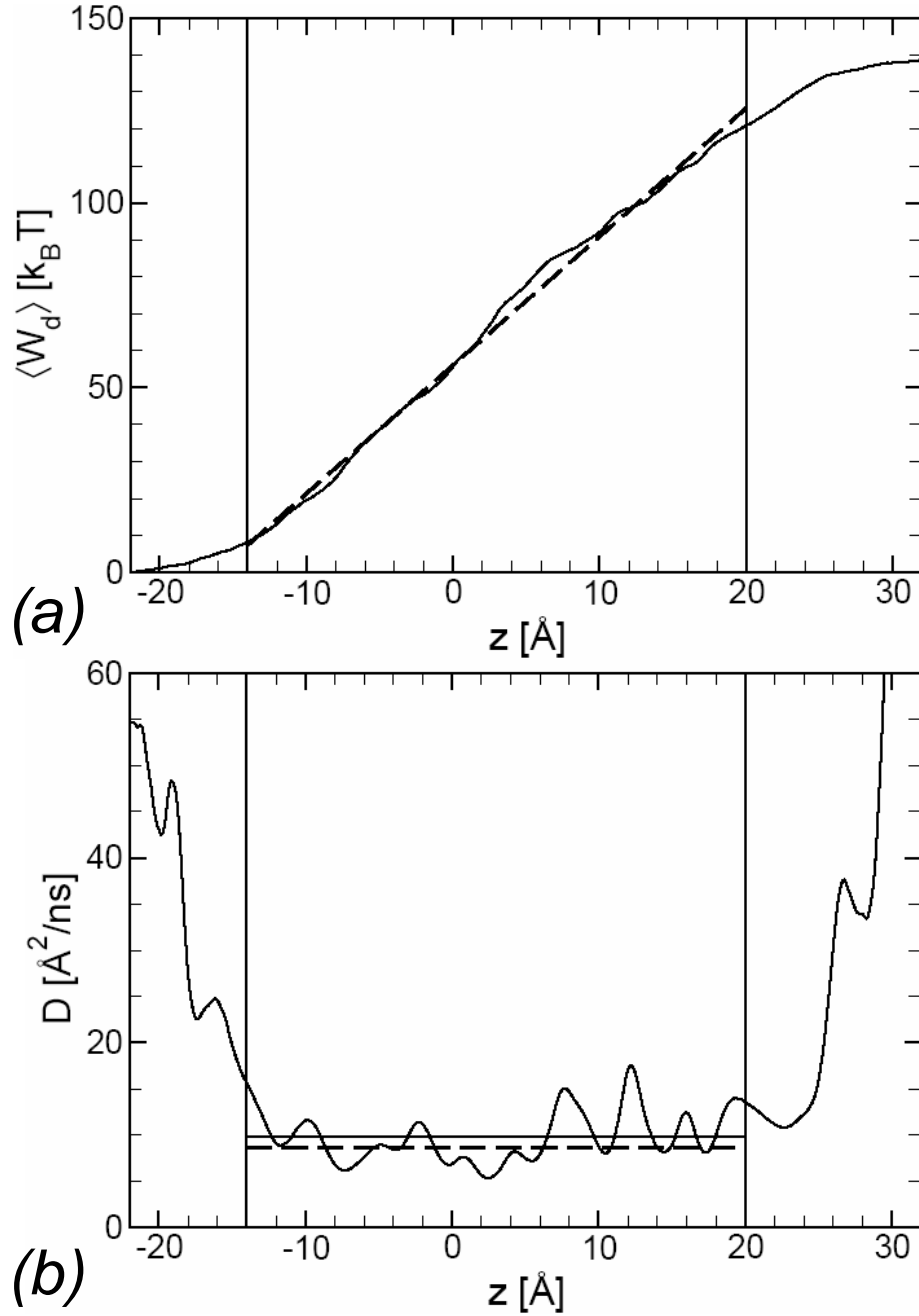


Figure 4.9: (a) The mean dissipative work  $\overline{W_d}$  and (b) its corresponding diffusion coefficient  $D$ . The solid lines correspond to position dependent  $\overline{W_d}(z)$  and its corresponding  $D(z)$  (the horizontal solid line in the bottom panel is the average of  $D(z)$ ). The dashed line reveals a constant diffusion coefficient  $D \sim 8.6 \text{ Å}^2/\text{ns}$ , corresponding to a linear fit of  $\overline{W_d}(z)$  in the narrow region of the channel. The diffusion coefficient of glycerol in bulk (for  $z < -30 \text{ Å}$ ) was found to be  $D_{bulk} \gtrsim 95 \text{ Å}^2/\text{ns}$ .

The mean dissipative work (see Fig. 4.9a), obtained from Eq. 3.24c), is proportional to the friction coefficient. Therefore, the friction of the glycerol inside the channel is, as expected, larger than in the vestibule and bulk area.  $\overline{W_d}$  in the narrow region linearly increases with  $z$ , yielding a constant friction coefficient. The diffusion coefficient (Fig. 4.9b), calculated from Eq. (3.25), is found to be  $D \sim 8.6 \text{ \AA}^2/\text{ns}$ . Using the same approach in the bulk (outside of the plotted region), one gets  $D_{bulk} \gtrsim 95 \text{ \AA}^2/\text{ns}$ . The latter value matches rather well with the experimentally measured bulk value for glycerol in water  $D_{bulk}^{exp} \sim 110 \text{ \AA}^2/\text{ns}$  [126].

#### 4.3.4 Mean First Passage Time

Having both the potential of mean force and its corresponding diffusion coefficient, one can easily calculate the mean first passage time that a molecule needs to pass through the channel [5]. Due to the attractive periplasmic well, it is assumed that when the glycerol molecule exits towards the periplasmic side ( $z_p$ ) another molecule takes immediately its place (reflective periplasmic boundary). If it exits to the cytoplasm ( $z_c$ ), the glycerol is quickly phosphorylated (absorbent cytoplasmic boundary). With these assumptions, the mean first passage time can be calculated using the expression [127]:

$$\tau(z_p, z_c) = \int_{z_p}^{z_c} \frac{dz}{D(z)} \int_{z_p}^z dz' e^{U(z)-U(z')} \quad (4.1)$$

In the regime of linear dependence of the mean square displacement with time, the effective diffusion coefficient of the molecule passing through the channel of length  $L = z_c - z_p$  can be calculated from the expression:

$$D_{eff} = \frac{L^2}{2\tau} \quad (4.2)$$

where  $\tau \equiv \tau(z_p, z_c)$  is the mean first passage time from eq. (4.1).

For the glycerol to pass from one vestibule into the other, i.e. from  $z = -14 \text{ \AA}$  to  $z = +20 \text{ \AA}$ , the glycerol needs to pass a  $16 k_B T$  barrier at the SF region. The mean first passage time calculated for this interval yielded a value of  $\tau_{gly}(-14, +20) = 3.32 \text{ ms}$ . Equation (4.2) yielded an effective diffusion coefficient  $D_{eff} \sim 0.17 \text{ \AA}^2/\mu\text{s}$ .

### 4.3.5 Channel Permeability

The ability of a channel to conduct water, in the absence of a hydrostatic pressure difference across the membrane, is given by the ratio of the net flux to the osmotic pressure, or, equivalently, by the ratio of the molar flux and the solute concentration gradient  $\Delta c_s$  that induces the osmotic pressure:

$$\Phi_w = -p_f \Delta n_s \quad (4.3)$$

where  $\Delta n_s = N_A \Delta c_s$  is the concentration difference expressed in water molecules per unit volume and  $p_f$  is the *osmotic permeability coefficient* of a channel. Similarly, the channel permeability to a solute can be defined as the ratio of the solute flux to the concentration difference across the membrane:

$$\Phi_s = p_s \Delta n_s \quad (4.4)$$

Note that the flux of the solute is in the direction of the concentration gradient, whereas the water is opposed to it.

If the membrane is impermeable to both water and solute, but is porous, the osmotic permeability per pore is related to the permeability ( $P_f$  or  $P_s$ ) of the membrane, the membrane area  $A$  and the number of pores  $n$  contained in it, by the expression:

$$p_{f,s} = \frac{P_{f,s} A}{n} \quad (4.5)$$

In the steady state transport of a molecule across a channel, characterized by a PMF  $U(z)$  and a corresponding diffusion coefficient  $D(z)$ , the flux is constant and in the absence

of an external force, is given by [122]

$$\begin{aligned}\Phi &= A_p \rho_p - A_c \rho_c, \\ A_i &= D_i / (z_c - z_p) \cdot e^{U(z_i)} \left[ \int_p^c e^{U(z)} dz \right]^{-1}, \\ \rho_i &= S_i n_i, \quad i = \{p, c\}\end{aligned}\tag{4.6}$$

where  $S_i = S(z_i)$  and  $n_i = n(z_i)$  are the area of the cross-section of the channel and the concentration in molecules per unit volume on the periplasmic ( $i = p$ ) and cytoplasmic ( $i = c$ ) sides. In the case when there is no hydrostatic pressure difference across the membrane  $U(z_p) = U(z_c) = U_0$  and therefore  $A_p = A_c = A_0$ . Assuming the same area of the cross-sections at the two ends of the channel  $S_p = S_c = S_0$  and a constant diffusion coefficient  $D$ , one can determine the permeability of a channel from the sole knowledge of  $U$  and  $D$ :

$$p = A_0 S_0 \tag{4.7}$$

where  $A_0$  depends on  $U$  and  $D$  according to Eq. (4.6).

More specifically, the potential of mean force  $U(z)$  and its corresponding diffusion coefficient  $D(z)$  enter in the Eq. (4.7) as the flux of the glycerol through the channel. Considering the cross-section area  $S_0 = 100 \text{ \AA}^2$  [103] and the diffusion coefficient  $D = 8.6 \text{ \AA}^2/\text{ns}$ , the GlpF permeability to glycerol becomes  $p_{gly} \sim 0.78 \times 10^{-19} \text{ cm}^3/\text{s}$ . The large underestimation of  $p_{gly}$  comes from the flux term, which is proportional to the net barrier height, i.e. the barrier height with respect to the energy level in the bulk (vestibules). In order for the permeability coefficient to have the same order of magnitude with the experimental value ( $10^{-16} \text{ cm}^3/\text{s}$ ) [101], the net barrier should be about 3 times smaller than the calculated  $12.5 k_B T$  value. One explanation might be that the periplasmic well size is underestimated due to the lack of sampling in the vestibule region. The PMF in this region was calculated along an axis guiding the glycerol into the channel. The  $\psi$  angle was also restrained so that flips of the molecule were not allowed. Since this region is wider than the size of the



glycerol, these flippings could occur, meaning that a glycerol molecule could require more energy to leave the vestibule, i.e. the periplasmic well is larger.

### 4.3.6 Conclusions

The PMF of the glycerol through the GlpF channel and its corresponding diffusion coefficient was determined using the FR method. The PMF exhibited two main features: a small periplasmic well and a  $16 k_B T$  barrier located at the narrowest region of the channel, the selectivity filter. The SF region radius was found to be larger in the presence of glycerol, as compared to when the channel is filled with water, suggesting that the protein molds around the glycerol molecule. Hence, its flexibility has a crucial role in the glycerol transport through the GlpF channel as well. Moreover, results show that the PMF is also dependent on the axial orientation of the glycerol. Thus, the molecule can pass the SF only under certain orientations:  $\varphi = \{10^\circ, 100^\circ, 190^\circ, 280^\circ\}$ . Although the NPA motif is very important in water transport (the water dipole flipping occurs here), it seems to have no major role in the glycerol transport.

The computed free energy profile is qualitatively different from the one reported from unidirectional SMD pulls, using Jarzynski's equality and cumulant approximation [5]. The latter exhibits features that were not reproduced (e.g. a binding site in the SF region).

The diffusion coefficient was found to be almost constant inside the narrow region of the channel:  $D_{ch} = 8.6 \text{ \AA}^2/\text{ns}$ . The bulk value is about 10 times larger:  $D_{bulk} \sim 95 \text{ \AA}^2/\text{ns}$ , and matches rather well the experimental value  $D_{bulk}^{exp} \sim 110 \text{ \AA}^2/\text{ns}$  [126].

Having the PMF and  $D$ , we calculated (i) the mean first passage time that a glycerol molecule needs to pass through the GlpF channel  $\tau_{gly} \sim 3.3 \text{ ms}$ , and (ii) the channel permeability to glycerol  $p_{gly} \sim 0.8 \times 10^{-19} \text{ cm}^3/\text{s}$ . The latter is three orders of magnitude smaller than the experimental findings [101]. Since  $p_{gly}$  is proportional to the net barrier height (barrier height with respect to bulk free energy level), one explanation could be that the periplasmic well size is underestimated due to the poor sampling of the vestibule region.

# **5 In silico prediction of the oligomerization state of light harvesting complexes**

## 5.1 Introduction

In photosynthesis, assemblies of pigment-protein complexes absorb sunlight and convert its energy into a biochemical potential. In recent years, tremendous progress has been made towards identifying the *in vivo* structure of the photosynthetic apparatus, in particular that of purple bacteria [14, 130–132]. Calculations and spectroscopic measurements reveal that a close relationship exists between the efficiency of light harvesting and the geometrical arrangements of pigment-protein complexes [16, 133–135].

This raises the question as to how does the photosynthetic apparatus assemble within a cell membrane. The antenna light-harvesting complex 2 (LH2) of purple bacteria can be considered a paradigmatic model system to address this question, because (i) atomic-resolution crystal structures exist for LH2s with different organizations and (ii) mutagenesis and reconstitution assays allow for direct experimental studies of key factors in the assembly of LH2 complexes.

LH2s are composed of repetitions of a basic unit of two transmembrane polypeptides,  $\alpha$  and  $\beta$ . Each  $\alpha\beta$  heterodimeric subunit non-covalently binds three bacteriochlorophylls (BChls) and one carotenoid. Figure 5.1 shows the crystal structure of the basic subunit of LH2 from *Rhodospirillum (Rs.) molischianum* [12], *Rhodopseudomonas (Rps.) acidophila* [13] and from a species with unpublished crystallographic structure, hence-forth referred to as MOLI, ACI and X, respectively<sup>†</sup>. The corresponding primary structures are shown in Fig. 5.2. Both  $\alpha$  and  $\beta$  polypeptides consist of polar N- and C-termini and a central hydrophobic region. The N-termini lie on the cytoplasmic side of the membrane, the C-termini on the periplasmic side. Amino acids in the central hydrophobic region form two transmembrane  $\alpha$ -helices. B850 BChls are ligated to the highly conserved His residues near the C-terminus.

---

This chapter is based on published article [128] and manuscript [129].

<sup>†</sup>The PDB structure file of the unknown X heterodimer was provided by Dr. Richard Cogdell (University of Glasgow).

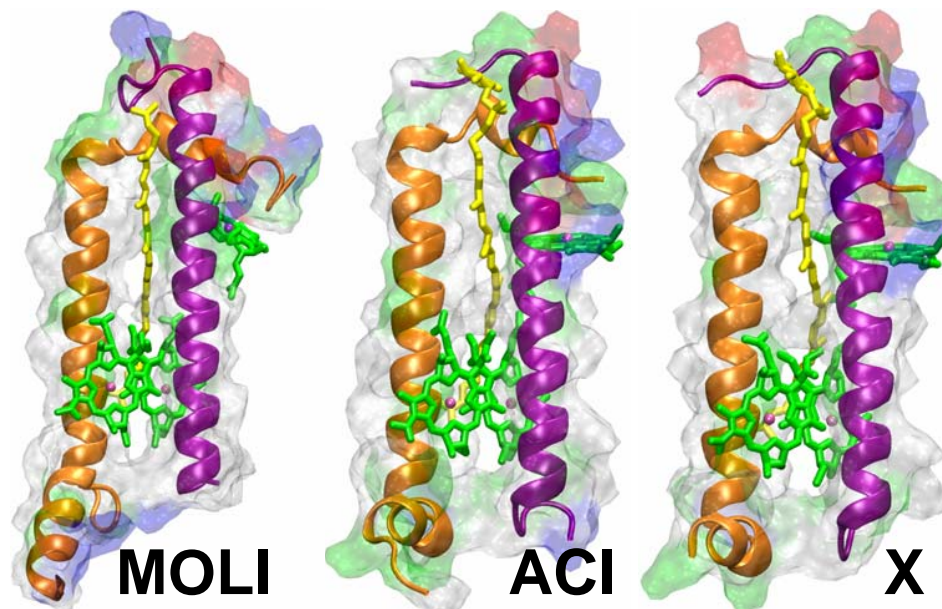


Figure 5.1: Structure of a subunit of LH2 from *Rs. molischianum* (MOLI), from *Rps. acidophila* (ACI) and from an unknown species (X). Each subunit consists of one  $\alpha$  polypeptide (orange ribbon), one  $\beta$ -polypeptide (magenta ribbon), three BChls (green; phytyl chains truncated), and one carotenoid (yellow). The N-terminal domains are on top, the C-terminal domains on bottom. The surface of the subunit is superimposed onto the simplified representations.

<b>MOLI-<math>\alpha</math></b>	SNPKDDYKIWLVINPSTWLPVIWIVATVVVAIAVHAAVLAAPGFNWIALGAAKSAAK
<b>ACI-<math>\alpha</math></b>	MNQGKIWTVVNPAIGIPALLPSVTVIAILVHLAILSHTTWFPAWQGGVKK
<b>X-<math>\alpha</math></b>	?NQGKIWTVVPFAFGLPLMLGAVAITALLVHAAVLTHTTWYAAFLQ
<b>MOLI-<math>\beta</math></b>	AERSLSGLTEEEAIAVHDQFKTTFSAFIILA AVAHVLVWVWKPWF
<b>ACI-<math>\beta</math></b>	ATLTAEQSEELHKYVIDGTRVFLGLALVAHFLAFSATPWLH
<b>X-<math>\beta</math></b>	AEVLTSEQAEELHKHVIDGTRVFLVIAAIAHFLAFTLTPW

Figure 5.2: The primary structure of the light-harvesting polypeptides of LH2 from MOLI, ACI and X, aligned at the conserved His residue. Transmembrane helical domains are highlighted in yellow.

Interestingly, the crystallographic structures of LH2 reveal a different organization of subunits, a ring of nine  $\alpha\beta$ -subunits for ACI [13, 136], but of eight  $\alpha\beta$ -subunits for MOLI [12]. EM and AFM studies reveal nonameric organizations for LH2s from *Rhodovulum sulfidophilum* [137], *Rhodobacter (Rb.) sphaeroides* [138, 139], and *Rubrivivax gelatinosus* [140, 141], whereas a low-light variant form of LH2 from *Rps. palustris* reveals an eight-fold symmetry [142]. In all of the above cases, the octameric or nonameric organization appears to be homogeneous within a given species. Structural heterogeneity of LH2s within a species has been described only for *Rs. photometricum*, where AFM studies suggest that most LH2s are organized in either eight-, nine-, or ten-fold symmetry [143].

It needs to be pointed out that at present very little is known about the assembly of light harvesting complexes *in vivo*. Although a very recent study suggests that formation of RC-LH1 complexes is temporally separated from formation of LH2 complexes [144], one cannot exclude the possibility that both LH1 and LH2  $\alpha\beta$ -subunit are present at the same time during the assembly process, raising the possibility of variable complex sizes because of mixture of different apoprotein types.

Reconstitution assays [145] show that in many cases light harvesting complexes with very similar optical properties to the wild-type complexes can be reconstituted *in vitro* from their individual components [146, 147]. Truncated versions of natural proteins, chemically synthesized *de novo* proteins, and mutagenetic gene products have been studied, revealing residues essential to formation of  $\alpha\beta$ -subunits and full complexes [148–152]. A recent study demonstrated *in vivo* assembly of redesigned peptides from *Rb. sphaeroides* into fully functional light-harvesting complexes [153].

These results suggest that the building blocks of light harvesting complexes can self-assemble to form stable, functional complexes. Thus, one should be able to relate the observed differences in complex organization to the structure of their subunits. What features of the subunits govern the organization of the complete ring complexes, in particular the oligomerization state, i.e., the number of subunits employed in forming a ring?

This chapter investigates in how far the variations in oligomerization states can be explained by changes in the local interaction angle between neighboring subunits. In theory, one expects that subunits with a preferred interaction angle of  $45^\circ$  would assemble into a ring of eight subunits ( $8 \times 45^\circ = 360^\circ$ ), whereas subunits with a preferred interaction angle of  $40^\circ$  should form rings of nine subunits. It appears a remarkable feat for subunits of two helical proteins to control a difference in angle as small as  $5^\circ$  in the presence of conformational fluctuations and solvent dynamics.

As a first step in understanding how LH2 subunits assemble into precise ring structures, the focus is set on determining and characterizing the interaction between two LH2 subunits. Atomic models for two LH2 subunit dimers, namely for MOLI and ACI, can be readily built starting from their high resolution crystal structures available from the Protein Data Bank (entry codes 1LGH [12] and 1KZU [13, 154], respectively). In order to mimic their native environment, a pair of subunits for both MOLI and ACI was embedded into properly solvated lipid bilayers. Using all atom MD simulations, the dynamical behavior of the systems was investigated by monitoring their relative spatial separation and interaction angle. Next, the relationship between these two characteristics of the dimers was studied using forced detachment of the two LH2 subunits. The results yielded two representative separations for which the potential of mean force along the interaction angle was calculated.

The second step is to try to predict the oligomerization states into which the LH2 rings assemble from the sole knowledge of the preferred angle of two LH2 subunits. Three atomic models for each LH2 dimer, namely for MOLI, ACI and X were built from their crystallographic structures. The initial angles between the monomers was set to  $36^\circ$ ,  $40^\circ$  and  $45^\circ$  for each species. Finally, all the nine systems were embedded in their native lipid environments. Free energy profiles along the dimer interaction angles were determined through all atom steered molecular dynamics simulations.

## 5.2 Interaction Between Two LH2 Subunits

Since, presently, all atom MD simulations studies are restricted to the 10-100 nanosecond time scale, they cannot be applied directly to describe the complete assembly process between two LH2 subunits. Indeed, lateral diffusion of these protein units in the lipid membrane occurs on a much longer time scale than the one accessible by MD simulations. However, suitably designed MD simulations, combined with statistical mechanical analysis, can be used to determine the free energy profile (or PMF) [57] between the interacting subunits. The PMF then can be used as input in a suitable stochastic model (e.g., Fokker-Planck or Smoluchowski equation [59]) for describing the dynamics of the system at a mesoscopic or even macroscopic scale. To this end, one first needs to identify a small number of variables (e.g., distances and/or angles), referred to as *reaction coordinates* (see Sec. 3.2), that describe the relative separation and orientation of the subunits. Then, one should use a properly designed set of MD simulations to calculate the PMF  $U(z)$  of the system as a function of the reaction coordinates (RC)  $z$ . Once the PMF is determined, the generalized force exerted between the subunits is equal to  $F = -\partial U(z)/\partial z$ .

### 5.2.1 Reaction Coordinates $R$ and $\theta$ and PMF $U(\theta, R)$

The self-assembly of LH2 rings can be envisioned as a process in which preformed LH2 subunits are first inserted into the membrane, then brought within contact distance through diffusion in the lipid membrane, followed by final locking into the ring specific geometry. In a first approximation, this process can be modeled as a purely two dimensional one in which each subunit has a specific anisotropic 2D structure (e.g., a deformed disk) that at the end of the process forms an  $N$ -fold symmetric ring (e.g.  $N = 8$  for MOLI and  $N = 9$  for ACI).

Two reaction coordinates are defined to characterize the spatial interaction between LH2 subunits in their native membrane environment, namely  $R$  - the separation between

them, and  $\theta$  - their relative angular orientation. A more precise definition of reaction coordinates is illustrated in Fig. 5.3, displaying two MOLI subunits (from the equilibrated system) in a top view (left) from the N-terminal or cytoplasmic side. On the right-hand side of Fig. 5.3 is shown only the transmembrane helical domains of the two subunits ( $A_1$ ,  $B_1$ ) and ( $A_2$ ,  $B_2$ ), showing a clear separation of  $\alpha$  and  $\beta$  polypeptides.  $A_i$  ( $B_i$ ) denotes the center of mass (COM) of the  $\alpha$  ( $\beta$ ) apoprotein in subunit  $i = 1, 2$ . By definition,  $\theta$  is the angle made by the projections of the vectors  $\mathbf{v}_i = \overrightarrow{A_i B_i}$  on the  $xy$ -plane of the membrane. The separation distance reaction coordinate is defined as the distance between the COMs of the  $\alpha\beta$ -apoprotein heterodimer within the  $xy$ -plane, i.e.,  $R = |\mathbf{R}_1 - \mathbf{R}_2|$ , where  $\mathbf{R}_i$ ,  $i = 1, 2$ , are the projections of the position vectors of these COMs on the  $xy$ -plane. Note that, at any instant of time, the reaction coordinates are determined (through the COMs) by the coordinates and masses of a selected group of atoms from both subunits.

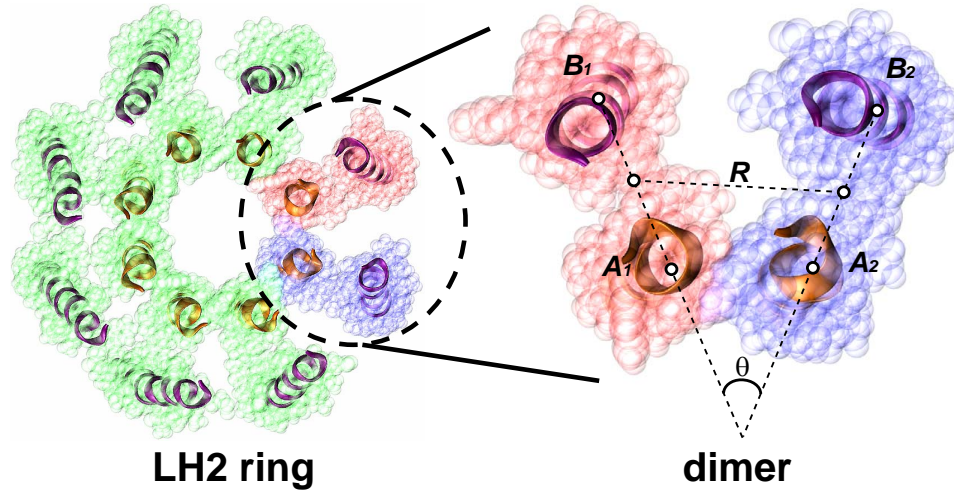


Figure 5.3: Definition of the reaction coordinates  $\theta$  and  $R$  illustrated for two subunits of LH2 from MOLI. (left) Top view from the N-terminal or cytoplasmic side in space-filling representation with transmembrane helices highlighted in ribbon representation. (right) Top view of two neighboring LH2 subunits'  $\alpha$  and  $\beta$  polypeptides with definition of the reaction coordinates.

In selecting the group of atoms that define the reaction coordinates one has to make sure that under equilibrium (or stationary) conditions, the reaction coordinates have well defined mean values and sufficiently small fluctuations. Otherwise, the reaction coordinates are ill



defined and cannot be used to characterize the system quantitatively. For example,  $R$  and  $\theta$  are well defined if one considers only the heavy atoms in the transmembrane helical (TMH) domains of the  $\alpha\beta$ -apoproteins. Adding to the selection all the heavy atoms in the N- and C-terminal domains, would lead to substantial increase in the fluctuations of  $\theta$  rendering the latter reaction coordinate meaningless. In the following, the convention used is that  $R$  and  $\theta$  refer to particular target values of the reaction coordinates, while  $\tilde{R} \equiv \tilde{R}(\mathbf{q})$  and  $\tilde{\theta} \equiv \tilde{\theta}(\mathbf{q})$  refer to the instantaneous values of the reaction coordinates determined from the corresponding positions  $\mathbf{q} \equiv \mathbf{q}_j(t)$  of all defining atoms.

The PMF  $U(z) \equiv U(\theta, R)$  is the free energy of the system formed by the two interacting subunits for well defined spatial  $R$  and angular  $\theta$  separations; the compact notation  $z \equiv \{\theta, R\}$  was introduced to match the notation used in Chp. 3. In principle, the PMF can be calculated from Eq. (3.2), by integrating out all degrees of freedom except for the reaction coordinates [57] [see Eq. (3.1)]. However, in practice, even the longest equilibrium MD trajectories will sample only a restricted region of the reaction coordinate domain (i.e., within the vicinity of the PMF minimum) of interest and the direct application of Eq. (3.1) is impractical.

First, the umbrella sampling [61] method, described in Sec. 3.5, is used for calculating the  $\theta$  dependent PMFs for a constrained separation distance  $R$  between the two LH2 subunits (see Sec. 5.2.5). Next, the “FR method” [52], described in Sec 3.8, is used to determine the PMFs  $U(\theta)$  along the angle reaction coordinate when  $R$  is integrated out (i.e. varies freely) (see Sec. 5.3.4).

## 5.2.2 System Modeling and MD Simulations

*ACI dimer* – two adjacent complete LH2 subunits (protein and cofactors) were extracted from the PDB structure 1KZU [13, 154]. After adding the missing hydrogens, the protein complex was inserted in a pre-equilibrated and hydrated POPC lipid bilayer. Finally, the system was solvated by adding extra water layers to the two sides of the lipid bilayer.

Besides the proteins and cofactors the system contained 4,014 water molecules and 169 POPC lipid molecules, with a total system size of 38,594 atoms. The +4.0 e charge of the system was neutralized by properly adding 4  $\text{Cl}^-$  counter ions. The system was built by using XPLOR [155] and VMD [96].

*MOLI dimer* – using VMD and its plugins, two complete neighboring units extracted from a fully equilibrated 8-fold LH2 MOLI ring, reported in [22] were inserted in a pre-equilibrated and hydrated POPE lipid bilayer, and then solvated by adding two pre-equilibrated 8 Å thick water layers to each side of the membrane. The +4.0 e charge of the system was neutralized by properly adding 4  $\text{Cl}^-$  counter ions. In addition to the protein and cofactors, the final system contained 8,299 water molecules and 128 POPE lipids, with a total system size of 44,997 atoms.

*Equilibrium MD simulations* – After proper energy minimization, both ACI and MOLI were equilibrated at 310 K and normal atmospheric pressure through 5 ns long MD simulations in the NPT ensemble. Periodic boundary conditions and full electrostatics via the particle mesh Ewald method were used. All MD simulations were carried out with the NAMD [156] program using the CHARMM 27 parameter set [37]. The simulations were carried out on local Linux Beowulf clusters, and the required time for 1 ns simulation running on 30 CPUs was about 1.5 days.

*SMD simulations* – A total of 20 SMD simulations were carried out for each system, starting from the same state that coincided with the last frame of equilibrium MD run. The details of the applied harmonic guiding potential for each of the runs are described in Sec. 5.2.4.

*Umbrella sampling simulations* – To determine the PMF  $U(\theta|R)$  for  $R = R_0 = 18$  Å and  $R = R_x = 25$  Å, umbrella sampling MD runs were carried out in a number of windows, starting from  $\theta = 33^\circ$  to  $\theta = 53^\circ$ , that resulted in well overlapping histograms. For ACI (MOLI) at both  $R_0$  and  $R_x$  the windows were centered on the following  $\theta_i$  angles, measured in degrees: 33, 35, 37, 38, 39, 40, 41, 42, 43, 45, 47, 49, 51 and 53 (35, 36, 37, 38, 39,

40, 41, 42, 43, 45, 47, 49, 51 and 53). The target angle was enforced by applying a 2D harmonic guiding potential  $V_{i,j}(\theta, R) = k_i(\theta - \theta_i)^2/2 + k_R(R - R_j)^2/2$  with  $j = \{0, x\}$ ,  $k_R = 80$  kcal/mol/Å<sup>2</sup>, and  $k_i$  tuned between  $(8 - 10) \times 10^3$  kcal/mol/rad<sup>2</sup> for achieving optimal sampling in each window. After exhaustive testings, simulations were performed for 0.7 ns for each window. Only the last 0.5 ns parts of the trajectories were used to construct the RC distribution histograms. The PMF  $U(\theta|R)$  was determined from WHAM [see Eqs. (3.10)].

The interaction between the two subunits of each system was determined and characterized in three steps: (1) In order to test the reliability and usefulness of the reaction coordinates, the time evolution of  $R$  and  $\theta$  was monitored during 5 ns long equilibrium MD simulations of the dimer system; (2) Study the relationship between the spatial separation  $R$  and relative orientation  $\theta$  of the two LH2 subunits as a result of their forced detachment; (3) Investigate the interaction angle between the two subunits of the dimer.

### 5.2.3 Equilibrium MD Simulations of MOLI and ACI Dimers

Since the reaction coordinates  $R$  (the relative separation) and  $\theta$  (the relative orientation) are defined through the COMs of the  $\alpha$ - and  $\beta$ -polypeptides for each subunit (see Fig. 5.3), their values depend on the selection of atoms used to determine the COMs. Extensive testing yielded that the most stable reaction coordinates correspond to the situation in which only the heavy atoms of the trans-membrane helical (TMH) domain of the  $\alpha\beta$ -apoproteins are considered. In this case, the reaction coordinates assume well defined mean equilibrium values. The distribution histograms of the values of the reaction coordinates for the last 2 ns of the MD trajectories are shown in Fig. 5.4 (thick lines). The peak positions in the histograms in Fig. 5.4 (corresponding to the most probable values of the reaction coordinates) are  $R_0 \approx 18.2 \pm 0.14$  Å,  $\theta_0 \approx 44.8^\circ \pm 1.7^\circ$  for MOLI, and  $R_0 \approx 17.8 \pm 0.15$  Å,  $\theta_0 \approx 38.1^\circ \pm 1.4^\circ$  for ACI. For comparison, Fig. 5.4 also includes the corresponding histograms for complete MOLI and ACI LH2 rings (thin lines), obtained from the last 3 ns part

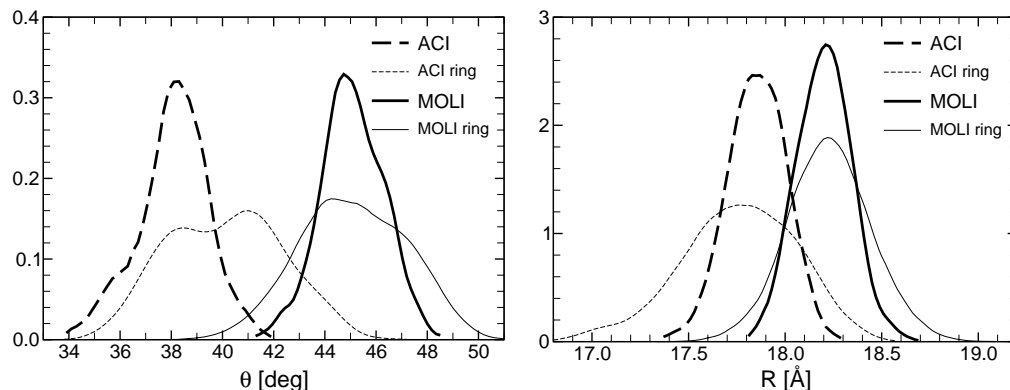


Figure 5.4: Histograms of the reaction coordinates  $\theta$  (left) and  $R$  (right) corresponding to equilibrium MD simulations as follows: MOLI dimer, 2 ns run (thick solid line), ACI dimer, 2 ns run (thick dashed line), MOLI LH2 ring, 3 ns (thin solid line) and ACI LH2 ring, 3 ns (thin dashed line).

of 5 ns long equilibrium MD runs. The peak positions of the spatial RC for the rings are  $R_0 \approx 18.2 \pm 0.2 \text{ Å}$  for MOLI and  $R_0 \approx 17.8 \pm 0.4 \text{ Å}$  for ACI. The mean angle RC values are  $\theta_0 = 45^\circ \pm 2.36^\circ$  and  $\theta_0 = 40^\circ \pm 2.36^\circ$  for MOLI and ACI LH2 ring, respectively. According to these results, there is a clear separation ( $\sim 7^\circ$ ) between the values of the preferred angle  $\theta_0$  for MOLI and ACI, respectively. Indeed, in the case of MOLI  $\theta_0$  is very close to the theoretically expected  $45^\circ$ , while for ACI the corresponding angle is about  $2^\circ$  smaller than the expected  $40^\circ$ . Moreover, for both systems the root-mean-square fluctuations of  $\theta_0$  are of the same magnitude (note the similar width of the corresponding angle distribution histograms in Fig. 5.4), being smaller than  $2^\circ$ . On the other hand, for both MOLI and ACI LH2 rings  $\theta_0$  coincides with the expected  $45^\circ$  and  $40^\circ$ , respectively, albeit the fluctuations in the angle are noticeably larger than for the subunit dimers. Furthermore, fluctuations in  $R$  are much smaller than fluctuations in  $\theta$ , and the mean value  $R_0$  is essentially the same (within less than half of one Å) for all four systems. If one tries to extend the protein atoms selection in the definition of the reaction coordinates, e.g., by including the CT (C-terminus) and NT (N-terminus) domains of the  $\alpha\beta$ -polypeptides, the resulting reaction coordinates become ill defined especially due to the sharp increase in the magnitude of the fluctuations. In such cases, test MD simulations have shown that the fluctuations both in  $\theta$

and  $R$  increase by more than a factor of 2 (results not presented here). These findings are consistent with the knowledge that in general the TMH regions of membrane proteins are more rigid than the outer membrane parts.

#### 5.2.4 Forced Detachment of the Subunits Using SMD

In principle, the study of the relationship between the spatial separation and relative orientation of the two LH2 subunits requires the knowledge of the 2D potential of mean force  $U(\theta, R)$ , that describes the statistical mechanical state of the system as a function of the two reaction coordinates. However, a brute force determination of the PMF (e.g., by direct application of the umbrella sampling method, described in Sec. 3.5) is computationally prohibitively expensive.

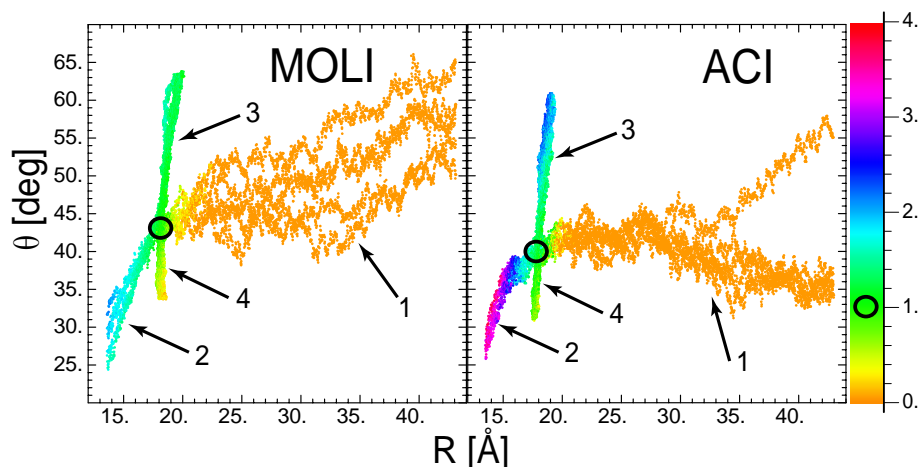


Figure 5.5: Two dimensional density plot of the volume overlap of the transmembrane protein regions of MOL (left) and ACI (right) dimers along the reaction coordinates  $\theta$  and  $R$  recorded in the SMD simulations described in the text; the trajectories corresponding to the four different sets of simulations are indicated by numbered arrows. The volumes are relative to the value corresponding to the equilibrium reaction coordinates  $\theta_0$  and  $R_0$  (see text), marked by the small circle.

Thus, to gain some insight into the mechanism that governs the relationship between the relative distance and orientation of the LH2 subunit dimer during the forced separation or compression of the subunits, four sets of 5 SMD simulations each were conducted, starting

from initial states that coincide with the last frames of the equilibrium simulations shown in Fig. 5.4, and characterized by  $R_0 = 18.2 \text{ \AA}$  and  $\theta_0 = 43.7^\circ$  for MOLI and  $R_0 = 18.1 \text{ \AA}$  and  $\theta_0 = 41^\circ$  for ACI. These starting values are marked by small circles in Fig. 5.5. In the first (second) set of simulations  $\theta$  was unconstrained while  $R$  was increased (decreased) with a constant rate of  $v_R = 0.1 \text{ \AA/ps}$ , by applying a harmonic guiding potential  $V(\tilde{R}|R) = k_R(\tilde{R} - R)^2/2$ , as described in Sec. 3.3, with  $k_R = 500 \text{ kcal/mol/\AA}^2$ . Similarly, in the third (fourth) set of simulations  $R$  was unconstrained while  $\theta$  was increased (decreased) with  $v_\theta = 0.1 \text{ deg/ps}$  through a harmonic guiding potential  $V(\tilde{\theta}|\theta) = k_\theta(\tilde{\theta} - \theta)^2/2$  with  $k_\theta = 5 \times 10^4 \text{ kcal/mol/rad}^2$ . In these potentials,  $R$  and  $\theta$  are the target values of the reaction coordinates while  $\tilde{R}$  and  $\tilde{\theta}$  represent the instantaneous value of the reaction coordinate as determined from the corresponding atomic coordinates. The SMD trajectories in the reaction coordinate plane (i.e., the loci of points with coordinates  $\{R(t), \theta(t)\}$ ) are shown in Fig. 5.5 for both MOLI and ACI. The different trajectories belonging to different sets of simulations are indicated by numbered arrows. Trajectory points are color-coded according to the current volume overlap of the trans-membrane parts of the two subunits, normalized to the corresponding initial state value.

The reaction coordinate trajectories exhibit distinctive features for each set of simulations, with manifest differences between the two systems.

As soon as the separation  $R$  between the two subunits is increased (set 1) past  $R_0$ , the TMH domains of the  $\alpha\beta$ -apoproteins separate for both MOLI and ACI, an event which is accompanied by a noticeable increase in the fluctuations of the angle reaction coordinate.

While in the case of ACI, the separation of the N-terminal and TMH domains seem to occur simultaneously as  $R$  increases, in contrast, for MOLI, the N-terminal domains of the subunits do not detach until  $R$  becomes larger than  $35 \text{ \AA}$ . On the other hand, the C-terminal domains remain connected in both systems even for separations as large as  $40 \text{ \AA}$ . This difference may also be responsible for the profile of the trajectories in set 1. For MOLI, as  $R$  increases, the trajectories cluster in fairly well distinguished paths, with occasional partial

overlaps, and are subject to large  $\theta$  fluctuations. By performing an additional four SMD runs with pulling speed  $0.01 \text{ \AA/ps}$  (i.e., ten times slower than the previous SMD simulations), it was found (not shown in Fig. 5.5) that the above distinctive features of the paths in set 1 for MOLI prevail for slower pulling speeds as well. However, for these new paths the deviation in the angle reaction coordinate appears to be somewhat larger than the one corresponding to faster pulls. The results suggest that the PMF  $U(\theta|R)$ , for a given  $R > R_0$ , has a broad global minimum with several (at least three) local minima separated by relatively small potential barriers. The location of the minimum is shifted to  $\theta > \theta_0$  values (see below). On the other hand, for ACI, as  $R$  increases, the trajectories remain clustered (albeit with enhanced  $\theta$  fluctuations) suggesting that the PMF  $U(\theta|R)$  has a potential well that is broader than the one for  $R_0$ . However, the dramatic deviation of one of the trajectories from the rest for  $R \gtrsim 35 \text{ \AA}$  suggests that for larger separation  $U(\theta|R)$  may develop a structure with at least two well separated local minima, with one equilibrium angle smaller and the other one larger than  $\theta_0$ .

The behavior of the reaction coordinate trajectories for the simulations in sets 2, 3 and 4 are qualitatively similar for both MOLI and ACI. Already a few  $\text{\AA}$  compression of  $R$  (set 2) increases the overlap of the protein subunits several times, accompanied by a decrease of the angle variable. The trajectories nicely overlap, indicating that the corresponding PMF  $U(\theta|R)$  is similar to the one corresponding to  $R_0$ , with the minimum shifted towards a smaller angle than  $\theta_0$ .

Finally, it is remarkable that for the simulations in sets 3 and 4, in which  $R$  is unconstrained, the latter shows only a slight increase with respect to  $R_0$  as the angle is increased (decreased) by as much as  $20^\circ$  ( $10^\circ$ ). Thus, based on the results of our SMD simulations, one may conclude that in general forced rotation of the relative orientation of the subunits has only very limited effect on their spatial separation. On the other hand, the modification of the distance between the subunits in general has strong impact on the angle between the subunits. This conclusion also provides support to the notion that *the preferred angle*

subunit 1 residue	subunit 2 residue(s)	average distance for breaking links	contact region
$\alpha_1$ GLY <sub>+8</sub>	$\alpha_2$ ALA <sub>+16</sub>	> 40 Å	C-terminus
$\beta_1$ LYS <sub>+7</sub>	$\alpha_2$ ALA <sub>+16</sub> , SER <sub>+19</sub> , ALA <sub>+20</sub>	> 40 Å	C-terminus
$\beta_1$ PRO <sub>+8</sub>	$\alpha_2$ GLY <sub>+15</sub> , SER <sub>+19</sub>	> 40 Å	C-terminus
$\beta_1$ TRP <sub>+9</sub>	$\alpha_2$ GLY <sub>+15</sub>	> 40 Å	C-terminus
$\alpha_1$ PHE <sub>+9</sub>	$\alpha_2$ ILE <sub>+12</sub>	36 - 38 Å	C-terminus
$\beta_1$ ARG <sub>-32</sub>	$\beta_2$ LEU <sub>-27</sub>	32 - 35 Å	N-terminus
$\beta_1$ LEU <sub>-30</sub>	$\alpha_2$ SER <sub>-18</sub>	21 - 22 Å	N-terminus
$\alpha_1$ VAL <sub>-22</sub>	$\alpha_2$ PRO <sub>-14</sub> , SER <sub>-18</sub>	21 - 24 Å	TM

Table 5.1: Average separation distances at which the most important inter-residue links between two LH2 subunits of MOLI are broken in SMD simulations.

subunit 1 residue	subunit 2 residue(s)	average distance for breaking links	contact region
$\alpha_1$ THR <sub>+7</sub>	$\alpha_2$ TRP <sub>+14</sub> , GLN <sub>+15</sub>	> 40 Å	C-terminus
$\alpha_1$ THR <sub>+8</sub>	$\alpha_2$ GLN <sub>+15</sub>	> 40 Å	C-terminus
$\alpha_1$ TRP <sub>+9</sub>	$\alpha_2$ TRP <sub>+14</sub> , GLN <sub>+15</sub>	> 40 Å	C-terminus
$\beta_1$ PRO <sub>+8</sub>	$\alpha_2$ GLN <sub>+15</sub>	> 40 Å	C-terminus
$\beta_1$ TRP <sub>+9</sub>	$\alpha_2$ TYR <sub>+13</sub>	34 - 38 Å	C-terminus
$\alpha_1$ VAL <sub>-21</sub>	$\alpha_2$ ALA <sub>-18</sub> , PRO <sub>-13</sub>	20 - 22 Å	N-terminus
$\alpha_1$ VAL <sub>-22</sub>	$\alpha_2$ ALA <sub>-18</sub> , PRO <sub>-19</sub>	21 - 27 Å	N-terminus
$\alpha_1$ ALA <sub>+2</sub>	$\alpha_2$ LEU <sub>+4</sub>	23 - 25 Å	TM

Table 5.2: Average separation distances at which the most important inter-residue links between two LH2 subunits of ACI are broken in SMD simulations.

*between LH2 subunits is mainly determined by contact interactions between subunits.*

Steered molecular dynamics simulations also provide insights into the key interactions between the subunits: as two subunits are pulled apart, the links between the subunits break from the weakest to the strongest. A link is considered, arguably, to be broken, if the distance between the closest contact between two residues becomes larger than 3 Å. A list of strongest links, as presented in Tables 5.1 and 5.2, provides qualitative information on the basis of SMD simulation data. With increased sampling, more refined analysis in terms of e.g. interaction energies or free energy barriers would become possible. Inspection of the two tables shows that all links in the TMH domain become separated for distances larger than  $R = 25$  Å for both cases. In the N-terminal domain, all links become separated



for distances around 25 Å for ACI and around 35 Å for MOLI. This is consistent with the overlapping volume data shown in Fig. 5.5. On the other hand, there is one group of links in the C-terminal domain for both ACI and MOLI that persists even for separations larger than 40 Å. This group represents the strongest interactions conferring overall stability of the binding of the two subunits. An immediately recognizable difference is that most of the links in ACI are between the  $\alpha$ -apoproteins while they are between the  $\beta$ 1 and  $\alpha$ 2 in MOLI.

### 5.2.5 Calculation of PMF $U(\theta, R)$

The 1D PMFs  $U_i(\theta|R_i)$  along  $\theta$  were determined by using umbrella sampling and WHAM for both MOLI and ACI, for two representative separations, i.e., the equilibrium  $R_0 = 18$  Å and  $R_x = 25$  Å. While the choice for the equilibrium value is obvious, the reason for the  $R_x$  selection is that at this particular distance MOLI and ACI are in qualitatively different separation states. While, for  $R_x$ , the TMH domains of the  $\alpha\beta$ -polypeptides are already separated in both MOLI and ACI, the N-terminal domain is fully separated only in ACI, but not in MOLI; the C-terminal domains are still in contact for both systems.

The computed PMFs, using umbrella sampling and WHAM, are shown in Fig. 5.6. The relative statistical errors in the PMFs were estimated by employing the Monte Carlo bootstrap error analysis method as implemented in Alan Grossfield's WHAM software package [<http://dasher.wustl.edu/alan/wham/>]. The results suggest that, in general, for several hundreds of bins used to sample  $U(\theta|R_0)$ , the relative error is smaller than 10% for any angle. (For clarity the actual error bars are not shown in Fig. 5.6.) For  $R_0$  the PMF for both MOLI and ACI exhibits a nearly parabolic lineshape with minima that practically coincide with the peak positions in the corresponding equilibrium angle distribution histograms in Fig. 5.4, i.e., 45.2° and 38.5°, respectively. In fact, the PMFs calculated from those histograms as  $U_0(\theta) \propto -k_B T \ln[p_0(\theta)]$  matches rather well the bottom part of the full PMF obtained from umbrella sampling and WHAM (data not shown). At the equilibrium

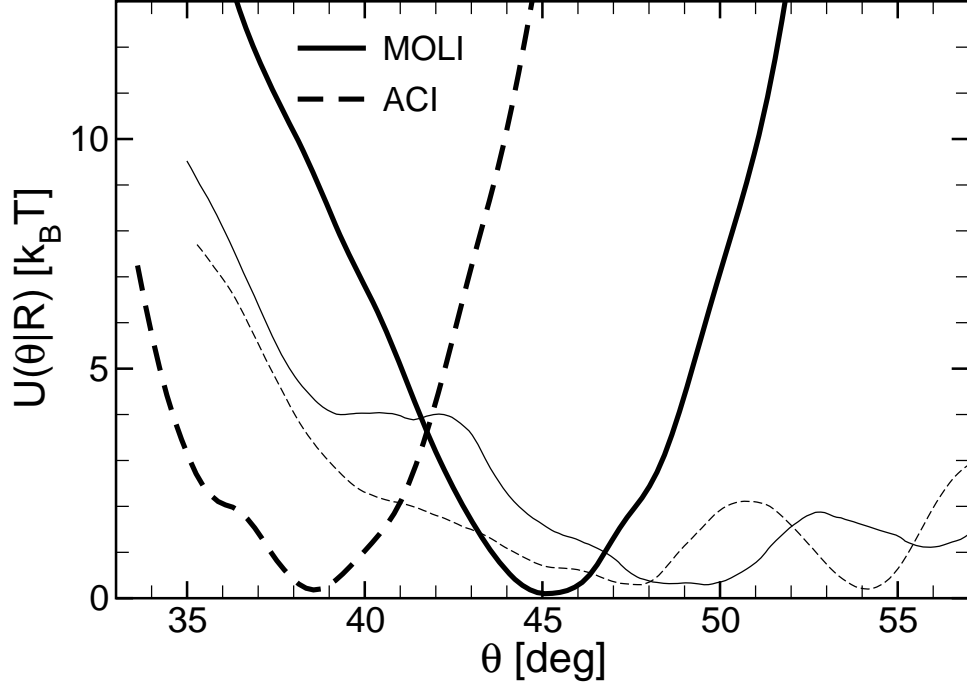


Figure 5.6: Calculated potentials of mean force  $U(\theta|R)$  for both MOLI (solid lines) and ACI (dashed lines) dimers. The thick (thin) curves correspond to  $R = 18 \text{ \AA}$  ( $R = 25 \text{ \AA}$ ).

distance, the PMF of ACI and MOLI are rather similar, indicating that for both systems the angular constraints between adjacent subunits are of comparable strength. At least around their minima, both PMFs can be fitted well with a harmonic potential characterized by the same angular elastic (torsional) constant  $k_\theta \approx 0.18 \text{ kcal/mol/deg}^2$ . Compared to the  $R_0$  case, for  $R_x$  the PMF for both systems widens up and acquires additional features. For MOLI, the PMF exhibits a small plateau at angles around  $40^\circ$ , and a steep downhill region for  $\theta \gtrsim 42^\circ$  that ends in a broad minimum around  $49^\circ$ , followed by a modest potential barrier at  $\sim 53^\circ$ . These features are consistent with the SMD results. Indeed, the angle spread of the SMD trajectories at  $R_x$  extends from  $\sim 40^\circ$  to  $\sim 55^\circ$ . The steep potential barrier in the PMF at  $\lesssim 40^\circ$  explains the lack of trajectory points below this value. Also, the higher trajectory points density along the plateau region and the broad minimum is expected. For ACI, the PMF for  $R_x$  shows a  $\sim 10^\circ$  wide, rather flat region starting from  $39^\circ$ . The fact that the corresponding SMD trajectory points are clustered only in the interval  $38^\circ < \theta < 44^\circ$

does not conflict with the PMF data but raises the question why are there no trajectory points up to angles of  $\sim 50^\circ$ ? There are several possible answers. First, the small number of SMD trajectories may not provide a proper sampling of the angles for  $R_x$ . Secondly, the  $\theta$  self diffusion coefficient may be very small, so that a flat PMF does not lead to significant dispersion. The second well in the PMF at  $54^\circ$  suggests that eventually a second branch of trajectories may appear oriented towards higher angle values as it appears indeed for  $R > 32 \text{ \AA}$ .

### 5.2.6 Preferred Angle

The results obtained so far show that two subunits in van-der-Waals contact (center-to-center distance of  $\sim 18 \text{ \AA}$ ) arrange at a preferred angle with each other. The PMF minima are located at  $45.2^\circ$  for MOLI and at  $38.5^\circ$  for ACI. Both values are very close to the theoretically expected values of  $45^\circ$  (8-fold symmetry) and  $40^\circ$  (9-fold symmetry). The free energy profiles closely match parabolic profiles, characterized by the same angular elastic (torsion) constant  $k_\theta \approx 0.18 \text{ kcal/mol/deg}^2$ . At the same time, the free energy profiles exhibit a clear angular separation: it requires about  $9.3 k_B T$  to force two subunits from MOLI into the angle of  $38.5^\circ$  preferred by subunits from ACI, while about  $14.4 k_B T$  are required to force two subunits from ACI into the angle of  $45.2^\circ$  preferred by subunits from MOLI.

These suggest that the preferred angle between the subunits plays an important role in guiding the assembly of light harvesting complexes into a particular ring size or oligomerization state.

To see whether the surface contact between the two subunits is important in defining a preferred angle, additional calculations where two subunits are constrained to a center to center distance of  $25 \text{ \AA}$  were performed. At this distance, subunits of MOLI are still connected at both their C- and N-terminal domains, and subunits of ACI are connected at their C-terminus only. However, the  $\alpha\beta$  polypeptides in the transmembrane region are no longer in contact for either subunit pairs. In this case, the free energy decreases at

larger angles with considerably more shallow minima than the free energy profiles at the equilibrium distance of 18 Å.

These findings suggest that the exact preferred angle between two subunits is largely defined by the surface interactions in the transmembrane region. Once the  $\alpha\beta$ -polypeptides in the transmembrane region become disconnected, the angular variation increases significantly (see Fig. 5.5). However, angular constraints, albeit more relaxed, still exist even at larger separations of the subunits.

Interestingly, the angle between two subunits is more constrained in ACI than in MOLI. Although the parabola of the free energy profiles at equilibrium distance (18 Å) are similar, the five trajectories of ACI during the steered molecular dynamics simulations show relatively little angular spread after the transmembrane regions are disconnected and even after the links in the N-terminal domain are broken. Much more angular spread can be observed in the case of equivalent SMD runs for MOLI, even for cases when the links in both the C- and N-terminal regions are still intact. As already stated, the configuration space may not have been sampled well due to the fast pulling and small sample size during this simulation.

### 5.3 Prediction of the LH2 Ring Size

The results obtained in the previous section show that the preferred angle plays a crucial role in guiding the assembly of LH2 complexes into a particular ring size. Therefore, the next step in understanding how these rings assemble is to try to determine their oligomerization state by identifying the preferred angle of a dimer from the sole knowledge of the atomic structure of a single subunit. First, a preferred angle was determined from energy minimization of each system. Second, the dimer angle was monitored during free molecular dynamics simulations. Finally, using the FR method, the PMF profile  $U(\theta)$ , determined along the angle RC, was obtained for several systems from each species, in order to iden-

tify the most probable size (i.e. the number of subunits) of the ring that the corresponding species would form. Combining the results from the three approaches, a prediction on the oligomerization state of the light-harvesting complexes was made.

### 5.3.1 System Modeling and MD Simulations

In order to build the MOLI, ACI and X dimer systems from the sole knowledge of the high resolution crystallographic structure of a monomer, one needs the angle  $\theta$  between the two monomers and the radius  $\rho$  of the corresponding ring structures. The radius is related to the distance  $R$  between two neighboring subunits (see Fig. 5.3). In the case when the subunits are tightly packed (i.e. part of a ring),  $R$  is equivalent to the size (in the plane of the ring) of a subunit. Therefore,  $R$  should be independent of the number of subunits contained by a ring. Indeed, the crystallographic structures yield the distance between neighboring subunits  $R_0 \sim 18 \text{ \AA}$  for both MOLI and ACI. Since X is similar to ACI, one can safely assume the distance  $R_0$  between monomers is the same in both systems. The angle  $\theta$  and radius  $\rho = \frac{R}{2} / \sin \frac{\theta}{2}$  determine uniquely the position of the second subunit in the dimer.

Using the VMD [96] *Membrane* plugin, each dimer was immersed in a previously equilibrated hexagonal patch of POPC lipid bilayer with solvated lipid headgroups. Lipids within  $0.8 \text{ \AA}$  of the dimers were removed. The membrane-dimer complex was then solvated in water using the VMD plugin *Solvate*. Two layers of water were added on both sides of the membrane, each of them having  $8 \text{ \AA}$  in thickness. The VMD plugin *Autoionize* was used to randomly place ions such that the resulting total charge of the systems was zero. The characteristics of the built systems are summarized in Table 5.3.

All simulations were performed using the parallel molecular dynamics program NAMD 2.5 [156], CHARMM27 forcefield for lipids and proteins [37], and TIP3 model for water. A cutoff distance of  $12 \text{ \AA}$  (switching function starting at  $10 \text{ \AA}$ ) for van der Waals interactions was assumed. Hexagonal periodic boundary conditions were used with flexible cell enabled, unless otherwise mentioned. Minimum distance between periodic images of the

$\theta_0$	No. of atoms			No. of lipids			No. of waters		
	MOLI	ACI	X	MOLI	ACI	X	MOLI	ACI	X
36°	66569	58654	58381	207	198	197	11577	9418	9409
40°	63918	61249	61271	200	213	214	11006	9613	9613
45°	60196	55763	54879	184	188	183	10480	8901	8867

Table 5.3: The total number of atoms, the number of lipids and of water molecules, of the built dimer systems at initial angles  $\theta_0 = \{36^\circ, 40^\circ, 45^\circ\}$ .

dimer was 32 Å in all cases. Long-range electrostatic interactions were computed using Particle Mesh Ewald method, with the density of the grid points around  $1/\text{\AA}^3$ . Rigid hydrogen bonds was set and, therefore, an integration time step of 2 fs and SHAKE constraint on all hydrogen atoms were used. Impulse-Verlet algorithm was employed in integrating the equations of motions. Both, short-range nonbonded and long-range electrostatic interactions were determined at each time step.

Langevin dynamics was utilized to keep constant temperature  $T = 310$  K in all simulations. The Langevin damping coefficient was set to  $5 \text{ ps}^{-1}$ . Also, the simulations were conducted at constant pressure of 1 atm. using the Nosé-Hoover Langevin piston method with a decay period of 100 fs and a damping timescale of 100 fs.

*Energy minimizations* – For each of the 9 dimer systems, the energy minimization process was carried out using the subsequent 3-step protocol: (1) The system was energy minimized for 50,000 steps by keeping the heavy atoms of the dimer fixed. During this process all bad contacts between atoms have been eliminated, without any change in the relative orientation  $\theta_0$  of the dimer subunits; (2) The system was gradually heated up with 1 K/ps for about 200 ps by continuing to keep the heavy atoms of the dimer fixed. At the end of this process the lipids appeared uniformly distributed around the dimer; (3) By releasing all dimer atoms the entire system was energy minimized for 100,000 steps.

*Free MD simulations* – MOLI, ACI and X dimers, constructed at initial angles  $\theta_0 \in \{36^\circ, 40^\circ, 45^\circ\}$ , embedded in the lipid bilayer and then solvated in water, were equilibrated in the following three stages: (1) During 100,000 energy minimization steps the heavy

atoms of the dimers were fixed, assuring that the angle RC remains set to its initial  $\theta_0$  value and that only the environment (lipids and water) and hydrogen atoms are minimized. 50,000 energy minimization steps followed, having the angle RC restrained by a harmonic potential to the value  $\theta_0$ ; (2) An equilibration of the environment, with the same harmonic restrain of the angle RC to  $\theta_0$ , was performed. A gradual heating of the system of 1 K/ps up to 310 K, with the flexible cell parameter disabled, was performed. After 0.5 ns of equilibration, the lipids arranged themselves nicely around the dimers; (3) The  $\theta$  angle restrain release was followed by free MD simulations for 10 ns of the entire system at 310 K in the NPT ensemble.

*SMD simulations* – Steered molecular dynamics (SMD) simulations were employed to calculate the PMF  $U(\theta)$  for  $\theta \in [30^\circ, 50^\circ]$ , using the FR method. The approach requires both *forward* (F) and *reverse* (R) pulls (see Sec. 3.8) between two states of the system that have the angle RC  $\theta$  equilibrated around a desired value  $\theta_{eq}$ . Different systems, built at initial angles  $\theta_0 = \{36^\circ, 40^\circ, 45^\circ\}$ , as previously described, were used to calculate three PMFs  $U(\theta)$  for each species. In order to guarantee fast equilibrations and Gaussian distribution of the works, the stiff spring approximation (see Sec. 3.4) was employed. All pulls were constant velocity ( $v = 20$  deg/ns) SMD simulations implemented in NAMD [156] using its Tcl interface.

The FR method gives the free energy difference between the two equilibrated states. For best accuracy of  $U(\theta)$  along the studied range, the  $[30^\circ, 50^\circ]$  interval was split into 10 subintervals of 2 degrees each. A single protocol was employed for all the systems, requiring the following simulations: (i) Eleven 600 ps long equilibration runs of the entire system in the states between subintervals ( $\theta_{eq} = \{30^\circ, 32^\circ, 34^\circ, \dots, 50^\circ\}$ ). These equilibrations restrained the angle RC to  $\theta_{eq}$  using a harmonic guiding potential, with  $k_\theta = 5 \times 10^5$  kcal/mol/rad<sup>2</sup>. The last 400 ps of the equilibration simulations were used to select five states, 100 ps apart, as initial configurations for the F and R pulls. (ii) Five F and five R SMD simulations along each subinterval, starting from the saved initial configurations

for each  $\theta_{eq}$  (total of  $2 \times 50$  simulations) were performed. The F and R pulls correspond to increasing  $\theta$  with constant velocity, and to decreasing it, respectively. Due to the large number of pulls required to determine  $U(\theta)$  for  $[30^\circ, 50^\circ]$ , the pulls were employed on the smallest number of subintervals that would clearly yield the position of the PMF minimum.

### 5.3.2 Energy Minimization

For each of the studied species (MOLI, ACI and X) three dimer systems were built at initial angles of  $\theta_0 = \{36^\circ, 40^\circ, 45^\circ\}$ . During the 100,000 energy minimization steps of each system, the dimer angle  $\theta$  was monitored and plotted in Fig. 5.7. The values of the angles obtained at the end of the minimization process are recorded in the third column of Table 5.4. For MOLI, the systems with  $\theta_0 = 36^\circ$  and  $\theta_0 = 40^\circ$  minimized to values of  $\theta_{final}$  around  $40^\circ$ . The other system remained at its initial angle  $\theta_0 = 45^\circ$ , reproducing the dimer angle reported in the crystal structure [12]. For ACI, the result is fairly the same, except that the  $45^\circ$  system's angle dropped down to  $43.3^\circ$ . The  $36^\circ$  and  $40^\circ$  systems yielded final angles that suggest a 9-ring for ACI, as reported in the crystal structure [13]. For the unknown structure X, the results clearly suggest a 9-ring formation, all the dimer systems yielding the energy minimization final angle  $\theta_{final}$  around  $40^\circ$ .

### 5.3.3 Free MD Simulation

In the second approach, the dimer angle  $\theta$  was monitored along a 10 ns free molecular dynamics simulation of each system. Figure 5.8 shows in the right panels the angle evolution along the 10 ns trajectories, and in the left panels the angle distribution and mean value over the last 7 ns of free MD simulation. The mean values  $\langle \theta \rangle$  and the angle distributions peak position  $\theta_{peak}$  are recorded in the Table 5.4.

The angle of MOLI dimer system at  $\theta_0 = 36^\circ$  (Fig. 5.8a) is almost constantly decreasing, indicating that this system might be unstable, probably because the dimer angle is so



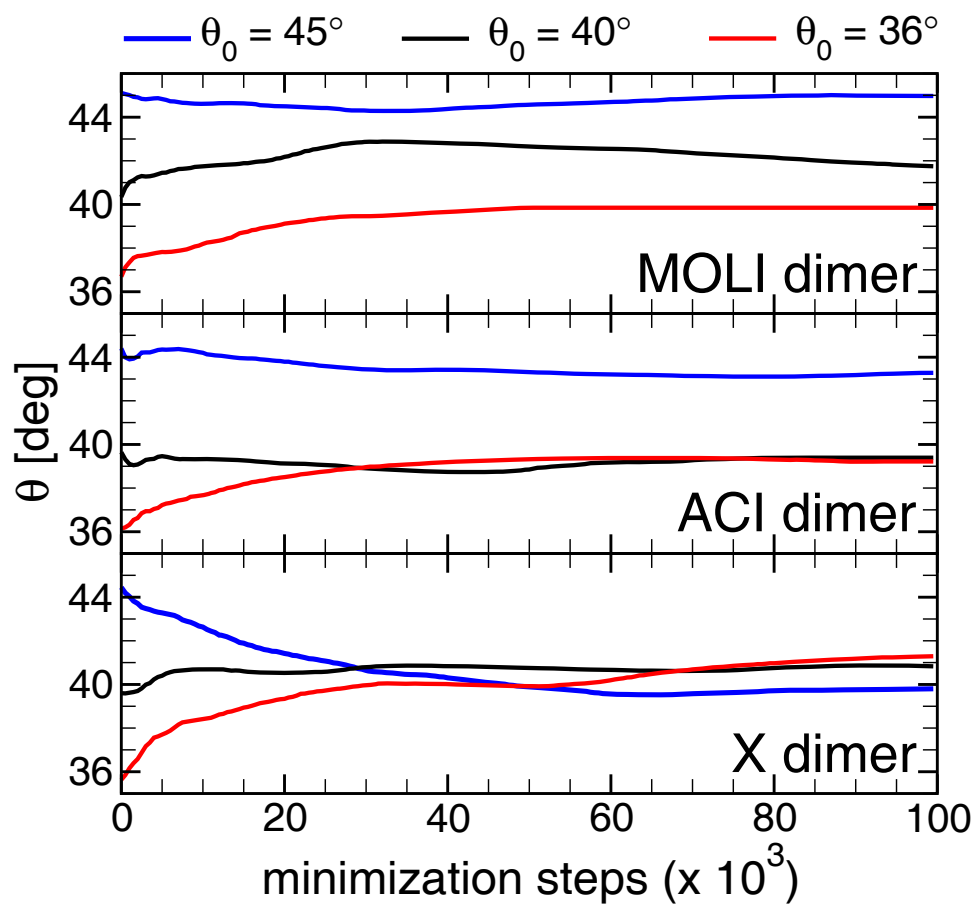


Figure 5.7: Evolution of the angle reaction coordinate  $\theta$  during energy minimization of MOLI (top), ACI (center) and X dimers (bottom), respectively. The curves for each dimer correspond to initial  $\theta_0$  angle values of  $36^\circ$  (red line),  $40^\circ$  (black line) and  $45^\circ$  (blue line).

Species	Angle	En.min.	Free MD		PMF $U(\theta)$
	$\theta_0$	$\theta_{final}$	$\langle\theta\rangle$	$\theta_{peak}$	$\theta_{min}$
MOLI	36	39.9	31.7	31.2	35.0
	40	41.8	40.3	41.1	42.6
	45	45.0	40.7	41.2	39.9
ACI	36	39.2	38.0	38.3	36.0
	40	39.4	39.3	39.5	39.1
	45	43.3	37.7	37.5	41.3
X	36	41.3	43.7	43.5	37.7
	40	40.8	36.5	36.5	38.5
	45	39.8	39.6	41.1	41.2

Table 5.4: Angle values determined from the three approaches for the nine systems: (i) final angle value  $\theta_{final}$  after energy minimization, (ii) mean value of  $\theta$  (averaged over the last 7 ns) and the peak position of its distribution, obtained from the 10 ns free MD simulations, and (iii) angle value  $\theta_{min}$  corresponding to the global minimum of the PMF  $U(\theta)$ . All angles are reported in degrees.

far from its expected value of  $45^\circ$  (MOLI forms and 8-fold LH2 ring). The  $40^\circ$  system's dimer angle value fluctuates for 7.5 ns around  $41^\circ$  and then suddenly drops down, maintaining its fluctuations for the last 2 ns around  $38^\circ$  (Fig. 5.8b). The angle of the  $45^\circ$  system mainly varies around the mean value  $\langle\theta\rangle \sim 40^\circ$ , having just a brief drop of  $3^\circ$  right after passing 6 ns of simulation (Fig. 5.8c). Surprisingly, the free MD simulations suggest that the most stable configuration of the built MOLI dimers corresponds to a 9-ring, and not to the expected 8-ring light harvesting complex.

In the case of ACI, the three MD simulations are rather similar. The dimer angles fluctuate around  $\langle\theta\rangle \sim 38^\circ$  for the  $36^\circ$  (Fig. 5.8d) and  $45^\circ$  (Fig. 5.8f) systems, and around  $40^\circ$  for the system having initial angle of  $\theta_0 = 40^\circ$  (Fig. 5.8e). The latter case has the smallest fluctuations ( $\sim 1.4^\circ$ ), indicating that the system built at the expected dimer angle of  $40^\circ$  is most preferred by the subunits. All the three MD simulations suggest that a light harvesting complex of 9 subunits (as expected) will be formed from the ACI subunits.

The X dimer system with initial angle  $\theta_0 = 36^\circ$  (Fig. 5.8g) quickly increases its angle value to  $\sim 45^\circ$ , but after  $\sim 1$  ns stabilizes around  $43^\circ$ . The dimer angle of the  $40^\circ$  system (Fig. 5.8h) goes very rapidly to  $36^\circ$  and then continues to fluctuate around this value, except

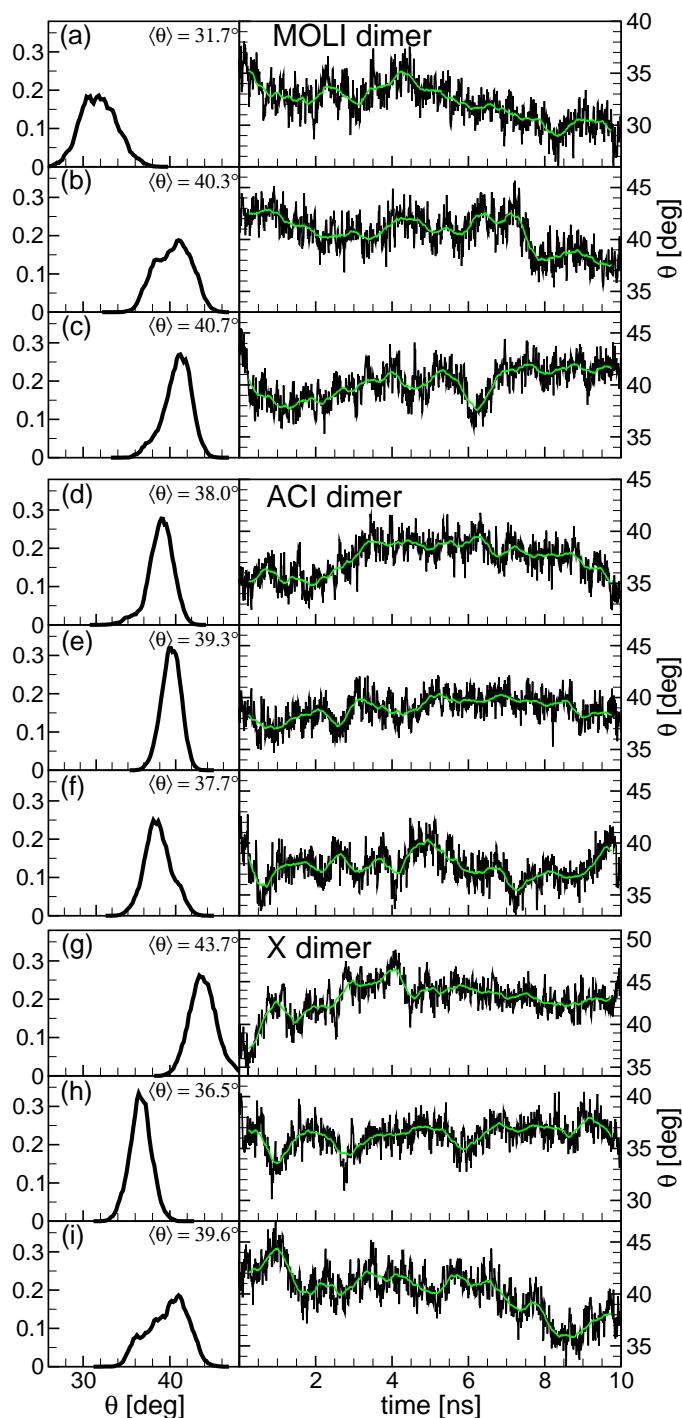


Figure 5.8: Distribution histogram (left panels) and time-evolution (right panels) of the angle reaction coordinate  $\theta$  for MOLI (a,b,c), ACI (d,e,f) and X (g,h,i) dimers corresponding to free MD simulations starting from conformations characterized by  $\theta_0 = 36^\circ$  (a,d,g),  $\theta_0 = 40^\circ$  (b,e,h) and  $\theta_0 = 45^\circ$  (c,f,i). For each case the listed mean value  $\langle \theta \rangle$  was calculated using the last 7 ns of the corresponding MD trajectory.

for a few "dips" around 1 ns, 2.8 ns and 5.8 ns. In the case of the  $45^\circ$  system (Fig. 5.8i), the  $\theta$  angle declines and then fluctuates around  $41^\circ$  for about 5 ns. During the next 3.5 ns its value decrease to  $\sim 36^\circ$  is promptly followed by an increase up to  $\sim 38^\circ$ . Although the mean value of the dimer angle  $\langle \theta \rangle \sim 40^\circ$  indicates a 9-ring light-harvesting complex formation, the stability of this system is questionable due to these large variations in  $\theta$ . The three MD simulations are suggesting that a 10-, 9- and, perhaps, an 8-ring could form from the subunits of the X species.

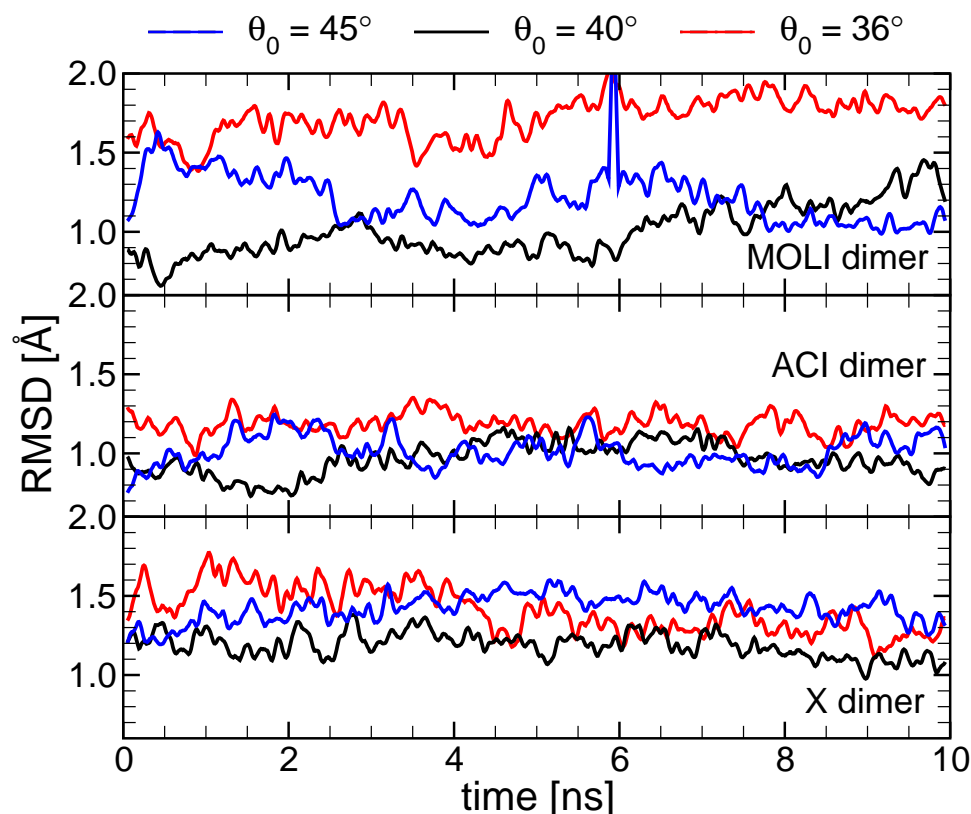


Figure 5.9: Time dependence of the RMSD of the dimer transmembrane  $C_\alpha$  atoms with respect to the crystal structure in case of MOL1 (top) and ACI (center), and to the build structure in case of the X dimer (bottom).

The root mean square deviation (RMSD) values were calculated along the 10 ns trajectories, for the transmembrane  $C_\alpha$  atoms of the dimers, with respect to the crystal structure for MOL1 and ACI, and to the built structure for X (since there is no crystal structure for the X dimer) (see Fig.5.9). For MOL1 and ACI, one can easily notice that the RMSD in-

creases when the  $\theta$  angle is further from the crystallographic value. For MOLI (top panel), as expected, the  $36^\circ$  system has the largest RMSD, because the  $\theta$  angle (and so the dimer configuration) is the furthest from the expected  $45^\circ$  value. For ACI (central panel), since  $\theta$  fluctuates around  $40^\circ$  for the three simulations, all the RMSDs are small and fairly similar. For X (bottom panel), the RMSD is calculated with respect to the built structure, so it rather reflects how much the dimers change along the free MD simulations. Indeed, the greatest RMSD is found for the  $45^\circ$  system, where the largest  $\theta$  angle variations are found, but the difference between the three RMSDs is not that significant.

### 5.3.4 Potential of Mean Force $U(\theta)$

Third and final approach used to determine the preferred angle between two subunits is based on the potential of mean force (PMF) profile along the dimer angle reaction coordinate (RC)  $\theta$ . The PMFs of the nine systems, i.e. MOLI, ACI and X species built at initial dimer angles  $\theta_0 = \{36^\circ, 40^\circ, 45^\circ\}$ , were obtained from five forward (F) and five reverse (R) constant velocity steered molecular dynamics (cv-SMD) simulations, described in Sec. 5.3.1. Each  $U_{\theta_0}(\theta)$  was built by stitching together free energy profiles calculated on two degree intervals. Figure 5.10 shows the obtained PMFs along the angle  $\theta$  for MOLI (top), ACI (center) and X (bottom) dimers built at  $\theta_0$  angles of  $36^\circ$  (dash-dotted lines),  $40^\circ$  (dotted lines) and  $45^\circ$  (solid lines). The angle values  $\theta_{min}$  where the free energy profiles have their minimum are recorded in the Table 5.4.

Clearly, the location of the  $U(\theta)$  minima for the nine systems is not exactly  $36^\circ$ ,  $40^\circ$  or  $45^\circ$ , corresponding to the preferred angles of the 10-, 9- or 8-ring complexes, respectively. This is not surprising though, if one analyzes the dimer angle distributions during free MD simulations of built LH2 rings (see Fig. 5.11). The preferred dimer angle is between  $42^\circ$  and  $48^\circ$  for the MOLI 8-ring, and between  $37^\circ$  and  $43.5^\circ$  for the ACI 9-ring. Moreover, for ACI there is not even a single dimer in the light harvesting ring that has a preferred angle within  $1^\circ$  of the expected  $40^\circ$ . For both LH2 complexes, the sum of all the dimer

angles at any time is  $360^\circ$ , yielding the average preferred angle equal to the expected angle of  $45^\circ$  and  $40^\circ$  for the 8- and 9-ring, respectively. Therefore, one would expect the PMF calculations along the dimer angle  $\theta$  to yield  $\theta_{min}$  values within  $3^\circ - 3.5^\circ$  from the preferred angle.

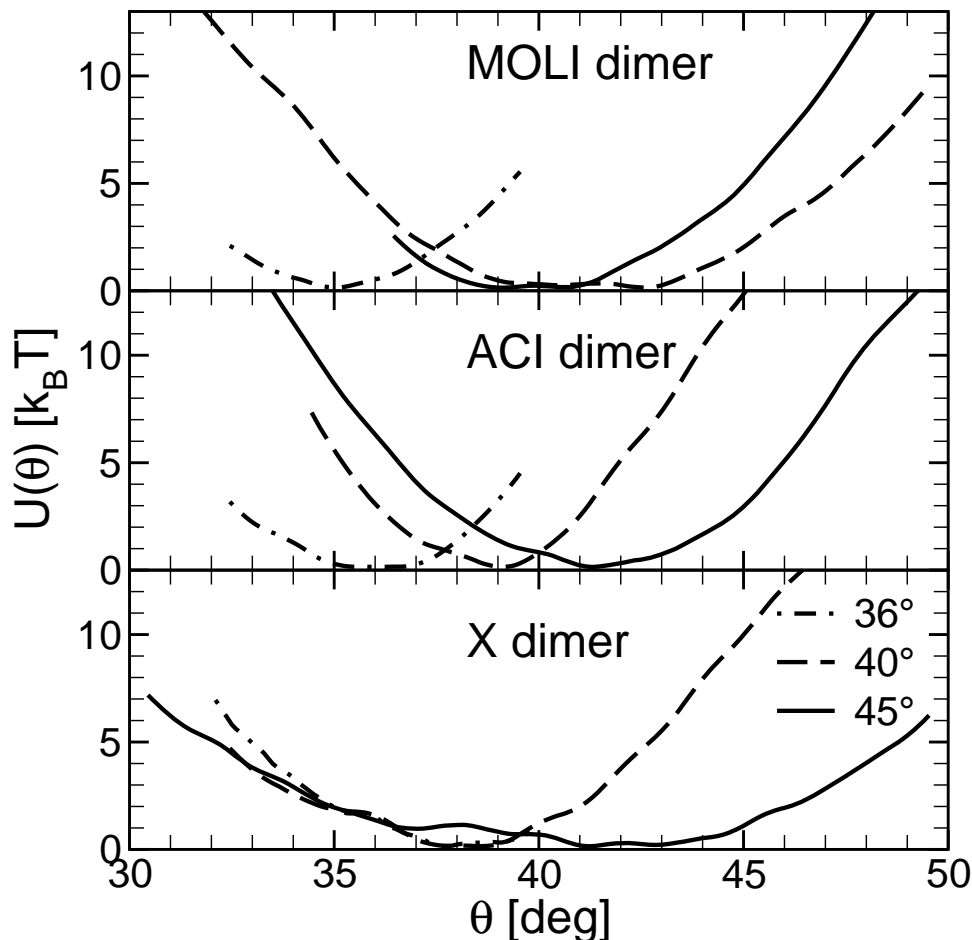


Figure 5.10: Potential of mean force  $U(\theta)$  calculated using the FR method described in the text for MOL (dashed line), ACI (dash-dotted line) and X (solid line) dimers.

On the other hand, due to the high velocity forward and reverse pulls used in determining the PMFs, the lipids surrounding the dimer complex are “frozen” during these short (time) SMD simulations. Since the dimer complex is strongly interacting with its local lipid environment, the change in the dimer angle cannot be too large during one fast pull. The maximum possible change in  $\theta$ , without deforming the dimer complex, was established to be around  $2^\circ$ . Hence, the determined PMFs are depending to some degree on the

initial configuration of the system, i.e. on the initial dimer angle  $\theta_0$ . Thus, shifts of  $\theta_{min}$  are, in general, within a few degrees from  $\theta_0$ . Consequently, one should rationalize  $\theta_{min}$  in terms of direction and relative size of shift from the initial value  $\theta_0$ , rather than in terms of closeness to a certain preferred angle.

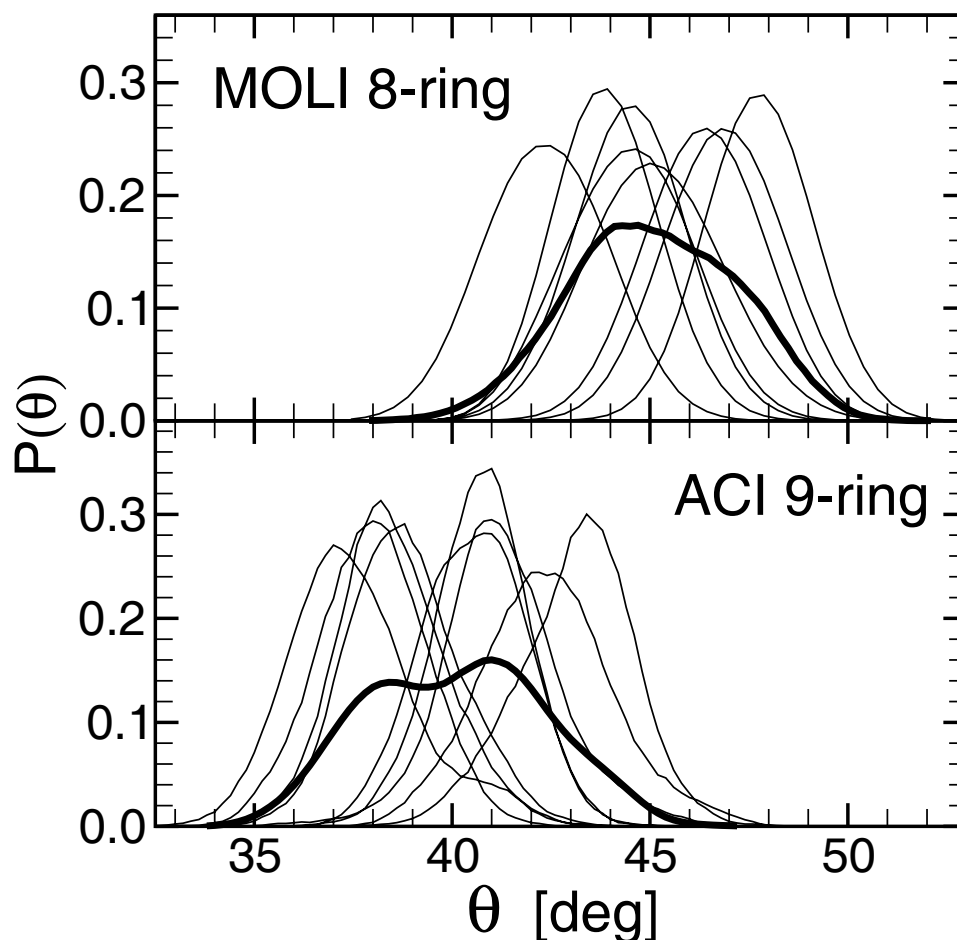


Figure 5.11: Distribution function  $P(\theta)$  of the relative orientation angle  $\theta$  between adjacent subunits (thin lines) corresponding to the eight subunit MOLI (top panel) and nine subunit ACI (bottom panel) LH2 rings. The distribution histograms for were determined from 3 ns long equilibrium MD trajectories. The average angle distributions for all subunit pairs are also shown (solid lines).

$U_{45^\circ}$  from MOLI indicates that the subunits would most likely form a 9-mer, but the wide minimum of  $U_{40^\circ}$ , spreading from  $39^\circ - 43^\circ$ , suggests that the subunits could form not only 9-rings, but 8-rings as well. This result is also implied by the shifts of the two PMFs towards each other. The  $\theta_{min}$  (PMF minimum position) values indicate that a 9-ring

is preferred. Although  $U_{36^\circ}$  keeps the position of its minimum around  $\theta_0$  ( $\theta_{min} \sim 35^\circ$ ), this system could be discarded on the grounds that it seems unstable (large  $\theta$  angle fluctuations) during the 10 ns free MD simulation. Based on the results from the three approaches, the MOLI subunits would most likely form 9-rings, although the expected 8-rings are possible as well.

For ACI,  $U_{40^\circ}$  keeps its minimum close to  $\theta_0$ . As in the case of MOLI, the  $45^\circ$  system shifts its minimum towards  $40^\circ$  ( $\theta_{min} \sim 41^\circ$ ). The two PMFs clearly indicate that the ACI subunits would organize into a 9-ring. The  $2^\circ$  broad minimum around  $\theta_0$  of  $U_{36^\circ}$  includes the possibility of the 10-ring formation as well. But, taking into account the energy minimization and free MD simulations as well, the 9-mer LH2 structure is clearly preferred by the ACI subunits.

In the case of X,  $U_{36^\circ}$  and  $U_{45^\circ}$  are both shifting their minima towards  $40^\circ$ , while the  $40^\circ$  system's minimum remains around  $\theta_0$  ( $\theta_{min} \sim 38.5^\circ$ ).  $U_{45^\circ}$  shows that it would take less than  $2 k_B T$  to change the dimer angle from  $40^\circ$  to  $36^\circ$  or  $45^\circ$ . The large angle differences recorded by the corresponding free MD simulations are consistent with the PMF findings, indicating that this system might be unstable and could be discarded. Considering all the three approaches, most likely that the X subunits would form 9-mers.

## 5.4 Conclusions

Understanding the assembly of a protein complex requires that one addresses a set of inter-related questions: (1) What is the temporal order in which parts are put together? (2) What interactions stabilize the protein complex or parts of the complex? (3) What factors govern the (reproducible) stable geometry of the complex?

First part of the chapter focused on the assembly of a two-subunit complex from two fully formed  $\alpha\beta$ -subunits with three BChl and one carotenoid bound. Spectroscopic observations support the assumption that this is an important step in the formation of light



harvesting complexes as they show a progression from monomeric BChl (777 nm) to a BChl dimer bound to an  $\alpha\beta$ -subunit (820 nm) to a complex formed of two  $\alpha\beta$ -subunits (851 nm) [157, 158]. It is, however, unknown whether carotenoids are already incorporated at this step and how closely the structure of the  $\alpha\beta$ -subunit at this point matches the structure of the  $\alpha\beta$ -subunit in the crystal structures, from which we took the coordinates.

An initial analysis of fluctuations in a two-subunit complex equilibrated in a lipid-water environment revealed that the hydrophobic core region is rather rigid. Fluctuations in the centers of mass of the core regions in free molecular dynamics runs were small enough to allow for a meaningful definition of global coordinates (distance, angle) of the subunits as introduced above.

In order to probe the assembly process of two subunits, two sets of calculations with different philosophies were performed. In one set of calculations, the subunits were constrained at different angles, allowing the systems to equilibrate under these constraints. Using umbrella sampling, information about the free energy profile as a function of angle was extracted. Pulling two subunits apart through steered molecular dynamics gave complementary information into key events of the binding/unbinding process and a rough order of interaction strengths.

The results suggested that the *preferred angle between two subunits* plays an important role in guiding the assembly of light harvesting complexes into a particular oligomerization state. Determining this angle from the sole knowledge of the atomic structure of a single subunit was the focus of the second part of this chapter.

In order to determine the preferred angle of a dimer, three sets of calculations were performed. First, a set of energy minimizations of dimers built at three different initial angle values was employed. Then, the systems were freely equilibrated. Finally, the “FR method” was applied to calculate the potential of mean force profiles along the angle for the three systems from each species.

The combined results of the three methods indicate that all the three species (MOLI,

ACI and X) will most likely form *light-harvesting ring complexes containing nine subunits*. For MOLI, both the energy minimization and the PMF approaches suggest the distinct possibility of the expected 8-ring [12] formation as well. This study makes a very important prediction for MOLI, namely that not only 8-, but also 9-ring LH2s are also possible. The expected 9-mer LH2 ring of ACI [13] was the only predicted oligomerization state. For X, the equilibrations and the PMFs suggest that subunits could structure into 10-rings as well.

Computationally inexpensive SMD calculations were used to obtain qualitative, but not quantitative information about the factors stabilizing the connection between two subunits. The results from these calculations suggest that links in the terminal domains play an important role in stabilizing the complex. The strongest interactions are all found in the C-terminal domain. Interactions in the TMH domain are weakest (see Tables 5.1,5.2). Braun *et al.* [153] demonstrated formation of functional light harvesting complexes from polypeptides in which all amino acids in the TMH domain except for the ligating His were replaced by alternating pairs of alanine and leucine residues. An addition of four amino acids in the C-terminal domain resulted in a complete loss of light harvesting complex formation. This study supports a crucial role the C-terminal domain in connecting subunits to form complete light harvesting complexes, while indicating a less important role of the TMH domain.

# References

- [1] T. W. Allen, O. S. Andersen, and B. Roux, *Biophys. J.*, **90**(10), 3447–3468 (2006). Ion permeation through a narrow channel: Using gramicidin to ascertain all-atom molecular dynamics potential of mean force methodology and biomolecular force fields.
- [2] T. Bastug, A. Gray-Weale, S. M. Patra, and K. S., *Biophys. J.*, **90**(7), 2285–2296 (2006). Role of protein flexibility in ion permeation: A case study in gramicidin A.
- [3] T. W. Allen, O. S. Andersen, and B. Roux, *J. Gen. Physiol.*, **124**(6), 679–690 (2004). On the importance of atomic fluctuations, protein flexibility, and solvent in ion permeation.
- [4] S. Edwards, B. Corry, S. Kuyucak, and S. H. Chung, *Biophys. J.*, **83**(3), 1348–1360 (2002). Continuum electrostatics fails to describe ion permeation in the gramicidin channel.
- [5] M. O. Jensen, S. Park, E. Tajkhorshid, and K. Schulten, *Proc. Natl. Acad. Sci. U.S.A.*, **99**(10), 6731–6736 (2002). Energetics of glycerol conduction through aquaglyceroporin GlpF.
- [6] L. Janosi, I. Kosztin, and A. Damjanovic, *J. Chem. Phys.*, **125**(1), 014903 (2006). Theoretical prediction of spectral and optical properties of bacteriochlorophylls in thermally disordered LH2 antenna complexes.
- [7] T. Renger, V. May, and O. Kühn, *Phys. Rep.*, **343**(3), 138–254 (2001). Ultrafast excitation energy transfer dynamics in photosynthetic pigment-protein complexes.
- [8] V. Chernyak, W. M. Zhang, and S. Mukamel, *J. Chem. Phys.*, **109**(21), 9587–9601 (1998). Multidimensional femtosecond spectroscopies of molecular aggregates and semiconductor nanostructures: The nonlinear exciton equations.
- [9] S. Mukamel, *Principles of Nonlinear Optical Spectroscopy*, Oxford University Press, New York, 1995.
- [10] V. May and O. Kühn, *Charge and Energy Transfer Dynamics in Molecular Systems*, WILEY-VCH, Berlin, 2000.
- [11] H. van Amerongen, L. Valkunas, and R. van Grondelle, *Photosynthetic Excitons*, World Scientific, Singapore, 2000.

- [12] J. Koepke, X. Hu, C. Muenke, K. Schulten, and H. Michel, *Structure*, **4**, 581–597 (1996). The crystal structure of the light-harvesting complex II (B800-850) from *Rhodospirillum rubrum*.
- [13] G. McDermott, S. Prince, A. Freer, A. Hawthornthwaite-Lawless, M. Papiz, R. Cogdell, and N. Isaacs, *Nature*, **374**, 517–521 (1995). Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria.
- [14] X. Hu, T. Ritz, A. Damjanovic, F. Autenrieth, and K. Schulten, *Q. Rev. Biophys.*, **35**(1), 1–62 (2002). Photosynthetic apparatus of purple bacteria.
- [15] M. Yang, R. Agarwal, and G. R. Fleming, *J. Photochem. Photobiol. A-Chem.*, **142**(2-3), 107–119 (2001). The mechanism of energy transfer in the antenna of photosynthetic purple bacteria.
- [16] T. Sundstrom, V. Pullerits and R. van Grondelle, *J. Phys. Chem. B*, **103**, 2327–2346 (1999). Photosynthetic light-harvesting: reconciling dynamics and structure of purple bacterial LH2 reveals function of photosynthetic unit.
- [17] H. M. Wu, M. Ratsep, R. Jankowiak, R. J. Cogdell, and G. J. Small, *J. Phys. Chem. B*, **101**(38), 7641–7653 (1997). Comparison of the LH2 antenna complexes of *Rhodospseudomonas acidophila* (strain 10050) and *Rhodobacter sphaeroides* by high-pressure absorption, high-pressure hole burning, and temperature-dependent absorption spectroscopies.
- [18] L. M. P. Beekman, R. N. Frese, G. J. S. Fowler, R. Picorel, R. J. Cogdell, I. H. M. vanStokkum, C. N. Hunter, and R. vanGrondelle, *J. Phys. Chem. B*, **101**(37), 7293–7301 (1997). Characterization of the light-harvesting antennas of photosynthetic purple bacteria by stark spectroscopy .2. LH2 complexes: Influence of the protein environment.
- [19] S. Georgakopoulou, R. N. Frese, E. Johnson, C. Koolhaas, R. J. Cogdell, R. van Grondelle, and G. van der Zwan, *Biophys. J.*, **82**(4), 2184–2197 (2002). Absorption and CD spectroscopy and modeling of various LH2 complexes from purple bacteria.
- [20] O. J. G. Somsen, R. van Grondelle, and H. van Amerongen, *Biophys. J.*, **71**(4), 1934–1951 (1996). Spectral broadening of interacting pigments: Polarized absorption by photosynthetic proteins.
- [21] G. D. Scholes and G. R. Fleming, *J. Phys. Chem. B*, **104**(8), 1854–1868 (2000). On the mechanism of light harvesting in photosynthetic purple bacteria: B800 to B850 energy transfer.
- [22] A. Damjanovic, I. Kosztin, U. Kleinekathofer, and K. Schulten, *Phys. Rev. E*, **65**(3 Pt 1), 031919 (2002). Excitons in a photosynthetic light-harvesting system: a combined molecular dynamics, quantum chemistry, and polaron model study.

- [23] Z. He, V. Sundstrom, and T. Pullerits, *J. Phys. Chem. B*, **106**(44), 11606–11612 (2002). Influence of the protein binding site on the excited states of bacteriochlorophyll: DFT calculations of B800 in LH2.
- [24] X. Hu, A. Damjanovic, T. Ritz, and K. Schulten, *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 5935–5941 (1998). Architecture and mechanism of the light-harvesting apparatus of purple bacteria.
- [25] J. A. Ihalainen, J. Linnanto, P. Myllyperkio, I. H. M. van Stokkum, B. Ucker, H. Scheer, and J. E. I. Korppi-Tommola, *J. Phys. Chem. B*, **105**(40), 9849–9856 (2001). Energy transfer in LH2 of Rhodospirillum rubrum, studied by subpicosecond spectroscopy and configuration interaction exciton calculations.
- [26] J. Linnanto, J. E. I. Korppi-Tommola, and V. M. Hellenius, *J. Phys. Chem. B*, **103**(41), 8739–8750 (1999). Electronic states, absorption spectrum and circular dichroism spectrum of the photosynthetic bacterial LH2 antenna of Rhodospseudomonas acidophila as predicted by exciton theory and semiempirical calculations.
- [27] T. Meier, Y. Zhao, V. Chernyak, and S. Mukamel, *J. Chem. Phys.*, **107**(10), 3876–3893 (1997). Polarons, localization, and excitonic coherence in superradiance of biological antenna complexes.
- [28] J. Ray and N. Makri, *J. Phys. Chem. A*, **103**(47), 9417–9422 (1999). Short-range coherence in the energy transfer of photosynthetic light-harvesting systems.
- [29] S. J. Jang and R. J. Silbey, *J. Chem. Phys.*, **118**(20), 9324–9336 (2003). Single complex line shapes of the B850 band of LH2.
- [30] X. Hu and K. Schulten, *Physics Today*, **50**, 28–34 (1997). How nature harvests sunlight.
- [31] J.-P. Zhang, R. Fujii, P. Qian, T. Inaba, T. Mizoguchi, and Y. Koyama, *J. Phys. Chem. B*, **104**, 3683–3691 (2000). Mechanism of the carotenoid-to-bacteriochlorophyll energy transfer via the  $S_1$  state in the LH2 complexes from purple bacteria.
- [32] I. Mercer, I. Gould, and D. Klug, *J. Phys. Chem. B*, **103**(36), 7720–7727 (1999). A quantum mechanical/molecular mechanical approach to relaxation dynamics: Calculation of the optical properties of solvated bacteriochlorophyll-a.
- [33] N. Makri, *J. Phys. Chem. B*, **103**(15), 2823–2829 (1999). The linear response approximation and its lowest order corrections: An influence functional approach.
- [34] K. Schulten and M. Tesch, *Chem. Phys.*, **158**, 421–446 (1991). Coupling of protein motion to electron transfer: Molecular dynamics and stochastic quantum mechanics study of photosynthetic reaction centers.
- [35] S. A. Egorov, K. F. Everitt, and J. L. Skinner, *J. Phys. Chem. A*, **103**(47), 9494–9499 (1999). Quantum dynamics and vibrational relaxation.

- [36] A. D. MacKerell Jr., D. Bashford, M. Bellott, et al., *FASEB J.*, **6**(1), A143–A143 (1992). Self-consistent parameterization of biomolecules for molecular modeling and condensed phase simulations.
- [37] A. D. MacKerell Jr., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, *J. Phys. Chem. B*, **102**(18), 3586–3616 (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins.
- [38] M. Schlenkrich, J. Brickmann, A. D. MacKerell Jr., and M. Karplus in *Biological Membranes: A Molecular Perspective from Computation and Experiment*, K. M. Merz and B. Roux, Eds.; Birkhauser, Boston, 1996; pages 31–81.
- [39] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, *J. Chem. Phys.*, **79**, 926–935 (1983). Comparison of simple potential functions for simulating liquid water.
- [40] S. E. Feller, Y. H. Zhang, R. W. Pastor, and B. R. Brooks, *J. Chem. Phys.*, **103**(11), 4613–4621 (1995). Constant pressure molecular dynamics simulation — the Langevin piston method.
- [41] T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.*, **98**, 10089–10092 (1993). Particle mesh Ewald. An  $N \cdot \log(N)$  method for Ewald sums in large systems.
- [42] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, *J. Comp. Chem.*, **26**(16), 1781–1802 (2005). Scalable molecular dynamics with NAMD.
- [43] S. Miyamoto and P. A. Kollman, *J. Comp. Chem.*, **13**(8), 952–962 (1992). SETTLE: An analytical version of the SHAKE and RATTLE algorithm for rigid water molecules.
- [44] M. G. Cory, M. C. Zerner, X. Hu, and K. Schulten, *J. Phys. Chem. B*, **102**(39), 7640–7650 (1998). Electronic excitations in aggregates of bacteriochlorophylls.
- [45] J. Ridley and M. Zerner, *Theor. Chim. Acta*, **32**, 111–134 (1973). An intermediate neglect of differential overlap technique for spectroscopy: Pyrrole and the azines.
- [46] M. Zerner, G. Loew, R. Kirchner, and U. J. Mueller-Westerhoff, *J. Am. Chem. Soc.*, **102**, 589–599 (1980). An INDO technique for spectroscopy of transition metal complexes.
- [47] J. Linnanto and J. Korppi-Tommola, *J. Comp. Chem.*, **25**(1), 123–138 (2004). Semiempirical PM5 molecular orbital study on chlorophylls and bacteriochlorophylls: comparison of semiempirical, ab initio, and density functional results.

- [48] A. Damjanovic, H. M. Vaswani, P. Fromme, and G. R. Fleming, *J. Phys. Chem. B*, **106**(39), 10251–10262 (2002). Chlorophyll excitations in photosystem I of *Synechococcus elongatus*.
- [49] HyperChem(TM), Hypercube, Inc., 1115 NW 4th Street, Gainesville, Florida 32601, USA.
- [50] Gaussian 98. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. M. V. G. Zakrzewski, R. E. Stratmann, J. C. Burant, J. M. M. S. Dapprich, A. D. Daniels, K. N. Kudin, M. C. Strain, J. T. O. Farkas, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. A. C. Pomelli, S. Clifford, J. Ochterski, G. A. Petersson, Q. C. P. Y. Ayala, K. Morokuma, D. K. Malick, A. D. Rabuck, J. B. F. K. Raghavachari, J. Cioslowski, J. V. Ortiz, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, D. J. F. R. L. Martin, T. Keith, M. A. Al-Laham, C. Y. Peng, C. G. A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. G. Johnson, M. W. W. W. Chen, J. L. Andres, M. Head-Gordon, E. S. Replogle, and J. A. Pople; Gaussian Inc., Pittsburgh, PA, (1998).
- [51] K. Visscher, H. Bergstrom, V. Sundström, C. Hunter, and R. van Grondelle, *Photosynth. Res.*, **22**, 211–217 (1989). Temperature dependence of energy transfer from the long wavelength antenna Bchl-896 to the reaction center in *Rhodospirillum Rubrum*, *Rhodobacter Sphaeroides* (w.t. and M21 mutant) from 77 to 177 K, studied by picosecond absorption spectroscopy.
- [52] I. Kosztin, B. Barz, and L. Janosi, *J. Chem. Phys.*, **124**(6), 064106 (2006). Calculating potentials of mean force and diffusion coefficients from nonequilibrium processes without Jarzynski’s equality.
- [53] M. Forney, L. Janosi, and I. Kosztin, manuscript in preparation (2007). Reconstruction of free energy profiles from fast non-equilibrium processes.
- [54] B. Roux, *Curr. Opin. Struct. Biol.*, **12**(2), 182–189 (2002). Theoretical and computational models of ion channels.
- [55] A. R. Leach, *Molecular Modelling: Principles and Applications (2nd edition)*, PrenticeHall, Upper Saddle River, New Jersey, 2001.
- [56] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, New York, 2001.
- [57] D. Frenkel and B. Smit, *Understanding Molecular Simulation from Algorithms to Applications*, Academic Press, California, 2002.
- [58] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, New York, 1987.
- [59] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications*, Springer-Verlag Telos, 3rd ed., 1996.

- [60] T. Simonson, G. Archontis, and M. Karplus, *Accounts Chem. Res.*, **35**(6), 430–437 (2002). Free energy simulations come of age: Protein-ligand recognition.
- [61] B. Roux, *Comp. Phys. Commun.*, **91**(1-3), 275–282 (1995). The calculation of the potential of mean force using computer-simulations.
- [62] Torrie and Valleau, *J. Comp. Phys.*, **23**, 187–199 (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling.
- [63] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten, *J. Mol. Graphics*, **19**(1), 13–25 (2001). Steered molecular dynamics investigations of protein function.
- [64] S. Park and K. Schulten, *J. Chem. Phys.*, **120**(13), 5946–5961 (2004). Calculating potentials of mean force from steered molecular dynamics simulations.
- [65] C. Jarzynski, *Phys. Rev. Lett.*, **78**(14), 2690–2693 (1997). Nonequilibrium equality for free energy differences.
- [66] C. Jarzynski, *Phys. Rev. E*, **56**(5), 5018–5035 (1997). Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach.
- [67] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante, *Science*, **296**(5574), 1832–1835 (2002). Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski’s equality.
- [68] F. Douarche, S. Ciliberto, A. Petrosyan, and I. Rabbiosi, *Europhys. Lett.*, **70**(5), 593–599 (2005). An experimental test of the Jarzynski equality in a mechanical experiment.
- [69] C. Bustamante, J. Liphardt, and F. Ritort, *Physics Today*, **58**(7), 43–48 (2005). The nonequilibrium thermodynamics of small systems.
- [70] J. Gore, F. Ritort, and C. Bustamante, *Proc. Natl. Acad. Sci. U.S.A.*, **100**(22), 12564–12569 (2003). Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements.
- [71] F. Ritort, C. Bustamante, and I. Tinoco, *Proc. Natl. Acad. Sci. U.S.A.*, **99**(21), 13544–13548 (2002). A two-state kinetic model for the unfolding of single molecules by mechanical force.
- [72] D. A. Hendrix and C. Jarzynski, *J. Chem. Phys.*, **114**(14), 5974–5981 (2001). A ”fast growth” method of computing free energy differences.
- [73] G. Hummer, *J. Chem. Phys.*, **114**(17), 7330–7337 (2001). Fast-growth thermodynamic integration: Error and efficiency analysis.



- [74] G. Hummer and A. Szabo, *Proc. Natl. Acad. Sci. U.S.A.*, **98**(7), 3658–3661 (2001). Free energy reconstruction from nonequilibrium single-molecule pulling experiments.
- [75] G. Hummer and A. Szabo, *Accounts Chem. Res.*, **38**(7), 504–513 (2005). Free energy surfaces from single-molecule force spectroscopy.
- [76] E. Atilgan and S. X. Sun, *J. Chem. Phys.*, **121**(21), 10392–10400 (2004). Equilibrium free energy estimates based on nonequilibrium work relations and extended dynamics.
- [77] S. X. Sun, *J. Chem. Phys.*, **118**(13), 5769–5775 (2003). Equilibrium free energies from path sampling of nonequilibrium trajectories.
- [78] D. M. Zuckerman and T. B. Woolf, *Phys. Rev. Lett.*, **89**(18), 180602 (2002). Theory of a systematic computational error in free energy differences.
- [79] D. M. Zuckerman and T. B. Woolf, *Chem. Phys. Lett.*, **351**(5-6), 445–453 (2002). Overcoming finite-sampling errors in fast-switching free-energy estimates: extrapolative analysis of a molecular system.
- [80] D. J. Evans and D. J. Searles, *Adv. Phys.*, **51**(7), 1529–1585 (2002). The fluctuation theorem.
- [81] G. M. Wang, E. M. Sevcik, E. Mittag, D. J. Searles, and D. J. Evans, *Phys. Rev. Lett.*, **89**(5), 050601 (2002). Experimental demonstration of violations of the second law of thermodynamics for small systems and short time scales.
- [82] P. Vidossich, M. Cascella, and P. Carloni, *Proteins*, **55**(4), 924–931 (2004). Dynamics and energetics of water permeation through the aquaporin channel.
- [83] G. E. Crooks, *Phys. Rev. E*, **61**(3), 2361–2366 (2000). Path-ensemble averages in systems driven far from equilibrium.
- [84] S. Park, F. Khalili-araghi, E. Tajkhorshid, and K. Schulten, *J. Chem. Phys.*, **119**(6), 3559–3566 (2003). Free energy calculation from steered molecular dynamics simulations using Jarzynski’s equality.
- [85] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, *J. Comp. Chem.*, **13**(8), 1011–1021 (1992). The weighted histogram analysis method for free-energy calculations on biomolecules .1. the method.
- [86] M. Cascella, L. Guidoni, A. Maritan, U. Rothlisberger, and P. Carloni, *J. Phys. Chem. B*, **106**(50), 13027–13032 (2002). Multiple steering molecular dynamics applied to water exchange at alkali ions.
- [87] R. Amaro, E. Tajkhorshid, and Z. Luthey-Schulten, *Proc. Natl. Acad. Sci. U.S.A.*, **100**(13), 7599–7604 (2003). Developing an energy landscape for the novel function of a (beta/alpha)<sub>8</sub> barrel: Ammonia conduction through hisf.

- [88] R. Amaro and Z. Luthey-Schulten, *Chem. Phys.*, **307**(2-3), 147–155 (2004). Molecular dynamics simulations of substrate channeling through an [alpha]-[beta] barrel protein.
- [89] C. H. Bennett, *J. Comp. Phys.*, **22**, 245–268 (1976). Efficient estimation of free-energy differences from Monte-Carlo data.
- [90] T. Rodinger and R. Pomes, *Curr. Opin. Struct. Biol.*, **15**(2), 164–170 (2005). Enhancing the accuracy, the efficiency and the scope of free energy simulations.
- [91] N. Lu, J. K. Singh, and D. A. Kofke, *J. Chem. Phys.*, **118**(7), 2977–2984 (2003). Appropriate methods to combine forward and reverse free-energy perturbation averages.
- [92] N. D. Lu and T. B. Woolf, *Mol. Phys.*, **102**(2), 173–181 (2004). Overlap perturbation methods for computing alchemical free energy changes: variants, generalizations and evaluations.
- [93] N. D. Lu, D. Wu, T. B. Woolf, and D. A. Kofke, *Phys. Rev. E*, **69**(5), 05772 (2004). Using overlap and funnel sampling to obtain accurate free energies from nonequilibrium work measurements.
- [94] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, *Phys. Rev. Lett.*, **91**(14), 140601 (2003). Equilibrium free energies from nonequilibrium measurements using maximum-likelihood methods.
- [95] M. R. Shirts and V. S. Pande, *J. Chem. Phys.*, **122**(14), 144107 (2005). Comparison of efficiency and bias of free energies computed by exponential averaging, the bennett acceptance ratio, and thermodynamic integration.
- [96] W. Humphrey, A. Dalke, and K. Schulten, *J. Mol. Graphics*, **14**, 33–38 (1996). VMD – Visual Molecular Dynamics.
- [97] P. Procacci, S. Marsili, A. Barducci, G. F. Signorini, and R. Chelli, *J. Chem. Phys.*, **125**(16), 164101 (2006). Crooks equation for steered molecular dynamics using a Nose-Hoover thermostat.
- [98] L. Janosi, J. Das, and I. Kosztin, manuscript in preparation (2007). Energetics and dynamics of glycerol and water transport in the aquaporin channel GlpF.
- [99] G. M. Preston, T. P. Carroll, W. B. Guggino, and P. Agre, *Science*, **256**(5055), 385–387 (1992). Appearance of water channels in xenopus oocytes expressing red-cell Chip28 protein.
- [100] K. B. Heller, E. C. Lin, and T. H. Wilson, *J. Bacteriol.*, **144**, 274–278 (1980). Substrate specificity and transport properties of the glycerol facilitator of escherichia-coli.

- [101] M. J. Borgnia and P. Agre, *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 2888–2893 (2001). Reconstitution and functional comparison of purified GlpF and AqpZ, the glycerol and water channels from *escherichiacoli*.
- [102] M. Borgnia, S. Nielsen, A. Engel, and P. Agre, *Ann. Rev. Biochem.*, **68**, 425–458 (1999). Cellular and molecular biology of the aquaporin water channels.
- [103] D. Fu, A. Libson, L. J. Miercke, C. Weitzman, P. Nollert, J. Krucinski, and R. M. Stroud, *Science*, **290**(5491), 481–486 (2000). Structure of a glycerol-conducting channel and the basis for its selectivity.
- [104] M. Ø. Jensen, E. Tajkhorshid, and K. Schulten, *Structure*, **9**, 1083–1093 (2001). The mechanism of glycerol conduction in aquaglyceroporins.
- [105] F. Tian and T. A. Cross, *J. Mol. Biol.*, **285**(5), 1993–2003 (1999). Cation transport: An example of structural based selectivity.
- [106] P. C. Jordan, *Biophys. J.*, **58**(5), 1133–1156 (1990). Ion-water and ion-polypeptide correlations in a gramicidin-like channel - a molecular-dynamics study.
- [107] B. Roux, S. Berneche, and W. Im, *Biochemistry*, **39**(44), 13295–13306 (2000). Ion channels, permeation, and electrostatics: Insight into the function of KcsA.
- [108] S. Kuyucak, O. S. Andersen, and S. H. Chung, *Rep. Prog. Phys.*, **64**(11), 1427–1472 (2001). Models of permeation in ion channels.
- [109] A. Finkelstein and O. S. Andersen, *J. Membr. Biol.*, **59**(3), 155–171 (1981). The gramicidin A channel: a review of its permeability characteristics with special reference to the single-file aspect of transport.
- [110] T. W. Allen, O. S. Andersen, and B. Roux, *Proc. Natl. Acad. Sci. U.S.A.*, **101**(1), 117–122 (2004). Energetics of ion conduction through the gramicidin channel.
- [111] A. B. Mamonov, R. D. Coalson, A. Nitzan, and M. G. Kurnikova, *Biophys. J.*, **84**(6), 3646–3661 (2003). The role of the dielectric barrier in narrow biological channels: A novel composite approach to modeling single-channel currents.
- [112] L. E. Townsley, W. A. Tucker, S. Sham, and J. F. Hinton, *Biochemistry*, **40**(39), 11676–11686 (2001). Structures of gramicidins A, B, and C incorporated into sodium dodecyl sulfate micelles.
- [113] B. Corry and S. H. Chung, *Eur. Biophys. J.*, **34**(3), 208–216 (2005). Influence of protein flexibility on the electrostatic energy landscape in gramicidin A.
- [114] T. W. Allen, O. S. Andersen, and B. Roux, *Biophys. Chem.*, **124**(3), 251–267 (2006). Molecular dynamics - potential of mean force calculations as a tool for understanding ion permeation and selectivity in narrow channels.

- [115] R. Coalson and M. Kurnikova, *Nanobiosci., IEEE Transact. on*, **4**(1), 81–93 (2005). Poisson-Nernst-Planck theory approach to the calculation of current through biological ion channels.
- [116] M. J. Borgnia, D. Kozono, G. Calamita, P. C. Maloney, and P. Agre, *J. Mol. Biol.*, **291**(5), 1169–1179 (1999). Functional reconstitution and characterization of AqpZ, the E-coli water channel protein.
- [117] P. Pohl, S. M. Saparov, M. J. Borgnia, and P. Agre, *Proc. Natl. Acad. Sci. U.S.A.*, **98**(17), 9624–9629 (2001). Highly selective water channel activity measured by voltage clamp: Analysis of planar lipid bilayers reconstituted with purified AqpZ.
- [118] S. M. Saparov, S. P. Tsunoda, and P. Pohl, *Biol. Cell*, **97**(7), 545–550 (2005). Proton exclusion by an aquaglyceroprotein: a voltage clamp study.
- [119] F. Q. Zhu, E. Tajkhorshid, and K. Schulten, *Biophys. J.*, **83**(1), 154–160 (2002). Pressure-induced water transport in membrane channels studied by molecular dynamics.
- [120] F. Q. Zhu, E. Tajkhorshid, and K. Schulten, *Biophys. J.*, **86**(1), 50–57 (2004). Theory and simulation of water permeation in aquaporin-1.
- [121] M. Ø. Jensen and O. G. Mouritsen, *Biophys. J.*, **90**(7), 2270–2284 (2006). Single-channel water permeabilities of Escherichia coli aquaporins AqpZ and GlpF.
- [122] I. Kosztin and K. Schulten, *Phys. Rev. Lett.*, **93**(23), 238102 (2004). Fluctuation-driven molecular transport through an asymmetric membrane channel.
- [123] S. E. Feller and A. D. MacKerell Jr., *J. Phys. Chem. B*, **104**(31), 7510–7515 (2000). An improved empirical potential energy function for molecular simulations of phospholipids.
- [124] E. Tajkhorshid, P. Nollert, M. Ø. Jensen, L. J. W. Miercke, J. O’Connell, R. M. Stroud, and K. Schulten, *Science*, **296**, 525–530 (2002). Control of the selectivity of the aquaporin water channel family by global orientational tuning.
- [125] M. Ø. Jensen, E. Tajkhorshid, and K. Schulten, *Biophys. J.*, **85**(5), 2884–2899 (2003). Electrostatic tuning of permeation and selectivity in aquaporin water channels.
- [126] D. R. Lide, *CRC Handbook of Chemistry and Physics*, CRC Press, Boca Raton, Florida, 75th ed., 1994.
- [127] A. Szabo, S. K., and S. Z., *J. Chem. Phys.*, **72**(8), 4350–4357 (1980). First passage time approach to diffusion controlled reactions.
- [128] L. Janosi, H. Keer, I. Kosztin, and T. Ritz, *Chem. Phys.*, **323**(1), 117–128 (2006). Influence of subunit structure on the oligomerization state of light-harvesting complexes: A free energy calculation study.

- [129] L. Janosi, H. Keer, T. Ritz, and I. Kosztin, manuscript in preparation (2007). In silico prediction of LH2 ring size from the crystal structure of a single subunit.
- [130] R. Cogdell, A. Gardiner, A. Roszak, C. J. Law, J. Southall, and N. W. Isaacs, *Photosynth. Res.*, **81**, 207–214 (2004). Rings, ellipses and horseshoes: How purple bacteria harvest solar energy.
- [131] S. Scheuring, J. Sturgis, V. Prima, A. Bernadac, D. Lévy, and J. Rigaud, *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 11293–11297 (2004). Watching the photosynthetic apparatus in native membranes.
- [132] S. Bahatyrova, R. Frese, C. Siebert, J. Olsen, K. van der Werf, R. van Grondelle, R. Niederman, P. Bullough, C. Otto, and C. Hunter, *Nature*, **430**, 1058–1062 (2004). The native architecture of a photosynthetic membrane.
- [133] G. Fleming and R. van Grondelle, *Curr. Opin. Struct. Biol.*, **7**, 738–748 (1997). Femtosecond spectroscopy of photosynthetic light-harvesting systems.
- [134] T. Ritz and K. Schulten, *Physikalische Blätter*, **57**, 49–53 (2001). Physik der photosynthese.
- [135] T. Ritz, A. Damjanovic, and K. Schulten, *ChemPhysChem*, **3**(3), 243–248 (2002). The quantum physics of photosynthesis.
- [136] M. Papiz, S. Prince, T. Howard, R. Cogdell, and N. Isaacs, *J. Mol. Biol.*, **326**, 1523–1538 (2003). The structure and thermal motion of the B800-850 LH2 complex from *Rps. acidophila* at 2.0 Å resolution and 100 K: new structural features and functionally relevant motions.
- [137] G. Montoya, M. Cyrklaff, and I. Sinning, *J. Mol. Biol.*, **250**, 1–10 (1995). Two-dimensional crystallization and preliminary structure analysis of light harvesting II (B800-850) complex from the purple bacterium *Rhodovulum sulfidophilum*.
- [138] T. Walz, S. Jamieson, C. Bowers, B. P.A., and C. Hunter, *J. Mol. Biol.*, **282**, 833–845 (1998). Projection structures of three photosynthetic complexes from *Rhodobacter sphaeroides*: LH2 at 6 Å, LH1 and RC-LH1 at 25 Å.
- [139] S. Scheuring, J. Seguin, S. Marco, D. Levy, B. Robert, and J. Rigaud, *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 1690–1693 (2003). Nanodissection and high-resolution imaging of the *Rhodopseudomonas viridis* photosynthetic core complex in native membranes by AFM.
- [140] J. Ranck, T. Ruiz, G. Pehau-Arnaudet, B. Arnoux, and F. Reiss-Husson, *Biochim. Biophys. Acta*, **1506**, 67–78 (2001). Two-dimensional structure of the native light-harvesting complex LH2 from *Rubrivivax gelatinosus* and of a truncated form.
- [141] S. Scheuring, F. Reiss-Husson, A. Engel, J.-L. Rigaud, and J.-L. Ranck, *EMBO J.*, **20**, 3029–3035 (2001). High resolution topographs of the *Rubrivivax gelatinosus* light-harvesting complex 2.

- [142] N. Hartigan, H. A. Tharia, F. Sweeney, A. M. Lawless, and M. Z. Papiz, *Biophys. J.*, **82**(2), 963–977 (2002). The 7.5-Å electron density and spectroscopic properties of a novel low-light B800 LH2 from *Rhodospseudomonas palustris*.
- [143] S. Scheuring, J. Rigaud, and J. Sturgis, *EMBO J.*, **23**, 4127–4133 (2004). Variable LH2 stoichiometry and core clustering in native membranes of *Rhodospirillum photometricum*.
- [144] M. Koblizek, J. Shih, S. Breitbart, E. Ratcliffe, Z. Kolber, N. Hunter, and R. Niederman, *Biochim. Biophys. Acta*, **1706**, 220–231 (2005). Sequential assembly of photosynthetic units in *Rhodobacter sphaeroides* as revealed by fast repetition rate analysis of variable bacteriochlorophyll *a* fluorescence.
- [145] P. Parkes-Loach, J. Sprinkle, and P. Loach, *Biochemistry*, **27**, 2718–2727 (1988). Reconstitution of the B873 light-harvesting complex of *Rhodospirillum rubrum* from the separately isolated  $\alpha$  and  $\beta$  polypeptides and bacteriochlorophyll *a*.
- [146] J. Todd, P. Parkes-Loach, J. Leykam, and P. Loach, *Biochemistry*, **37**, 17458–17468 (1998). In vitro reconstitution of the core and peripheral Light-Harvesting complexes of *Rhodospirillum rubrum* from separately isolated components.
- [147] J. Todd, P. Recchia, P. Parkes-Loach, J. Olsen, G. Fowler, P. McGlynn, C. Hunter, and P. Loach, *Photosynth. Res.*, **62**, 85–98 (1999). Minimal requirements for in vitro reconstitution of the structural subunit of light-harvesting complexes of photosynthetic bacteria.
- [148] J. Kehoe, K. Meadows, P. Parkes-Loach, and P. Loach, *Biochemistry*, **37**, 3418–3428 (1998). Reconstitution of core light-harvesting complexes of photosynthetic bacteria using chemically synthesized polypeptides. 2. determination of structural features that stabilize complex formation and their implications for the structure of the subunit complex.
- [149] K. Meadows, P. Parkes-Loach, J. Kehoe, and P. Loach, *Biochemistry*, **37**, 3411–3417 (1998). Reconstitution of core light-harvesting complexes of photosynthetic bacteria using chemically synthesized polypeptides. 1. minimal requirements for subunit formation.
- [150] C. Davis, P. Bustamante, J. Todd, P. Parkes-Loach, P. McGlynn, J. Olsen, L. McMaster, and P. Hunter, C.N. and Loach, *Biochemistry*, **12**, 3671–3679 (1997). Evaluation of structure-function relationships in the core light-harvesting complex of photosynthetic bacteria by reconstitution with mutant polypeptides.
- [151] P. Parkes-Loach, A. Majeed, C. Law, and P. Loach, *Biochemistry*, **43**, 7003–7016 (2004). Interactions stabilizing the structure of the core light-harvesting complex (LH1) of photosynthetic bacteria and its subunit (B820).

- [152] J. D. Olsen, B. Robert, A. Siebert, P. A. Bullough, and C. N. Hunter, *Biochemistry*, **42**(51), 15114–15123 (2003). Role of the c-terminal extrinsic region of the alpha polypeptide of the light-harvesting 2 complex of *Rhodobacter sphaeroides*: A domain swap study.
- [153] P. Braun, J. Olsen, B. Strohmman, C. Hunter, and H. Scheer, *J. Mol. Biol.*, **318**, 1085–1095 (2002). Assembly of Light-Harvesting bacteriochlorophyll in a model transmembrane helix in its natural environment.
- [154] S. M. Prince, M. Z. Papiz, A. A. Freer, G. McDermott, A. M. Hawthornthwaite-Lawless, R. J. Cogdell, and N. W. Isaacs, *J. Mol. Biol.*, **268**(2), 412–423 (1997). Apoprotein structure in the LH2 complex from *Rhodospseudomonas acidophila* strain 10050: modular assembly and protein pigment interactions.
- [155] C. D. Schwieters, J. J. Kuszewski, T. N. and G. M. Clore, *J. Magn. Res.*, **160**, 66–74 (2003). The XPLOR-NIH NMR molecular structure determination package.
- [156] L. Kale, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan, and K. Schulten, *J. Comp. Phys.*, **151**, 283–312 (1999). NAMD2: Greater scalability for parallel molecular dynamics.
- [157] A. Pandit, R. Vischers, I. van Stokkum, R. Kraayenhof, and R. van Grondelle, *Biochemistry*, **40**, 12913–12924 (2001). Oligomerization of Light-Harvesting I antenna peptides of *Rhodospirillum rubrum*.
- [158] A. Végh and B. Robert, *FEBS Lett.*, **528**, 222–226 (2002). Spectroscopic characterisation of a tetrameric subunit form of the core antenna protein from *Rhodospirillum rubrum*.

# Vita

Lorant Janosi was born on February 22, 1979 in Gheorgheni, Romania. He graduated third in his class from the Physics Department of “Babes-Bolyai” University in Cluj-Napoca, Romania, in July 2001. He earned his Masters Degree in Solid State Physics from the “Babes-Bolyai” University in July 2002, and his Ph.D. degree in Physics from the University of Missouri-Columbia in December, 2007.