Free energy calculations of protein-ligand complexes with computational molecular dynamics

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Georg-August-Universität zu Göttingen

Göttingen 2008

vorgelegt von

Maik Götte

aus Warburg

D 7
Referent: Prof. Dr. Ralf Ficner
Koreferent: Prof. Dr. Reinhard Jahn
Tag der mündlichen Prüfung (Rigorosum): 29./30.10.2008

Vorveröffentlichungen der Dissertation

Teilergebnisse dieser Arbeit wurden im folgenden Beitrag vorab veröffentlicht:

Publikationen

Goette, M. and Grubmüller, H. Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes *Journal of Computational Chemistry*, 2008, in press

Contents

1

Introduction

Evolution is the main driving force for the survival of organisms in changing environments.[1] The information about adaptation of organisms to these environments is stored in the DNA of their genome and is inherited to their offspring.

The human genome project[2,3] as well as further projects to resolve the genome of organisms like *drosophila melanogaster* or *caenorhabditis elegans* led to an abundant amount of information about the amino acid sequences of proteins in these organisms. This information about genetic sequences, and the implicit sequence of amino acids, is invaluable for the modern taxonomic identification.[4,5] Thus, such *DNA bar coding* complements the classical way of determining the taxonomy of organisms by their phenotype with a close inspection of also the genotype.

However, the amino acid sequence of proteins yields only limited information about the structure and function of biological macromolecules. Processes such as the catalytic function of enzymes, gating mechanisms in ion channels, the assembly of virus capsids or collagen fibres, and signal transduction or immune response pathways cannot be explained by these sequences.

To gain insight into such processes, the three-dimensional structure of a folded amino acid sequence, a functional protein, is necessary. The structure determination of proteins or nucleic acids is based on experimental techniques such as X-ray crystallography[6,7] or nuclear magnetic resonance (NMR),[8,9] which have made remarkable progress in solving high-resolution structures over the past years. Such structures are archived and accessible via the RSCB Protein Data Bank.[10] Furthermore, theoretical approaches for structure prediction from amino acid sequences are under constant development and have been monitored by CASP (Critical Assessment of Structure Prediction) over the last ten years.[11]

A particular snapshot of a protein structure is a point in configuration space. Accordingly, a fluctuating structure of a protein is described by a region on a high-dimensional complex free energy landscape, which is dynamically explored.[12] Such protein structure dynamics is the key to obtain insight into the function of biomolecules. A structure derived by classical X-ray crystallography is just an average and thus lacks dynamics. Moreover,

the free energy minimum of the structure derived with the help of such a protein crystal is governed by the crystallization conditions and possibly changed by the typically low temperature usage. However, experimental methods to probe the structural dynamics with atomic detail are available, even though with intrinsic limitations. Time-resolved X-ray crystallography[13,14] reveals conformational protein motions on the picosecond time scale, but wide-spread use is impeded due to the massive experimental effort involved. NMR relaxation measurements[15,16,17,18] have been used to probe protein dynamics on pico- to nanosecond and micro- to millisecond time-scales, and thus fail to reveal dynamics on the nano- to microsecond time scale.

Such time-resolved dynamics is needed to understand molecular recognition mechanisms which enable specific binding of drugs to target proteins, or binding of polymerases to specific DNA-sequences inducing the first step in protein synthesis. Moreover, as has been shown in a recent study of ubiquitin with a combination of residual dipolar coupling (RDC) NMR and Molecular Dynamics (MD) simulations,[19] conformational changes in proteins, and, in this context, the determination between induced fit and conformational selection mechanisms can be observed on the nano- to microsecond time scale.

All aspects of the previously described protein dynamics can be understood as the exploration of the high-dimensional complex free energy landscape, intrinsically accessible to the protein. Thus, the free energy, like the directly related entropy, is an ensemble property.[20] The biochemical processes in living organisms are driven by free energy gradients on such a free energy landscape. The accurate calculation of the free energies is therefore essential for understanding biomolecular functions like transport processes in cells. In contrast to the enthalpy which can be computed by the systems total energies from a MD simulation of arbitrary length, the calculation of the free energy and entropy requires a fully explored configuration (phase) space. For larger biomolecules, the computationally accessible time scales are limited and often insufficient to sample this complete phase space with MD simulations. Therefore, in these cases, the direct calculation of free energies and entropies from such ensembles is nearly impossible.

The accuracy in such free energy calculations from atomic-level simulations is often limited for larger systems such as proteins due to the sampling problems, mentioned before, and has been intensively reviewed.[21,22,23,24] Despite the considerable amount of studies for the different available methods, it is difficult to judge how these methods perform for larger and more complex systems such as biological macromolecules. A systematic comparison of the available methods with systems of different complexity will therefore help to develop a more accurate method for calculating free energy differences if larger biomolecules are involved.

Experimental techniques, such as fluorescence spectroscopy[25] or surface plasmon resonance spectroscopy[26,27] are used to obtain binding affinities of receptor/ligand complexes via the equilibrium dissociation constant (K_D) which is directly related to the absolute free energy difference of the binding process. However, measuring these binding affinities can be limited by experimental conditions, especially if large ligand concentrations are needed.

Computational approaches to assess the binding affinity of receptor/ligand complexes are thus very helpful when experimental measurements are brought to their limits, and therefore aim to have a predictive function to assess such binding affinities. They can be roughly divided into two classes, docking and free energy calculations from MD simulations.

In the docking approach, a binding site is defined in a, typically, rigid protein. A ligand is then fitted into this binding site by flexible rotation of functional groups within the ligand, and rotation of the ligand itself. Electrostatic and van-der-Waals interactions are calculated for the different conformations and a scoring function evaluates the ligand conformations, which energetically fit best. Several applications with different scoring functions, such as Autodock,[28] FlexX,[29] or Gold[30] have been developed over the past years. It is common to these applications, that an implicit solvent environment is used and the flexibility of the protein is often neglected. Despite the advantage to screen large libraries of ligands in a short amount of time, the different scoring functions often lead to different, inconsistent, results, especially in cases where water molecules in the binding pocket are important.[31,32]

The class of free energy calculations from MD simulations can further be separated into

three general fields: linear interaction energy (LIE), MM/PBSA, and thermodynamic perturbation. The LIE method[33] is based on linear response theory.[34,35] The binding free energy difference is approximated by the interaction energy between the ligand and the protein, derived from equilibrium MD simulations whereas these interaction energies are modified by three constants. These constants have been derived from experimental observations. Despite the reasonable free energy estimates, derived by this method, it still neglects the transfer free energy between different solvents (for details see Ref. 36). The main difference between LIE and the MM/PBSA method[37,38] is that in the latter, the solvent is treated implicitly and the electrostatic components are obtained from a dielectric continuum model with a dielectric constant for the solute and the solvent. Despite the speedup in simulation time due to the missing explicit solvent electrostatics, this method strongly depends on the dielectric constants which differ dramatically for most of the solutes, and, additionally, are hard to obtain.

In contrast to the LIE and MM/PBSA methods, thermodynamic perturbation theory aims to compute the free energy difference between two states instead of absolute free energies, but without approximations such as continuum models or the need for empirical observations. The two basic ideas in thermodynamic perturbation theory are thermodynamic integration[39] and free energy perturbation.[40] Over the years, several approaches to compute free energies have been developed, assessed, and applied, often focusing on smaller systems.[41,42,43,44,45,46,47,48,49,50] For larger systems such as proteins, the sampling problem becomes increasingly severe, mainly due to insufficient overlap of the involved thermodynamic phase space densities. This problem also complicates the reliable assessment of accuracy and convergence, simply due to the difficulty of obtaining a reliable reference value. All methods, relying on thermodynamic perturbation theory, are brought to their limits, when precise free energy differences with precise error estimates for large systems and perturbations, such as amino acid sidechains, are addressed. These methods as well as their limitations are briefly introduced in Chapter 2.3. However, the recently developed non-equilibrium methods to compute free energy differences offer the chance to develop

improved methods. The calculation of binding free energies is thus still a field of active research.

One aim of this thesis was to carry out a comprehensive evaluation of the available methods to compute free energy differences from all atom simulations for three test systems at different levels of complexity, namely the interconversion of ethane to methanol, of tryptophane to glycine in a tripeptide and of m_3GpppG to m^7GpppG in the globular protein snurportin 1. Additionally, an improved method to compute free energy differences from non-equilibrium simulations was developed and tested with the systems mentioned above. This newly developed method was then used to compute the binding free energy differences between the wildtype and eight point mutations in the hemagglutinin peptide HA 307-319 bound towards the major histocompatibility complex (MHC) Class II receptor, which is one of the key elements in immune response.

In contrast to the broadband affinity of the MHC proteins to peptides, the transport of RNA in and out of the nucleus is regulated by very specific recognition mechanisms, involving methyl groups as the only modification. A second aim of this thesis was to address the binding specificity of snurportin 1 to the different RNA-cap methylization states. It is still not clear how this specificity is obtained on the molecular level and why the hypermethylization leads to a higher affinity in the transport protein snurportin 1. Apart from the calculations of binding free energy differences of the two involved snRNA caps, the structural mechanisms involved in the function of snurportin 1 when bound to either the methylated or the hypermethylated cap as well as without a ligand can be addressed and answered by MD simulations.

This thesis is organized as follows. After a short sketch of the available methods used in this thesis (Chapter 2), a method to compute free energy differences from non-equilibrium simulations is described and extensively tested (Chapter 3). In Chapter 4, the free energy differences of eight mutants of the influenza viral peptide hemagglutinin HA 307-319 bound to a major histocompatibility complex (MHC) Class II protein are calculated with the newly developed method and are compared to experimentally derived binding affinities.

Following the free energy calculations of snurportin 1 in Chapter 3, the structural changes of the globular protein snurportin 1 upon ligand removal are determined and analyzed. Moreover, the contribution of the solvation shell of the hypermethylated RNA-cap to the binding affinity to snurportin 1 is studied (Chapter 5). The biological background and the questions addressed are described in detail in the introductory sections of the respective chapters.

2

Theory and Methods

This chapter outlines the general theoretical framework of this thesis and the common methods applied. Details of the particular simulation setup and the employed free energy calculations are given in chapters 3 to 5, which address the different biological systems these free energy calculations have been applied to.

2.1 Molecular Dynamics Simulations

Biolmolecular systems usually consist of several tens to thousands of amino acid residues concerning proteins, or up to millions of base pairs concering DNA. Over ten times more atoms are involved in such systems, and they span dynamics from bond vibrations at the femtosecond timescale up to large-scale conformational changes occuring within microseconds or on even longer time-scales.

From the established atomistic simulation methods, Molecular Dynamics (MD) is the only one that is able to handle the high complexity of protein structures and is able to assess relevant time scales up to microseconds on modern parallel computers.

MD simulations describe the time evolution of a molecular system, e.g., a protein, by numerically solving Newton's equations of motion for all atoms in the system. Such simulations can accurately describe the dynamics of biological relevant systems by using three approximations; (a) the Born-Oppenheimer approximation, where nuclear and electronic motions are decoupled, (b) the approximation that nuclei can be treated as classical particles, and (c) the use of an empirical force field to describe the interaction between particles. These approximations are described below.

The Born-Oppenheimer Approximation

The dynamical evolution of any system is described by the time-dependent Schrödinger equation,

$$i\hbar\frac{\delta\psi}{\delta t} = \mathcal{H}\psi, \tag{2.1}$$

where $\mathcal{H}$ denotes the Hamiltonian, i.e., the sum of potential and kinetic energy operator, ψ the wave function, and $\hbar = h/2\pi$ with h as Planck's constant. The wave function ψ is a function of the coordinates of all nuclei and all electrons. In the Born-Oppenheimer approximation[51] the fast degrees of freedom are separated from the slow ones. Due to the high mass of nuclei compared to that of the electrons, the former move much slower than the latter. This leads to the approximation that electrons can be considered to move in the field of fixed nuclei, i.e. that the dynamics of the electrons instantaneously adjust to the much slower nuclei. Accordingly, the electronic wave function ψ_e depends only parametrically on the nuclear coordinates and the total wave function ψ_{tot} , which hence can be separated into a nuclear (ψ_n) and an electronic wave function:

$$\psi_{tot}(\mathbf{r}, \mathbf{R}) = \psi_n(\mathbf{R})\psi_e(\mathbf{r}; \mathbf{R}). \tag{2.2}$$

Here, $\mathbf{R} = (\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N)$ denotes the coordinates of the N nuclei and $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M)$ the coordinates of the M electrons, respectively. This approximation yields a separation of Eq. 2.2 into a time-dependent Schrödinger equation for the motion of the nuclei and a time-independent Schrödinger equation for the electronic dynamics. This potential is called a potential energy surface (PES) because it is a function of the nucleic position. Within the Born-Oppenheimer approximation, the nuclei move on this PES obtained by solving the time-independent electronic problem,

$$\mathcal{H}_e\psi_e = E_e\psi_e, \tag{2.3}$$

with $\mathcal{H}_e$ as the electronic Hamiltonian and E_e the lowest energy eigenvalue, which therefore parametrically depends on the nuclear positions $\mathbf{R}$.

Classical Dynamics

The huge size of biomolecules like proteins or DNA renders the solution of the time-dependent Schrödinger equation (Eq. 2.1) for the nuclear motion still prohibitively expensive. Therefore, as an additional approximation in MD simulations, the nuclear dynamics is described classically and obeys Newton's equations of motion

$$-\nabla_i V(\mathbf{R}) = m_i \frac{d^2\mathbf{R}_i(t)}{dt^2}, \tag{2.4}$$

where $\mathbf{F}_i = -\nabla_i V(\mathbf{R})$ is the force onto atom i as a function of all atomic coordinates, and $\mathbf{R}_i$ and m_i are the coordinates and mass of atom i, respectively. The acceleration $d^2\mathbf{R}_i(t)/dt^2$ leads to a change in the atoms velocity and position within a discrete time step Δt, chosen sufficiently short such as to capture the fastest motions in the system. These motions are usually the bond and angle vibrations. Especially bond vibrations involving the light hydrogen atoms occur at the femtosecond timescale and therefore restrict the timestep to about 1 fs. To extend the timestep length beyond 1 fs, a number of algorithms to constrain bond lengths and therefore remove these fast degrees of freedom, e.g., SHAKE[52] and LINCS[53] have been developed. The latter was used for all underlying simulations in this thesis. For an efficient numerical integration of Newtons's equation of motion, the leap-frog algorithm[54] is applied,

$$\mathbf{R}(t + \Delta t) = \mathbf{R}(t) + \mathbf{v}\left(t + \frac{1}{2}\Delta t\right)\Delta t \tag{2.5}$$

$$\mathbf{v}\left(t + \frac{1}{2}\Delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\Delta t\right) + \frac{\mathbf{F}_i(t)}{m_i}\Delta t, \tag{2.6}$$

where the position $\mathbf{R}(t)$ and force $\mathbf{F}_i(t)$ are calculated together with the velocities $\mathbf{v}(t - \Delta t/2)$.

Force Fields

Although the Born-Oppenheimer approximation allows the treatment of an electronic wave function as a function of the nuclear coordinates, the evaluation of the potential $V(\mathbf{R})$ by solving the time-independent Schrödinger equation (Eq. 2.3) for the electrons is still required. This, however, is currently still too expensive for the large number of electrons in biomolecular systems, rendering their extended quantum-mechanical MD simulation unfeasible.

Therefore, as a third and last approximation, the potential energy of the system as a function of the nuclear coordinates is expressed as the sum of simple analytical functions. These functions in combination with a corresponding set of empirical parameters, compose the molecular mechanical (MM) force field,

$$\begin{aligned} V(\mathbf{R}) = & \sum_{\text{bonds } i} \frac{1}{2} k_i (l_i - l_i^0)^2 \\ & + \sum_{\text{angles } i} \frac{1}{2} k_i (\theta_i - \theta_i^0)^2 \\ & + \sum_{\text{impropers } i} \frac{1}{2} k_i (\xi_i - \xi_i^0)^2 \\ & + \sum_{\text{dihedrals } i} \frac{1}{2} V_i (1 + \cos(n\phi_i - \phi_i^0)) \\ & + \sum_{\substack{\text{pairs } ij \\ i \neq j}} (V_{\text{LJ}}^{ij} + V_{\text{Coul}}^{ij}), \end{aligned} \tag{2.7}$$

where l_i, θ_i, and ξ_i are the individual bond lengths, angles, and out-of-plane-angles, and l_i^0, θ_i^0, and ξ_i^0 their equilibrium values, respectively. The k_i denote the respective force constants. V_i states the barrier height, n the multiplicity, ϕ_i the plane-angle and ϕ_i^0 the phase of the respective dihedral. The electrostatic interaction between atomic (partial) charges $q_{i,j}$ is described by the Coulombic law

$$V_{\text{Coul}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 \epsilon_r r_{ij}}. \tag{2.8}$$

The Pauli repulsion and van der Waals attraction are given by the Lennard-Jones term,[55]

$$V_{\mathrm{LJ}}(r_{ij}) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{2.9}$$

where r_{ij} is the distance between the two particles i and j. ϵ_{ij} and σ_{ij} define the depth and position of the potential minimum, respectively. These force field components are illustrated in Fig. 2.1.

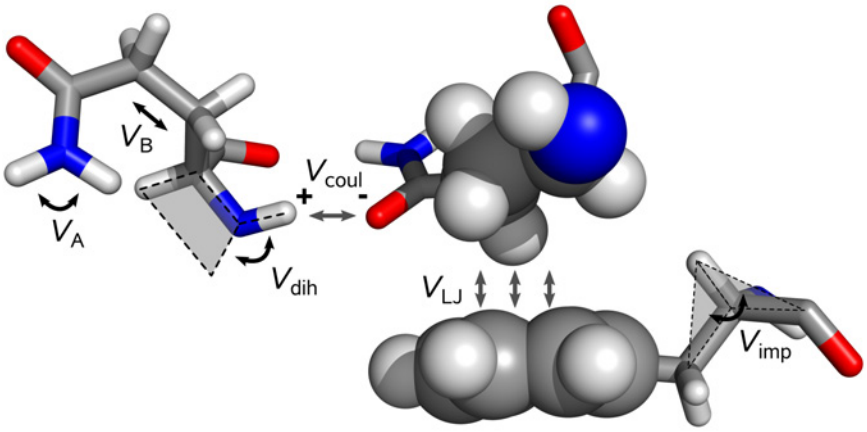

Fig. 2.1: Illustration of the interactions in a typical mechanical force field. The black arrows denote the bonded interaction potentials: bond-stretching, V_{B}, angle-bending, V_{A}, dihedral (out-of-plane), V_{dih}, and improper (plane-plane), V_{imp}. The grey arrows denote the non-bonded potentials: Coulomb, V_{coul} and Lennard-Jones, V_{LJ}.

Numerous force fields have been developed, differing in the way they were parameterized. The exact functional forms and individual parameters vary between the different force fields and may deviate from the more general Eq. 2.7. In this thesis, the OPLS-AA[56] and the AMBER99[57,58] force fields have been used. These force fields contain a large set of parameters, which are usually determined by multi-dimensional fitting to experimental data and to the results from quantum mechanical calculations. Whenever available, experimental and theoretical data for the condensed phase are used, since biomolecular systems are generally not in the gas phase.

The force fields described above have been developed for proteins on the basis of their

building blocks, amino acids. Furthermore, the AMBER99 force field has been developed for the nucleotides forming nucleic acids (DNA/RNA). Other molecules, present in many biological systems, require additional parameters. Since the force field parameters are determined by multi-dimensional fitting, single parameters of atoms have limited physical meaning, and only make sense if the force field is regarded as a whole. The parameterization of molecules not included in a force field, therefore, needs a careful adjustment to that respective force field. Depending on the desired accuracy, such parameters can be either derived *ab initio* from experimental data and quantum mechanical calculations, or if available, adopted from parameters of other available molecules – e.g. amino acids or nucleotides – in the force field. To circumvent the costly parameterization of single molecules, the general AMBER force field (GAFF)[59] has recently been developed and yields parameters for most organic molecules, compatible with the AMBER force field series.

2.2 Simulation details

About 70% of the mass of mammal bodies is composed of water. Approximately 2/3 of this water can be found within cells. Water therefore is the natural environment of proteins in a cell. To obtain a correct model system for MD simulations which resembles the *in vivo* system best, the proteins and their ligands used in the underlying simulations have been simulated in water and a 150 mM sodium-chloride salt concentration, mimicking the physiological solvent environment.

To be computationally feasible, MD simulations of biomolecular systems require much smaller system sizes than corresponding experiments. Hence, artifacts from the system boundaries have to be minimized by introducing appropriate boundary conditions. The common method to cope with these boundaries is the application of periodic boundary conditions. Here, the simulation system is surrounded by periodic images of itself, allowing atoms which leave the simulation box on one side to re-enter instantaniously on the opposite side. The box shape of such simulation systems is limited to space-filling geometrical forms,

like a triclinic, a dodecahedric, or a truncated dodecahedric unit cell. While the use of periodic boundaries is an effective approach to eliminate surface artifacts, care has to be taken to minimize arti.facts which arise from the artificial periodicity.[60,61,62] A cubic simulation box was used for all systems assessed in this thesis.

The calculation of the two non-bonded force terms, Coulomb (Eq. 2.8) and Lennard-Jones (Eq. 2.9) interactions, make up the computationally most expensive part of each integration step. It requires N^2 summations for the N atoms in a system. Hence, the computational effort quickly becomes too large to treat huge systems like proteins or nucleic acids in explicit solvent environment. Typically, a cut-off of 1.0 to 1.4 nm[63] around each particle is taken for the explicit non-bonded calculations to enhance the computational efficiency. Since the Lennard-Jones potential decays with r^{-6}, additional forces can be neglected for larger r. However, the use of such a cut-off for the Coulomb potential is known to cause severe artifacts[64,62,65] due to its slow $1/r$ decay.

Such long-range artifacts are avoided by the use of Ewald summation,[66] which originally was derived for the calculation of electrostatic potentials in periodic crystals. The similar periodicity in simulation systems with periodic boundaries therefore renders this method particularly well suited. When using Ewald summation, the electrostatic interaction is splitted into two contributions, of which those interactions within the cut-off are computed directly via Eq. 2.8 and the long-range interactions outside the cut-off are computed in reciprocal space. The use of fast Fourier transformations for the calculation of the reciprocal sum in the related Particle Mesh Ewald (PME)[67] is even more efficient than Ewald summation and scales with $N \log N$.

The numerical integration of Newton's equations of motion and the approximations in the evaluation of the non-bonded interactions introduce additional uncontrolled forces, which typically heat up the system and thus lead to a non-conserved total energy. This is prevented by adjusting the systems temperature T at each integration step via coupling the system to a heat bath with a reference temperature T_0 and a coupling time constant τ_T. The algorithm of Berendsen *et al.*[68] was used in this work, which rescales the atomic

velocities according to

$$v' = v\sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)}, \tag{2.10}$$

where Δt is the integration time step, v the initial and v' the resulting velocity, respectively. This temperature coupling method assumes the system to be thermally equilibrated, and removes temperature differences occuring between adjacent time steps from the system.

The proper protein function in experiments and living cells is usually not only coupled to a relatively constant temperature but also to a constant pressure of $\approx 1\,\text{bar}$. Similar to keeping the temperature of a simulation system constant by the described temperature coupling, a barostat is introduced to keep the pressure of the system constant. In this work, the Berendsen barostat[68] has been used, which controls the pressure by rescaling the atomic positions analogous to the velocity scaling for the temperature coupling.

If not explicitly mentioned elsewhere, the short-range Coulomb as well as the Lennard-Jones interactions were computed explicitely within a cutoff of 1.0 nm, and with Particle-Mesh-Ewald with a grid spacing of 0.12 nm and an interpolation order of 4 for long-range electrostatics.[67] All simulations were carried out in an NpT ensemble using Berendsen pressure and temperature coupling,[68] p=1 bar with a pressure coupling coefficient of τ_p=1 ps, and T_0=300 K with a temperature coupling coefficient of τ_T=0.1 ps. The simulations in this thesis were carried out with the GROMACS software-package (version 3).[69]

For the free energy calculations, a hardcore, i.e., an unmodified Lennard-Jones potential, and, where necessary, a softcore potential[70] was used. A softcore potential (V_{sc}) modifies the Lennard-Jones potential to allow the overlap of the van der Waals spheres of adjacent atoms. The softcore potential in this work was used with $\alpha = 0.25$ and, for hydrogens, with $\sigma = 0.3$. The two parameters are defined as

$$V_{sc}(r) = (1 - \lambda)V^A(r_A) + \lambda V^B(r_B) \tag{2.11}$$

with

$$r_A = \left(\alpha\sigma_A^6\lambda + r^6\right)^{\frac{1}{6}} \tag{2.12}$$

and

$$r_B = \left(\alpha\sigma_B^6(1-\lambda) + r^6\right)^{\frac{1}{6}}, \tag{2.13}$$

where V^A and V^B denote the Lennard-Jones potential (Eq. 2.9) for an arbitrary atom pair in state A and B, respectively, α is the soft-core parameter, σ the radius of interaction, and r_A and r_B the distance of an arbitrary atom pair in state A and B, respectively.

A softcore potential was used for perturbations involving changes of the number, i.e., appearing and disappearing, of atoms with van der Waals interactions. For the softcore potential a 3-step scheme was used to avoid Coulomb interaction singularities for overlapping atoms. Accordingly, the switching process from state A to state B was split into three steps, involving two intermediate states. In the first step, the charges of all perturbed atoms of state A were switched to zero. In the second step, we switched the Lennard-Jones parameters, masses, as well as all bond, angle, and dihedral parameters of the perturbed atoms. Masses of dummy atoms and the parameters of bonds involving dummy atoms were not changed. In the last step, the charges of the perturbed non-dummy particles were switched from zero to their state B value. If no softcore potential was necessary, the conventional 1-step scheme was used.

The simulation of a protein or nucleic acid requires the explicit spatial coordinates and initial velocities of every atom in such a molecule. While the initial velocities can be obtained from a Maxwell-Boltzmann distribution, the spatial coordinates were obtained from x-ray structures stored in the Protein Data Bank.[10] Such structures may lack heavy atoms from amino acid side chains or even whole loop regions, depending on their resolution. The addition of such side-chain heavy atoms can be obtained by smart software algorithms which compute the most likely coordinates for these atoms, depending on their environment. Loop modeling, in contrast, is a still highly problematic and difficult task. The intrinsically high flexibility of loops in proteins renders the determination of the atoms

spacial coordinates often challenging. Moreover, since the resolution of these structures is most often too low to contain hydrogen atoms, the structure has to be protonated before a simulation. Such structural repair and protonation has been performed with the software WHATIF[71] which optimizes the spatial placement of heavy atoms and protons by analyzing the environment for the energetically most favorable position at a given pH. The x-ray structures of the biomolecules, simulated in this work, are taken from the Protein Data Bank, and the respective PDB codes are given in the respective chapters 3 to 5.

The limited resolution of typically 1.0 to 3.0 Å and the necessary addition of missing atoms, protonation, and addition of solvent require further processing of the system. A short energy minimization using a steepest descent algorithm has been applied to each structure to remove possible van der Waals overlaps and bond-angle deformations, which could otherwise lead to unreasonably high forces in the simulations. Afterwards, the resulting structures have been simulated for 0.5 ns, where all heavy atoms of the solute have been restrained to their initial position, to allow the relaxation of the surrounding solvent. Furthermore, the unrestrained system was equilibrated to remove possible artifacts, e.g., non-native side-chain positions due to crystal-packing or other conditions, necessary for solving the structure. All simulations were performed with explicit water as referenced in the respective chapters.

2.3 Free energy calculations

The importance of the free energy for biochemical quantities has been described in detail in the introduction of this thesis. In contrast to the potential or kinetic energy, which can directly be computed from statistical averages of a MD trajectory, the (Gibbs or Helmholtz) free energy (G, F) cannot be computed from such a statistical average. The free energy and the entropy (S), which is connected to it via

$$G = H - TS, \tag{2.14}$$

is an ensemble property, which depends on the extent of configurational space (or phase space) accessible to the system. Therefore, computation of the *absolute* free energy of a biomolecular system is still a quite challenging task.

However, over the past few decades, statistical mechanical procedures have been developed for the calculation of *relative* free energies. The methods described below aim to compute the free energy changes, related to perturbations of a system, from MD simulations. Depending on the explicit method, the potential energy function $V(\mathbf{R})$ (Eq. 2.7) is changed, either slow or fast, from state A to B. Accordingly, the classical Hamiltonian $H(\mathbf{p}, \mathbf{R})$, is made a function of a coupling parameter λ, such that $H_{\mathrm{A}}(\mathbf{p}, \mathbf{R}, \lambda = 0)$ and $H_{\mathrm{B}}(\mathbf{p}, \mathbf{R}, \lambda = 1)$ describe the system in state A and B, respectively,

$$H(\mathbf{p}, \mathbf{R}, \lambda) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i(\lambda)} + V(\mathbf{R}, \lambda). \tag{2.15}$$

The free energies of state A and B are defined as

$$F_{\mathrm{A,B}} = -\frac{1}{\beta} \ln Z_{\mathrm{A,B}}, \tag{2.16}$$

where $Z_{\mathrm{A,B}}$ is the canonical partition function,

$$Z_{\mathrm{A,B}} = \int e^{-\beta(H_{\mathrm{A,B}}(\mathbf{p},\mathbf{R}))} d(\mathbf{p}, \mathbf{R}), \tag{2.17}$$

with the inverse temperature $\beta = \frac{1}{k_B T}$. The free energy difference therefore is

$$\Delta F = F_{\mathrm{B}} - F_{\mathrm{A}} = -\frac{1}{\beta} \ln \frac{Z_{\mathrm{B}}}{Z_{\mathrm{A}}}. \tag{2.18}$$

Methods to calculate free energy differences between two states A and B fall into two classes, equilibrium methods and, developed more recently, non-equilibrium methods. These methods will be briefly discussed in the following chapters. For a detailed overview of free energy calculations, Ref. 72 is suggested.

2.3.1 Equilibrium methods

Free Energy Perturbation (FEP)

Free Energy Perturbation (FEP) was developed by Zwanzig,[40] showing that the free energy difference can be computed directly via

$$\Delta F = -\frac{1}{\beta} \ln \left\langle e^{-\beta(H_{\mathrm{B}}(\mathbf{p},\mathbf{R})-H_{\mathrm{A}}(\mathbf{p},\mathbf{R}))} \right\rangle_{\mathrm{A}}, \tag{2.19}$$

where $\langle\rangle_{\mathrm{A}}$ denotes an ensemble average of the system at state A. FEP has been widely used to calculate free energy differences for amino acid substitutions in proteins.[73,74,75] This method, however, suffers particularly from sampling problems, which become increasingly severe for larger systems such as proteins, mainly due to insufficient overlap of the involved thermodynamic phase space densities. Because of the exponential growth of statistical uncertainty with decreasing phase space density overlap, this method requires excessive sampling, especially if the size of the perturbation is large.[76,77,78]

Slow Growth Thermodynamic Integration (SGTI)

Thermodynamic Integration (TI)[39] rests on the generalized force $(\delta H(\mathbf{p}, \mathbf{R}, \lambda)/\delta\lambda)$ with respect to the coupling parameter λ rather than on finite differences and can be computed by

$$\Delta F = \int_0^1 \frac{\delta H(\mathbf{p}, \mathbf{R}, \lambda)}{\delta \lambda} d\lambda. \tag{2.20}$$

Here λ is either varied continuously (slow-growth), or, similarly to FEP, sampled at discrete values (see next subchapter). One consequence of the continuous shift of λ in slow-growth TI (SGTI) is that, strictly speaking, the system is never in equilibrium, as is assumed in the construction of the method. Therefore, accurate results are expected only for small systems or very long simulation times.[79,80]

Discrete Thermodynamic Integration (DTI)

Discrete TI (DTI) circumvents the problems mentioned for SGTI in the previous paragraph by sampling at discrete λ values instead of a continuously variation:

$$\Delta F = \int_0^1 \left\langle \frac{\delta H(\mathbf{p}, \mathbf{R}, \lambda)}{\delta \lambda} \right\rangle_\lambda d\lambda \approx \sum_{i=0}^{N} \left\langle \frac{\delta H(\mathbf{p}, \mathbf{R}, \lambda)}{\delta \lambda} \right\rangle_{\lambda_i} \Delta\lambda, \tag{2.21}$$

Here, gradients of the free energy F are calculated at discrete λ values, and ΔF is obtained from numerical integration (summation) of these gradients. In regions of λ where these gradients are large, insufficient sampling and numerical integration may cause problems and, hence, require increased computational effort. This effect is often very large for λ=0 and λ=1, in particular when the number of atoms differs between the two involved states.[48] In this context, the Bennett Acceptance Ratio[81] typically provides an improved free energy estimate compared to numerical integration.[50]

2.3.2 Non-equilibrium methods

Exponential work averaging using a Gaussian approximation (EXP)

Jarzynski has shown that the Helmholtz free energy difference ΔF can be derived from a series of non-equilibrium work measurements or computations,[82,83]

$$e^{-\beta \Delta F} = \left\langle e^{-\beta W_\tau} \right\rangle, \tag{2.22}$$

where $\langle\rangle$ denotes an average over an ensemble of N trajectories which were started from an equilibrated canonical ensemble. Here, W_τ is the work

$$W_\tau = \int_0^1 \frac{\delta H_\lambda}{\delta \lambda} d\lambda \tag{2.23}$$

over a switching process of arbitrary length τ, which may be very short. In Eq. (2.23), λ is again the coupling parameter which switches the system during a simulation of length τ from state A to state B (defined by Hamiltonians H_A and H_B, respectively), e.g., via $H_\lambda = (1-\lambda)H_A + \lambda H_B$.

For very long switching times the system stays sufficiently close to equilibrium, such that the dissipated work is negligible and $\Delta F = W$. In this case, Eq. (2.23) describes Thermodynamic Integration (Eq. 2.20).[84,85]

Rather than spending computational effort on one long TI trajectory, Hendrix and Jarzynski suggested to carry out many short Fast Growth TI (FGTI) simulations from state A to B and to calculate the free energy difference between these two states via the Jarzynski equality (Eq. 2.22).[82,86] Unfortunately, however, from the resulting set of trajectories, those which occur only rarely carry most of the statistical weight. Therefore, the severe sampling problem persists[87,88] and renders the treatment of larger biomolecules often still as computationally demanding, as for FEP.

It has been suggested to alleviate this problem by combining forward (A→B) and reverse (B→A) switching simulations as well as by using Gaussian approximations of the work distribution.[89] For the method EXP, the distribution $P(W)$ of values for the work obtained from the trajectory ensemble is approximated by a Gaussian function,

$$P_{f,r}(W) \approx \frac{1}{\sigma_{f,r}\sqrt{2\pi}} \exp\left[-\frac{(W - W_{f,r})^2}{2\sigma_{f,r}^2}\right], \tag{2.24}$$

where $W_{f,r}$ and $\sigma_{f,r}$ are the means and the standard deviations of the work distributions, respectively. The index f denotes the forward ensemble, where $\lambda = 0 \rightarrow 1$, and r the reverse ensemble, where $\lambda = 1 \rightarrow 0$, respectively.

As has been shown by Hummer,[89] Eq. (2.22) yields for this approximation

$$\Delta F_f = W_f - \frac{1}{2}\beta\sigma_f^2 \tag{2.25}$$

and

$$\Delta F_r = -W_r + \frac{1}{2}\beta\sigma_r^2. \tag{2.26}$$

With the statistical accuracy of the two above expressions,

$$\bar{\sigma}_{f,r}^2 = \frac{\sigma_{f,r}^2}{N} + \frac{\beta^2\sigma_{f,r}^4}{2(N-1)}, \tag{2.27}$$

and using standard error propagation theory, we suggest to improve Hummer's estimate by using the weighted mean

$$\Delta F_{\mathrm{EXP}} = \left(\frac{\Delta F_f}{\bar{\sigma}_f^2} + \frac{\Delta F_r}{\bar{\sigma}_r^2}\right) \Bigg/ \left(\frac{1}{\bar{\sigma}_f^2} + \frac{1}{\bar{\sigma}_r^2}\right) \tag{2.28}$$

with statistical accuracy

$$\bar{\sigma}_{\mathrm{EXP}}^2 = \left(\frac{1}{\bar{\sigma}_f^2} + \frac{1}{\bar{\sigma}_r^2}\right)^{-1}. \tag{2.29}$$

The latter two expressions have been used as method EXP in our simulations. It should be noted that also Hummer[89] has derived an expression for $\sigma_f \neq \sigma_r$; however no error estimate has been derived for this case.

Bennett acceptance ratio with a maximum likelihood estimator (BAR)

Rather than relying on Jarzynski's equality, several methods are based on the more general Crooks Fluctuation Theorem (CFT, see Chapter 2.3.3).[90] Accordingly, the forward and reverse work distributions obey

$$\frac{P_f(W)}{P_r(-W)} = e^{\beta(W-\Delta F)}. \tag{2.30}$$

Following Eq. (2.30), Bennett's acceptance ratio[81] – which was originally proposed and applied for equilibrium Monte Carlo simulations – is also applicable to non-equilibrium work (NEW) calculations. Recently, this approach was combined with a maximum likelihood estimate,[91] and accurate free energy differences were obtained.[92]

Using Eq. (2.30), and assuming sufficiently smooth a priori distributions, a maximum likelihood approach on Bennett's Acceptance Ratio yields

$$\left\langle \frac{1}{1+\exp\left[\beta(W-\Delta F_{\mathrm{BAR}})\right]} \right\rangle_f = \left\langle \frac{1}{1+\exp\left[-\beta(W-\Delta F_{\mathrm{BAR}})\right]} \right\rangle_r , \tag{2.31}$$

assuming an equal number of forward (f) and reverse (r) simulations. From Eq. (2.31) ΔF_{BAR} is calculated numerically.

Following Shirts[91] and Anderson,[93] the statistical error of ΔF_{BAR} is estimated by

$$\frac{1}{\beta^2 N}\left[\left\langle \frac{1}{1+\cosh[\beta(W-\Delta F_{\mathrm{BAR}})]} \right\rangle^{-1} - 2\right], \tag{2.32}$$

where $\langle\rangle$ denotes the average over *all* (i.e., forward *and* reverse) work calculations. We will refer to this method as BAR in our work.

2.3.3 Crooks' Theorem

One goal of this work was to construct a non-equilibrium method for computing free energy differences between to states A and B, which performs better, especially with respect to the computation of precise errors, than the established methods discussed above. Because the newly developed method strongly exploits and relies on Crooks' fluctuation theorem (CFT),[90] this theorem will be described in more detail here.

The Crooks relation (Eq. 2.30) follows from the derivation of Jarzynski's equation (Eq. 2.22) using path-sampling ideas. A discrete trajectory $\mathbf{t}_0 \overset{H_1}{\rightarrow} \mathbf{t}_1 \overset{H_2}{\rightarrow} \cdots \overset{H_N}{\rightarrow} \mathbf{t}_N$, generated, e.g. by using Newtonian dynamics integrators, as described in section 2.1, proceeds for each step $\mathbf{t}_{i-1} \overset{H_i}{\rightarrow} \mathbf{t}_i$ according to Hamiltonian H_i. After each such time step, the Hamiltonian is changed in an designated way. If each of these time steps is path-independent, i.e., Markovian, in the full phase space, the probability P_f of computing a particular trajectory

can be obtained by

$$P_f\left(\mathbf{t}_0 \overset{H_1}{\to} \mathbf{t}_1 \overset{H_2}{\to} \cdots \overset{H_N}{\to} \mathbf{t}_N\right) = p_0(\mathbf{t}_0)\prod_{i=1}^{N} p_i(\mathbf{t}_i|\mathbf{t}_{i-1}), \tag{2.33}$$

where $p_0(\mathbf{t}_0) = \exp[-\beta(H_0(\mathbf{t}_0) - F_0)]$ is the normalized equilibrium Boltzmann probability density at the initial state $\mathbf{t}_0$ with the according free energy F_0, and $p_i(\mathbf{t}_i|\mathbf{t}_{i-1})$ the probability for a transition from $\mathbf{t}_{i-1}$ to $\mathbf{t}_i$ under the force of Hamiltonian H_i. If these transition probabilities satisfy the need for detailed balance, which is true for Newtonian dynamics,

$$\frac{p_i(\mathbf{t}_i|\mathbf{t}_{i-1})}{p_i(\mathbf{t}_{i+1}|\mathbf{t}_i)} = e^{-\beta[H_i(\mathbf{t}_i) - H_i(\mathbf{t}_{i-1})]}, \tag{2.34}$$

then the probability P_r of choosing the time-reversed path affected by the associated time-dependent Hamiltonian can be related to the probability P_f of the forward path:

$$\frac{P_f}{P_r} = \frac{P\left(\mathbf{t}_0 \overset{H_1}{\to} \mathbf{t}_1 \overset{H_2}{\to} \cdots \overset{H_N}{\to} \mathbf{t}_N\right)}{P\left(\mathbf{t}_N \overset{H_N}{\to} \mathbf{t}_{N-1} \overset{H_{N-1}}{\to} \cdots \overset{H_1}{\to} \mathbf{t}_0\right)} = \frac{p_0(\mathbf{t}_0)\prod_{i=1}^{N} p_i(\mathbf{t}_i|\mathbf{t}_{i-1})}{p_N(\mathbf{t}_N)\prod_{i=1}^{N} p_i(\mathbf{t}_{i-1}|\mathbf{t}_i)}. \tag{2.35}$$

Using Eq. 2.34 and the fact that the work W of bringing a system from state A to state B is the accumulated change in energy

$$W = \sum_{i=0}^{N-1} \left[H_{i+1}(\mathbf{t}_i) - H_i(\mathbf{t}_i)\right], \tag{2.36}$$

$$\frac{P_f}{P_r} = \exp\left[\beta \sum_{i=0}^{N-1} \left[H_{i+1}(\mathbf{t}_i) - H_i(\mathbf{t}_i)\right] - \beta(F_N - F_0)\right] = e^{\beta(W - \Delta F)} \tag{2.37}$$

is obtained, where $p_0(\mathbf{t}_0)$ and $p_N(\mathbf{t}_N)$ are substituted by F_0 and F_N, respectively, which are the free energies corresponding to Hamiltonians H_0 of state A and H_N of state B, with $\Delta F = F_N - F_0$.

As can be seen from Eq. 2.37, a backward path with a path probability differing exactly by a factor $\exp[\beta(W - \Delta F)]$ exists for every equivalent forward path. It can be seen that

the work distribution of the forward paths thus can also be obtained by sampling the reweighted backward paths. This reweighting factor apparently only depends on the work W and ΔF. Therefore it can be applied directly to obtain Crooks relation (2.30).

2.4 Analysis Methods

The positions and velocities of every atom of a simulation system are specified for every time step by the simulations trajectory. The structural changes of a protein can thus be obtained from such a trajectory, and are characterized by the methods introduced below.

Root Mean Square Deviations The root mean square deviation (RMSD) of a structure with atomic coordinates $\mathbf{r}_i$ with respect to a reference structure with its atoms coordinates $\mathbf{r}_i^0$ yields a quantitative measure for the structural difference between both,

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_i^0)^2}. \tag{2.38}$$

The definition of the RMSD requires the rotation and translation of the structure towards its best fit to the reference structure. The calculation of an appropriate RMSD of a proteins structure along its trajectory with respect to its crystal structure, e.g., yields a measure for conformational changes. Typically, an RMSD of 2–3 Å is caused by thermal fluctuations, whereas larger values point towards conformational changes.

To obtain informations about the globular motion of the structure, a subset of atoms from the system is chosen for the calculation of the RMSD via Eq. 2.38. This subset consists out of the amino acid backbone atoms, due to the noise induced by the fluctuations of amino acid side-chains. However, if a certain region in the protein is expected to contribute strongly to the total RMSD, a subset of atoms from that particular region can be used to calculate a more specific RMSD of that region.

Even though the RMSD can indicate general motions or conformational changes, the specific motions in phase space cannot be exactly determined. It is, therefore, not possible to correlate the globular motions of a protein from two ore more independent trajectories by their respective time resolved RMSD.

Principal Component Analysis One major technique to extract and classify information about large conformational changes from an ensemble of protein structures, generated either experimentally or theoretically, is principal component analysis (PCA). A detailed mathematical description of PCA is given in Ref. 94, 95. Principal component analysis is based on the observation that the largest part of positional fluctuations in biomolecules, like proteins, occurs along a small subset of collective degrees of freedom. The presence of a large number of internal constraints, defined by the atomic interactions in a biomolecule, leads to the dominance of this small subset of degrees of freedom (essential subspace) in the molecular dynamics of a protein. In particular, these interactions range from the strong covalent bonds to the weaker non-bonded interactions. PCA identifies the collective degrees of freedom that most contribute to the total amount of fluctuations. Typically, a small subset of 5–10% of the total degrees of freedom accounts for more than 90% of the total fluctuations within a protein.[94,96,97]

In general, PCA can be regarded as a multi-dimensional linear least squares fit procedure in configuration space. After fitting each configuration to a reference structure, the covariance matrix of the atoms positional fluctuations is build and diagonalized,

$$C = \left\langle (\mathbf{R}(t) - \langle \mathbf{R} \rangle)(\mathbf{R}(t) - \langle \mathbf{R} \rangle)^T \right\rangle , \tag{2.39}$$

where $\mathbf{R}(t)$ resembles the fitted ensemble (e.g. from a MD trajectory) of internal motions and $\langle\rangle$ an ensemble average. Here, $\mathbf{R}$ is a column vector of size $3N$, describing the coordinates of N atoms, and thus representing every structure of the ensemble. Because the collective motions of a protein are described very well by their backbone motions, the covariance matrix was made up by the proteins backbone atoms in this work. The sym-

metric $3N \times 3N$ matrix C is diagonalized by an orthogonal coordinate transformation D, containing the eigenvalues λ_i of matrix C. The ith column of D contains the normalized eigenvector, i.e. principal component, μ_i of matrix C corresponding to λ_i.

The eigenvalues λ_i describe the mean square fluctuations along the respective eigenvector μ_i. Hence, they contain each principal component's contribution to the total fluctuation. Sorting the eigenvectors μ_i according to their corresponding eigenvalue λ_i from large to small, therefore, yields a description of the collective motions of the system by the first eigenvectors.

These principal components comply with collective coordinates, including contributions from every atom of the protein, and were shown to make up for the functional dynamics of proteins in several cases.[94,98,99,100]

3

Assessment of Crooks Gaussian Intersection (CGI)

3.1 Introduction

In this chapter, we derive our new method to calculate free energies from non-equilibrium trajectories. The details of the available perturbation theory based methods, used below, have been described in Chap. 2.3. Here, we focus on the development of a new non-equilibrium method to compute free energy differences from MD simulations.

To compare the accuracy of our method with existing methods from both, the equilibrium and the non-equilibrium class, four established methods are considered and used for extensive test simulations. Recently proposed non-equilibrium methods include Jarzynski's work averaging and Bennett's Acceptance Ratio; established equilibrium methods are Slow Growth Thermodynamic Integration (SGTI) and Discrete Thermodynamic Integration (DTI).

3.2 Theory

The new method developed here, Crooks Gaussian Intersection (CGI) is an alternative set of methods which employs CFT (Eq. (2.30)) more directly. As well as for other non-equilibrium methods, for each trajectory the mechanical work for the switching process is calculated via Eq. (2.20). As can be seen from Eq. (2.30), ΔF is the work W for which $P_f(W) = P_r(-W)$, i.e, the intersection point of the two work distributions (Fig. 3.1).[88]

Direct determination of the intersection point (e.g., from histograms) is prone to large statistical errors, however, because only those work values are used which fall into the bin containing the intersection point. Particularly if the forward and reverse distributions exhibit only a small overlap, this number can be very small or even zero.

In analogy to method EXP, we here propose to use Gaussian approximations. This approach has the advantage that the assumption of a Gaussian work distribution has been shown to hold in the limit of large numbers of degrees of freedom.[101] Furthermore, as we will show below, this assumption can be tested through a Kolmogorov-Smirnov test.[102]

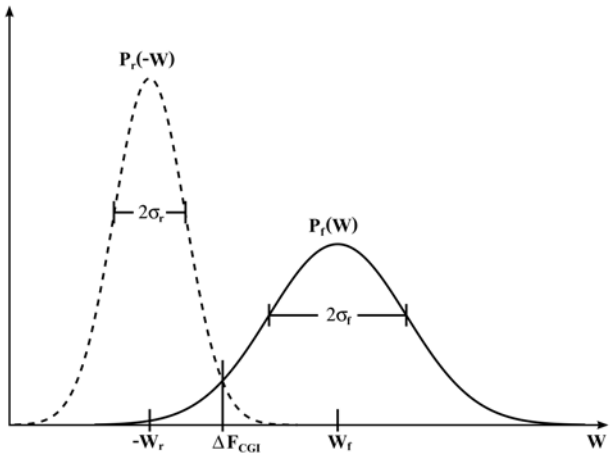

Fig. 3.1: Schematic Gaussian work distributions for the switching from state A to B for a forward ($P_f(W)$, solid line) and a reverse ($P_r(-W)$, dashed line) process. W_f, σ_f, $-W_r$ and σ_r are the means and standard deviations of $P_f(W)$ and $P_r(-W)$, respectively. These values are used to calculate the free energy difference ΔF_{CGI}.

The intersection point is given by

$$\Delta F_{CGI} = \frac{\frac{W_f}{\sigma_1^2} - \frac{-W_r}{\sigma_2^2} \pm \sqrt{\frac{1}{\sigma_1^2\sigma_2^2}\left(W_f + W_r\right)^2 + 2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\ln\frac{\sigma_2}{\sigma_1}}}{\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}}, \tag{3.1}$$

where W_f and $-W_r$ are the means, and σ_f and σ_r the standard deviations of the respective Gaussian functions, as also defined in Fig. 3.1.

We note that for $\sigma_f \neq \sigma_r$, two intersection points will generally be found, whereas ΔF is of course uniquely defined. Only for $\sigma_f^2 = \sigma_r^2$ a unique solution, $\Delta F_{\mathrm{CGI}} = (W_f - W_r)/2$,[103] is obtained. Typically, one intersection point is located between W_f and $-W_r$, and is close to $(W_f - W_r)/2$. The second intersection point is located far in the tail region and in this case is an artifact of our Gaussian approximation, which often fails in the tail region.[101] Accordingly, we chose the intersection value which is closest to $(W_f - W_r)/2$ as our estimate for ΔF_{CGI}.

If both Gaussians were too close to compute a proper intersection point, i.e., if W_f and $-W_r$ were closer to each other than to one of the two intersection values, we empirically

chose the mean of both as the best estimate for ΔF_{CGI}.

To estimate the statistical accuracy of ΔF_{CGI}, we resorted to a Monte Carlo approach. Accordingly, 10 000 synthetic sets of work values were generated, each set containing N values that were randomly chosen from a Gaussian distribution with mean W_f and standard deviation σ_f for the forward simulation, and, similarly, N values for the reverse calculations. Both, W_f, $-W_r$ and $\sigma_{f,r}$, were taken from the above Gaussian fits to the work distribution obtained from the simulations. For each of the 10000 synthetic forward/reverse work distributions, intersection points were calculated from the respective means and standard deviations. The standard deviation of this ensemble of synthetic intersection points was used as an estimate for the expected statistical error of ΔF_{CGI}.

3.3 Simulation details

Test systems To evaluate the accuracy and convergence of the described CGI method as well as the conventional ones, we considered two small test systems and a large one (see Tab. 3.1). The first, "E2M", involves the interconversion of one solvated ethane into a methanol molecule, which affects essentially all eight atoms (Fig. 3.2, E2M). The second, "W2G", is a tripeptide, Gly-Trp-Gly, which is "mutated" into Gly-Gly-Gly. Here, 19 atoms are involved, where 17 were transformed into dummy particles which keep their mass and bond parameters, but do not interact via Lennard Jones or via Coulomb interactions (Fig. 3.2, W2G). For both systems the OPLSAA force field was used.[56]

As a third, large test system of biological relevance we chose the protein snurportin 1 (SPN) (PDB entry 1XK5[104]). The ligand from the SPN crystal structure, m_3GpppG, was transformed to a modified form, m^7GpppG. Here, ten atoms are affected by the perturbation, and six become dummy particles (Fig. 3.2, SPN). For this system the AMBER99 force field[58,57] was used.

System	A-state	B-state	Force field	Water model	Ion conc. (mmol/L)	# atoms
E2M	C_2H_6	CH_3OH	OPLSAA	TIP4P	-	1432
W2G	GWG	GGG	OPLSAA	TIP4P	-	1389
SPN	m_3GpppG	m^7GpppG	AMBER99	TIP4P	150	63642

Tab. 3.1: Summary of simulation system details of the test-systems ethane (E2M), tripeptide (W2G), and snurportin 1 (SPN). The tripeptide in the A- and B-state is given in the one-letter-code for amino acids.

For both SPN ligands, m_3GpppG (resolved in the crystal structure) and m^7GpppG (not resolved in the crystal structure), no force field parameters were available. We therefore used the standard AMBER99 values for guanosine and ribose. The m_3G- and m^7G-nucleoside parameters were taken from Ref. 105, and the parameters for triphosphate, connecting the two nucleotides, from Ref. 106. Additionally, we scaled the charges of the molecule at the connecting phosphates such that a net charge of -2 was obtained.

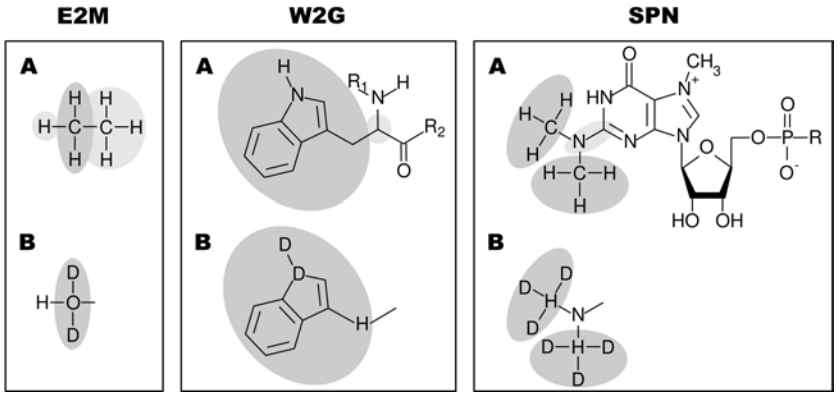

Fig. 3.2: Structures of the tested systems. **A**: A-state of the free energy calculations for E2M, W2G and SPN. **B**: Respective B-state. The dark grey ellipsoids denote the atoms, which contribute most to the perturbation. The light gray ellipsoids indicate regions of minor perturbations. R_x denotes the rest of the respective molecule, D the dummy particles. In **W2G B** the complete imidazole ring is composed of dummy particles.

Simulation setup All simulations were carried out in explicit solvent, except one SGTI simulation of E2M, which was additionally carried out in vacuum in order to compute

solvation free energies. All non-vacuum simulations were treated with the parameters as given in Chapter 2.2. The TIP4P water model[107] was used and, in the case of snurportin 1, a 150 mmol NaCl salt-concentration to mimic physiological solution. The E2M and W2G systems were equilibrated for one nanosecond; the SPN system was equilibrated for 50 ns. All test simulations were started from these equilibrated systems.

For the free energy calculations, we used a hardcore and, where necessary, a softcore potential. For the softcore potential a 3-step scheme as described in Chapter 2.2 was used. For the hardcore potential, the conventional 1-step scheme was used. The simulation length of all these steps was always chosen to be equal, and Gaussian error propagation was used to estimate the total error of ΔF.

Comparison to experiment As a control, solvation free energies were calculated for ethane and methanol, respectively, and compared to experiments. To this aim, nine SGTI simulations were carried out for system E2M in solvent and one *in vacuo*. All SGTI simulations were performed in a 1-step process over 10 ns in the forward and reverse direction, yielding a total simulation time of 200 ns.

Test of the Gaussian approximation To test the assumption that the obtained distributions of non-equilibrium work values can be approximated by Gaussian functions, 1000 forward and 1000 reverse FGTI simulations of τ=50 ps each were carried out for system E2M, using a 1-step switching process. After equilibrating the two states A and B for 11 ns, 1000 starting snapshots were extracted from the last 10 ns of each of the two trajectories. To the obtained set of work values, a Kolmogorov-Smirnov (KS-) test[102] was applied.

Convergence of trajectories: varying number and lengths For the two small systems, E2M and W2G, extended test simulations with comparable total simulation times were carried out. FGTI was used to compute trajectories for the NEW methods, and SGTI for comparison. To assess the effect of the number of used trajectories, the number of

trajectories for the non-equilibrium simulations was varied and a constant switching time τ=50 ps was used. 12, 25, 50, and 150 trajectories were used, yielding a total simulation time of 3.6, 7.5, 15, and 45 ns, respectively, for both the forward and reverse simulations. In the SGTI simulations, switching times τ were chosen to be 3.84, 7.5, 30, 37.5, 48, 60, and 75 ns, respectively, for the forward and reverse path. To allow proper comparison, the 3-step scheme was used for all simulations in this paragraph.

The effect of trajectory length (for a fixed number of trajectories) was assessed for all three systems, E2M, W2G, and SPN. The 3-step scheme was used for systems E2M and W2G; the 1-step scheme was used for system SPN. Switching times τ of 1, 5, 10, 25, 50, 80, 100, 128, 160, and 200 ps were used for all FGTI calculations. For each value of the switching time, 50 trajectories were calculated. The total simulation times for E2M and W2G were thus 0.3, 1.5, 3, 7.5, 15, 24, 30, 38.4, 48, and 60 ns, respectively.

For SPN, we additionally computed the interconversion of the two ligands m_3GpppG and m^7GpppG, bound to snurportin 1 and in solution to calculate their binding free energy difference. The total simulation times for these simulations were 0.6, 1, 2, 5, 10, 16, 20, 25.6, 32, and 40 ns.

Accuracy for given computational effort To evaluate the accuracy of all five methods SGTI, DTI, EXP, BAR, and CGI for given computational effort, the test systems E2M and W2G were used. A total simulation time of roughly 20 ns was spent for each method. The 3-step scheme was used for these simulations.

The SGTI simulation times for both systems were 9.6 ns each for the forward and reverse process, yielding a total simulation time of 19.2 ns. For the DTI simulations, the switching process was split into 11 equally spaced discrete λ values. For each of these values, the system was sampled for 600 ps, yielding a total simulation time of 18.6 ns. The first 200 ps of each simulation were discarded as additional equilibration time and were not used for the calculation of the intermediate free energies F_λ and their statistical accuracies.

For EXP, BAR, and CGI, identical trajectories were used as input. For the 1-step process

(system SPN), the A- and B-states were equilibrated for 1 ns. For the 3-step process (systems E2M and W2G), the intermediate states were also equilibrated for 1 ns. From the last 100 ps of each equilibration trajectory 50 snapshots were extracted as starting structures for the subsequent 50 ps FGTI simulations. Statistical independence was assessed via an autocorrelation analysis of $dH/d\lambda$ which yielded an autocorrelation time well below 100ps/50 = 2 ps. The total simulation time including the necessary equilibration runs was 21 ns.

3.4 Results

Comparison to experiment As a check, we compared computed free energies with experimental solvation free energies for E2M. The appropriate thermodynamic cycle (Fig. 3.3A) required interconversions of ethane into methanol both in solvent and in vacuum.

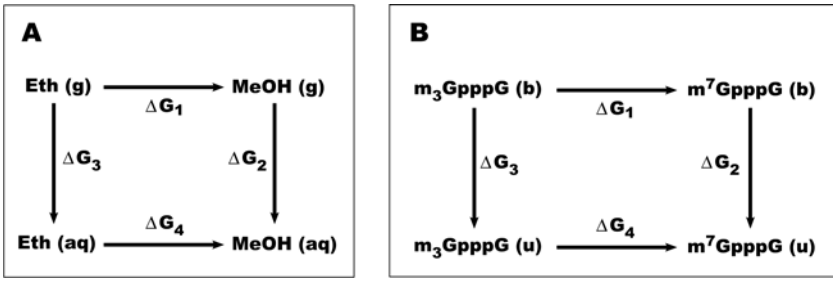

Fig. 3.3: Used thermodynamic cycles. **A**: switching from ethane (Eth) to methanol (MeOH) in vacuum (g) and aqueous solution (aq), **B**: switching from m_3GpppG to m^7GpppG bound to SPN (b) and unbound (u) in solution.

From nine SGTI simulations in solvent, ΔG_4=-21.91 $\pm$ 0.09 kJ/mol was obtained. Here, the error of the mean was estimated from the standard deviation (σ=0.27 kJ/mol) of the obtained nine free energy values. The observed value for the hysteresis was 0.09 kJ/mol on average, which agrees with this estimate. The vacuum free energy difference was ΔG_1=6.72 $\pm$ 0.03 kJ/mol, with the error estimated from the hysteresis. The resulting solvation free energy difference $\Delta\Delta G = \Delta G_1 - \Delta G_4$=28.63 $\pm$ 0.10 kJ/mol

agrees well with the experimental value of $\Delta\Delta G$=29.01 kJ/mol[108] as well as with earlier simulations, $\Delta\Delta G$=28.25 ± 1.13 kJ/mol.[41] We note that we consider Gibbs' free energies here, as the experiments refer to constant pressure conditions. Because only double differences are considered, the $p\Delta V$ term is not expected to affect our results.

Test of the Gaussian approximation Methods EXP and CGI rely on the assumption that the distribution of work values from non-equilibrium switching processes can be approximated by Gaussian functions. To test this assumption, 1000 independent simulations for the system E2M were performed, from which 1000 work values were obtained. Figure 3.4A shows the obtained distributions as well as the Gaussian fits.

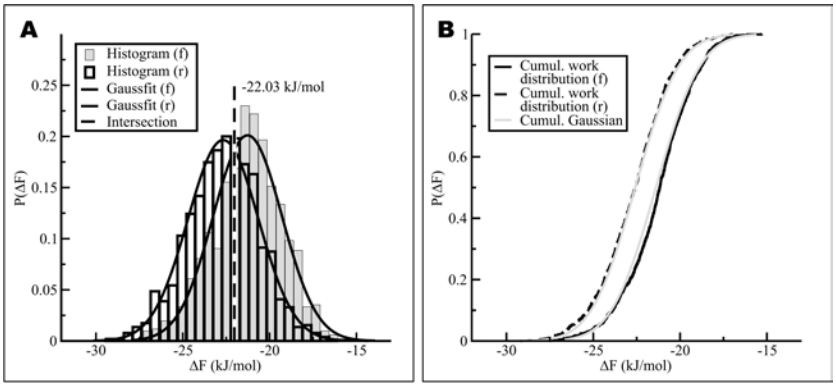

Fig. 3.4: **A**: Distribution of 1000 work measurements for the switching process of ethane to methanol in aqueous solution, forward (f) and reverse (r). A Gaussian function was fitted to the forward and reverse data, with their intersection yielding the free energy. **B**: Distribution functions for the same data, which are assessed by the Kolmogorov-Smirnov-test. For comparison, the respective cumulative distribution function (error function) of the respective Gaussian functions is shown in grey.

To test the hypothesis that these 1000 values are actually distributed according to a Gaussian function, the Kolmogorov-Smirnov-test was applied (Fig. 3.4B). The obtained significance levels of α_f=0.10 and α_r=0.50 for the forward and, respectively, reverse work distributions imply that the hypothesis cannot be rejected at the usual 5% (or lower) significance level. We thus assume that the Gaussian approximation holds for the case at

hand as well as for the other test systems considered here. We note that the limited set of work values lacks data points in the far tail regions of the distributions, which therefore cannot be reliably assessed by the KS-test. Because the intersection of the forward and reverse distribution falls typically within the "main body" of the Gaussian function and not in the tails, however, this uncertainty will not affect our free energy estimates.

Convergence for varying numbers of trajectories To assess the accuracy and convergence behavior of the three NEW methods, test simulations were carried out for the two test systems E2M and W2G. SGTI was used as a reference. Figure 3.5 shows the convergence of ΔF for the four methods as a function of the total computational effort spent. For SGTI, convergence is reached for both systems after about 40 ns. The mean and statistical accuracy of the last four energy values are ΔF=-21.89 $\pm$ 0.20 kJ/mol for E2M and ΔF=-16.54 $\pm$ 1.31 kJ/mol for W2G, respectively. We used these two mean values as our best reference estimates, against which convergence of the NEW methods is assessed.

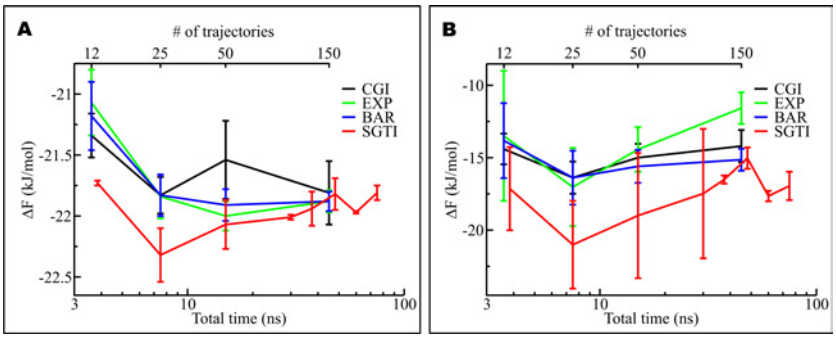

Fig. 3.5: Comparison of 3-step free energy calculations. **A**: System E2M, **B**: System W2G. The total simulation time is shown on a logarithmic time scale. The upper scale shows the corresponding number of FGTI trajectories used (uni-directional). For each SGTI free energy calculation, one trajectory was used.

As can be seen, beyond 7.5 ns, all NEW methods, except EXP for system W2G, yield free energy estimates, which agree with the reference within their respective statistical accuracy. Accordingly, for the systems at hand we consider 25 trajectories for the forward process and 25 for the reverse, respectively, sufficient to obtain reasonably converged results for

the NEW methods. Next, we studied the effect of trajectory length for a given number of 50 trajectories for systems E2M, W2G, and SPN. As can be seen in Figure 3.6, all NEW methods converge for all three systems at about 80 ps trajectory length. However, for the large system SPN, as well as for the large perturbation in system W2G, CGI turns out to converge markedly faster, such that accurate results are obtained already from quite short trajectories. Closer analysis reveals that CGI provides the same accuracy as EXP or BAR at about half of the computational effort. Furthermore, CGI provides reliable error estimates, whereas EXP drastically underestimates the statistical error.

Accuracy for given computational effort We compared the accuracy of both equilibrium (SGTI, DTI) and non-equilibrium methods (EXP, BAR, CGI) for given computational effort. For systems E2M and W2G, a total simulation time of roughly 20 ns was spent as described in the methods section. Table 3.2 shows the free energy contributions of the 3 steps as well as the total free energy difference obtained from the 3-step switching processes for system E2M. Table 3.3 shows the same results for system W2G. As can be seen from Table 3.2, for system E2M all applied methods yield free energy differences which converge to $\Delta G_{\mathrm{Tot}} \approx$ -21.9 kJ/mol. For a larger perturbation (system W2G), however, convergence to $\Delta G_{\mathrm{Tot}} \approx$ -15.2 kJ/mol with reasonable statistical accuracy was only reached by the NEW methods (Tab. 3.3). In addition to the DTI result using 11 discrete λ values listed in Tab. 3.3, DTI was also used with 21 discrete λ values, sampling each for 300 ps. However, no significant differences in the resulting free energies were seen (data not shown).

We finally applied the NEW techniques to SPN using the thermodynamic cycle shown in Figure 3.3B. As can be seen from Tab. 3.4, the free energy differences obtained for all three NEW methods are sufficiently accurate to allow a reliable calculation of the required double differences $\Delta\Delta G$. Despite the fact that quite small differences of large values are involved, CGI provides reliable values already for much shorter trajectory lengths than EXP and BAR. Comparison with the measured value of 5.61 - 11.22 kJ/mol[109,104] suggests that, given the relatively large experimental uncertainty, good agreement is obtained for

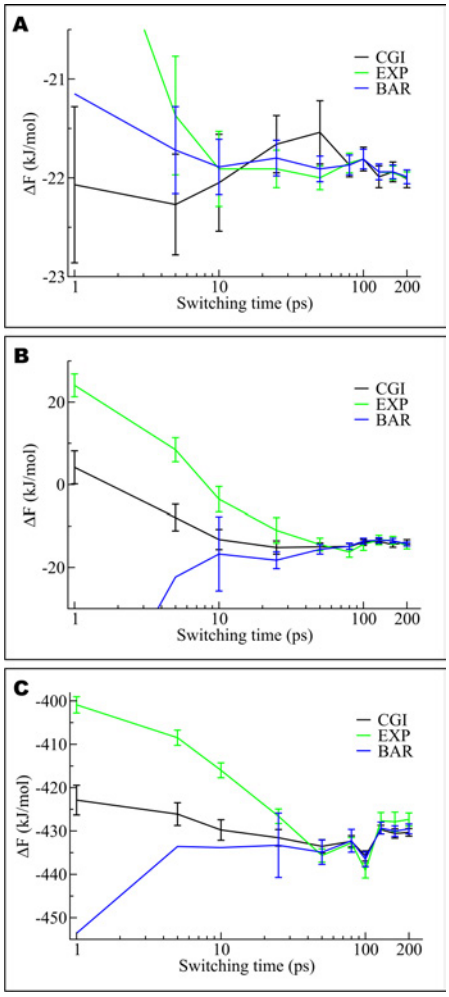

Fig. 3.6: Convergence of NEW free energies for differing FGTI trajectory lengths on a logarithmic scale. **A**: system E2M, **B**: system W2G, **C**: system SPN. Trajectory lengths vary from 1 ps to 200 ps. The errors of the BAR method for the very short trajectories (1-25 ps) are often very large and are therefore not shown in the plot.

Method	$\Delta G_{\mathrm{QQ_{off}}}$	σ	ΔG_{VdW}	σ	$\Delta G_{\mathrm{QQ_{on}}}$	σ	ΔG_{Tot}	σ
SGTI	-8.44	0.01	-0.77	0.09	-12.86	0.17	-22.07	0.19
DTI	-8.43	0.05	-0.70	0.49	-12.68	1.51	-21.81	1.59
EXP	-8.43	0.01	-0.77	0.05	-12.80	0.12	-22.00	0.13
BAR	-8.43	0.001	-0.78	0.002	-12.70	0.01	-21.91	0.01
CGI	-8.43	0.01	-0.78	0.24	-12.33	0.20	-21.54	0.31

Tab. 3.2: Free energy contributions of a 3-step switching process of system E2M. "$\mathrm{QQ_{off}}$" denotes switching the charges of the perturbed atoms in the state A to zero, "VdW" the interconversion of the Lennard Jones parameters from state A to B with softcore potentials, "$\mathrm{QQ_{on}}$" the switching of the charges to their values in state B, and "Tot" the sum of these three contributions, yielding the total free energy difference. All free energies are given in kJ/mol. The estimated error for all values is σ. For slow growth TI, σ provides a lower bound error estimate.

Method	$\Delta G_{\mathrm{QQ_{off}}}$	σ	ΔG_{VdW}	σ	$\Delta G_{\mathrm{QQ_{on}}}$	σ	ΔG_{Tot}	σ
SGTI	26.61	0.3	-52.00	4.17	6.39	1.13	-19.00	4.33
DTI	26.79	2.94	-48.42	9.46	7.68	1.66	-13.95	10.04
EXP	25.55	0.21	-46.84	1.52	6.86	0.05	-14.43	1.54
BAR	25.54	0.03	-48.38	1.24	7.24	0.02	-15.60	1.24
CGI	25.53	0.20	-47.63	0.92	7.24	0.16	-14.98	0.96

Tab. 3.3: Free energy contributions of a 3-step switching process of system W2G. Symbols are defined as in Tab. 3.2.

both, CGI and BAR.

Method	ΔG_{Cpx}	σ	ΔG_{Sol}	σ	$\Delta\Delta G$	σ
EXP	-427.35	1.51	-441.18	0.67	13.83	1.65
BAR	-429.52	1.29	-440.91	0.17	11.39	1.30
CGI	-430.34	0.86	-441.16	0.51	10.82	1.00

Tab. 3.4: Free energy contributions of a 3-step switching process of SPN. "Cpx" denotes the free energy difference calculated for the ligands bound to the protein, and "Sol" for the ligands in solvent, respectively. All free energies are given in kJ/mol. The estimated error for all values is σ.

3.5 Discussion & Conclusions

We compared established equilibrium (SGTI, DTI) and non-equilibrium (EXP, BAR) free energy calculation methods. Based on our results, an improved non-equilibrium method, CGI, has been derived and tested, which combines the advantages of existing methods. Interconversions between ethane and methanol (E2M), between tryptophane and glycine within a tripeptide (W2G), and between m_3GpppG into m^7GpppG bound to the globular protein snurportin 1 (SPN) were used as test cases. The main aim here was to assess recently proposed non-equilibrium methods, particularly for large perturbations which typically occur in the context of complex biomolecular systems.

For system E2M, the calculated free energy difference of $28.63 \pm 0.1\,\mathrm{kJ/mol}$ agrees well with the measured value of $29.01\,\mathrm{kJ/mol}$, which suggests that the used force field and simulation times provide sufficiently accurate results for our further assessments.

To assess their convergence behavior, the three NEW methods as well as SGTI were compared by calculating the free energy of the two systems E2M and W2G for different simulation times. Extended SGTI simulations were found to converge after about 40 ns and were therefore taken as reference. All NEW methods, except EXP for system W2G, yielded converged and, compared to SGTI, accurate free energies (within the error bars) already above 7.5 ns, i.e., for five times shorter simulation times.

The effect of trajectory length on the convergence behavior of the NEW methods was then studied in more detail by varying the trajectory length of the FGTI simulations performed for the systems E2M, W2G, and SPN. Using ensembles of 50 trajectories, all NEW methods converged for all studied systems for trajectory lengths above 80 ps. Remarkably, our new method CGI converged significantly faster, particularly for the more complex perturbations.

Both equilibrium (SGTI, DTI) and non-equilibrium methods (EXP, BAR, CGI) were then compared for given computational effort. To compute the free energy differences for the systems E2M and W2G, a total simulation time of roughly 20 ns was spent. For the very small system E2M, all five methods turned out to perform reasonably well. For such a small system, convergence after 20 ns is indeed expected. Importantly, for the larger perturbation of system W2G, the NEW methods provide significantly more accurate results than the equilibrium methods. Among the equilibrium methods, DTI was found to provide the most accurate free energy values; however, the very large statistical uncertainty involved is clearly a disadvantage.

As our new method CGI (as well as EXP) relies on the validity of the Gaussian approximation of work distributions, this assumption deserved particular attention. To test this assumption, Kolmogorov-Smirnov-tests were applied to the work distributions calculated for the systems mentioned above. 1000 forward and 1000 reverse work values were obtained from a corresponding number of FGTI trajectories, and their distributions were subjected to a Kolmogorov-Smirnov-test. Indeed, the hypothesis of an underlying Gaussian distribution could not be rejected, such that the Gaussian approximation for work distributions seems to be sufficiently accurate, which is in line with the accurate values obtained by CGI.

We attribute the improved accuracy of CGI over EXP to the fact that CGI is derived directly from the more general Crooks' theorem rather than from Jarzynski's equality, in particular, CGI exploits the relation between forward and reverse distributions, which is not implied in Jarzynski's equality.

As a check against experiment involving a large system, we finally calculated binding free energies for two ligands, m_3GpppG and m^7GpppG, bound to the globular protein snurportin 1 (SPN). The obtained values of $11.39 \pm 1.3\,\mathrm{kJ/mol}$ (BAR) and $10.82 \pm 1.0\,\mathrm{kJ/mol}$ (CGI), respectively, agree with the estimate from experiments of 6 - 11 kJ/mol, albeit such large experimental error does not allow to attribute a pronounced significance to this result. Nevertheless, the value of $13.83 \pm 1.65\,\mathrm{kJ/mol}$ obtained from method EXP, is clearly too large.

Overall, for the three test systems considered, the non-equilibrium methods were shown to outperform the traditional equilibrium methods. The best results were obtained for our newly proposed CGI method. We attribute the substantially faster convergence of the non-equilibrium methods to the fact that these do not rely on the assumption that the system is sufficiently close to equilibrium at all times. Rather, only the ensembles from which the required sets of trajectories are started need to be sufficiently equilibrated.

A further advantage of the non-equilibrium methods is that they are inherently parallel, and therefore scale optimally on parallel machines, in contrast to calculations of single long trajectories.

4

Major Histocompatibility Complex II

4.1 Introduction

The CGI method for computing free energy differences, developed in the previous chapter, has provided encouraging results, even for large sidechain perturbations or large globular proteins. In this chapter, the CGI method will be applied to compute free energy differences for several sidechain mutations in a large globular protein. To this end, the major histocompatibility complex was chosen as simulation system which is very important for the function of the immune system of humans and therefore is a well-suited target for drug development.

The immune system of vertebrates is a complex assembly of molecules and cells. Its major function is to protect the organism against microorganisms like viruses, bacteria, and parasites. The membrane bound major histocompatibility complex (MHC) is one of the key elements in the recognition and response mechanism, triggered by peptides of such microorganisms. There are two related, but structurally distinct, families of MHC proteins, divided into Class I and Class II.[110]

The recognition of peptides by MHC varies by individual differences in the amino acid sequence of the protein binding site. The human genome contains six different genes for MHC Class I and Class II proteins, respectively. The three loci on each of the two chromosomes for Class I genes are *human leukocyte antigen*(HLA)-A, HLA-B, and HLA-C. Those for Class II genes are HLA-DP, HLA-DQ and HLA-DR. These loci are highly polymorph and have been extensively studied, resulting in a detailed physical map of the region.[111] Every human therefore owns six MHC Class I and II receptors, mainly differing in the amino acid sequence of the peptide binding domain.

This enables the recognition of a multitude of peptides by the MHC proteins which lead to an immune response of the system. In the case of morbid germs this is a volitional reaction and very important for the survival of the organism. In the case of pollen or auto-immune response this feature is nowadays a growing civilizational problem. It is still not clear, how this broad peptide recognition with only 12 different receptors works on molecular

level. Due to the peptide ligands which have an intrinsically high variation in amino acids, the MHC Class II receptor is a very promising target for free energy calculations, yielding insight in the binding affinity differences on the molecular level.

MHC Class I presents peptides derived from endogenously synthesized proteins, such as viral compounds produced upon infection of the cell, to the extracelluar space. Cytotoxic T cells recognize these complexes on the surface and initiate lysis of the infected cell.[112] In contrast, MHC Class II presents peptides derived from degradation of endosomal proteins coming from, e.g., immunoglobuline-virus complexes, and stimulates helper T cells which then stimulate antibody production.[113] Therefore, MHC Class II receptors are expressed exclusively in B-cells, macrophages and dendritic cells.[114,115]

The MHC Class II protein consists of a 33 kD α-chain and a non-covalently bound 30 kD β-chain. Each of these chains contains two extracellular domains, one transmembrane segment, and a short intracellular tail. The petide binding region consists of two N-terminal domains (α1, β1) which display a remarkably high sequence variability. The folding of the α1 and β1 region results in a collective peptide-binding domain. Multiple MHC Class II structures as well as diseases, associated to different alleles, have been recently reviewed.[116]

The influenza virus is one of a larger number of viruses hosted by humans. Influenza belongs to the class of negative-strand RNA orthomyxoviruses and infects mucous membrane cells of the respiratory tract. Hemagglutinin (HA) which is part of the influenza viruses nucleocapsid envelope binds to sialic acid which is common in the membranes of these mucous membrane cells as well as in those of red blood cells. Peptide fragments of hemagglutinin, such as HA 307-319 are presented by MHC Class II proteins to helper T cells. The antibodies against hemagglutinin resulting from the induced immune response prevent cell infection, and thus neutralize the virus.[117]

The efficiency of vaccines against influenza varies between individuals, due to the highly polymorph binding sites of MHC Class II. It is therefore conceivable that a HA peptide, which is used as vaccine, binds well to many variants of MHC II. To design vaccines with increased broadband binding performance, it is important to understand the structural

determinants of the binding affinity changes and thus enhance the overall binding due to mutations in the peptide.

Here, we investigate the changes in binding affinity of certain mutants of the HA 307-319 peptide to a MHC Class II HLA-DR1 protein.

4.2 Methods

We used the structures 1DLH[118] and 1KLU[119] of the MHC Class II HLA-DR1 receptor from the PDB database[10] to construct the simulation system. The receptor in 1KLU has a better resolution (1.93 Å) than 1DLH (2.80 Å). However, since the 1KLU structure does not contain the proper hemagglutinin ligand, the peptide HA 307-319 (sequence: PKYVKQNTLKLAT) from 1DLH was fitted on the triosephosphate isomerase peptide (sequence: GELIGTLNAAKVPAD) in 1KLU, neglecting the N- and C-terminal residue of the latter for the fit. After removal of the original ligands, the modified 1KLU structure (Fig. 4.1) was used as starting structure for the MD simulations, together with the OPLS-AA[56] force field.

All simulations were carried out with the parameters as given in Chapter 2.2. The simulations were carried out in explicit solvent with the TIP4P water model[107] and a 150 mmol NaCl salt concentration to mimic a physiological environment. For the free energy calculations, the 3-step scheme described in Chapter 2.2 was used.

Eight point-mutations of the wildtype HA 307-319 were selected according to their measured[120] binding affinities, as described below. As can be seen from Tab. 4.1, the mutations were chosen such that they fit into the four regimes of better, equal, or worse binders, compared to the wildtype, as well as negative binders, which couldn't be specificly measured. Two mutations for each of these four classes are studied in this work.

To let the protein adapt to the exchanged ligand, the system was equilibrated for 20 ns. The resulting structure was used to generate eight systems with point mutations in their

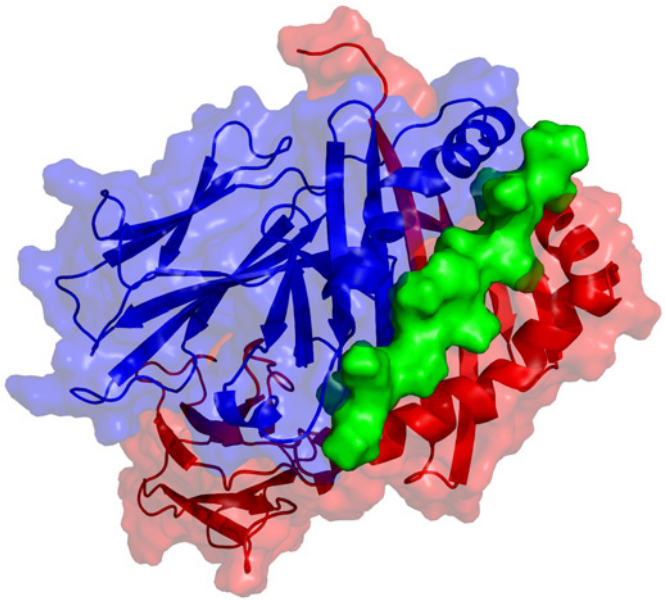

Fig. 4.1: Structure of MHC Class II with bound HA 307-319. The structure of the last frame from a 20 ns equilibration trajectory is shown. The α- and β-chains are colored in blue and red, respectively. The hemagglutinin peptide ligand is shown as surface representation in green.

B-state for the MHC Class II/HA 307-319-complex and the unbound HA 307-319 peptide according to Tab. 4.1. Figure 4.2 shows the positions of the amino acids in the complex which were chosen for mutations in HA 307-319.

Mutation	Y309A	Y309S	N313Q	T314A	T314Y	L315I	K316R	L317Q
Rel. affinity	<0.001	<0.001	0.92	3.0	0.026	1.1	4.4	0.009

Tab. 4.1: Selective mutations of HA 307-319. The relative binding affinity is compared to the wildtype affinity of 4.3 ± 0.43 nM. Values > 1 and < 1 denote a better and worse binding affinity, respectively. For the absolute binding affinity, the WT affinity has to be divided by the relative binding affinity of the according mutant.

To obtain an equilibrium ensemble to start the free energy calculations from, the A-state (wildtype) and the B-state (mutant) as well as two intermediate states of the sixteen systems were equilibrated for 5 ns each. From the last nanosecond of the resulting trajectories, 50 independent and equally distributed snapshots were taken to compute the free energy difference of each point mutation with the previously introduced Crooks Gaussian Inter-

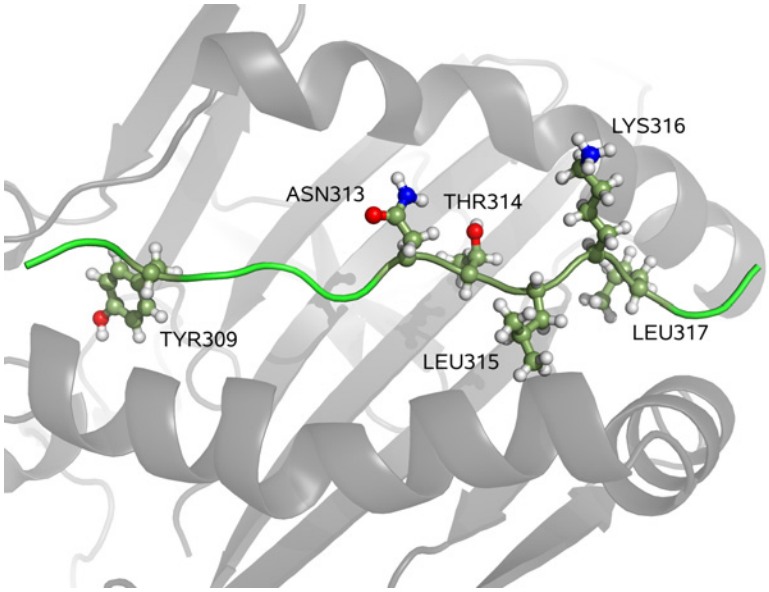

Fig. 4.2: Mutated residues in HA 307-319. The peptide is shown in green, the MHC Class II receptor in gray. The wildtype amino acids for the studied mutations are labeled by the 3-letter code and their position in the peptide.

section (CGI) method. The length of each of the underlying Fast Growth Thermodynamic Integration simulations was 50 ps. The single free energy contributions from (1) switching off the charges of the system in state A ($\Delta G_{\mathrm{QQ_{off}}}$), (2) transfering the Lennard-Jones and bonded parameters from state A to B (ΔG_{VdW}), and from (3) switching the charges to their final values in state B ($\Delta G_{\mathrm{QQ_{on}}}$) were added to obtain the total free energy difference (ΔG_{Tot}) between the two states A and B. The error (σ) of ΔG_{Tot} was computed by Gaussian error propagation as described in Chap. 3.

For comparison to experiment, the binding free energy double differences ($\Delta\Delta G$) for each mutation were calculated according to the thermodynamic cycle shown in Fig. 4.3. To compute $\Delta\Delta G$ for the experimental values,[120] the measured 50% inhibitory dose ($\mathrm{ID_{50}}$) for the wildtype and each mutant was used to estimate the equilibrium dissociation constant ($\mathrm{K_D}$). This rests on the assumption that, in the case of MHC proteins, it is likely that inhibitory peptides bind to the same binding site as native ones. Hence, the $\mathrm{ID_{50}}$ describes not only the 50% loss of activity of the protein at a certain ligand concentration, but also the

ligand concentration for which half of the ligand is bound to the protein (50% inhibitor, 50% ligand). The latter is equivalent to the definition of the dissociation equilibrium constant (K_D). The temperature for the experiments was not given in the literature and therefore assumed to be room temperature, 300 K. Taken together, these two assumptions allow the usage of the relationship between K_D and ΔG to compute the binding free energy double differences from the experimental values.

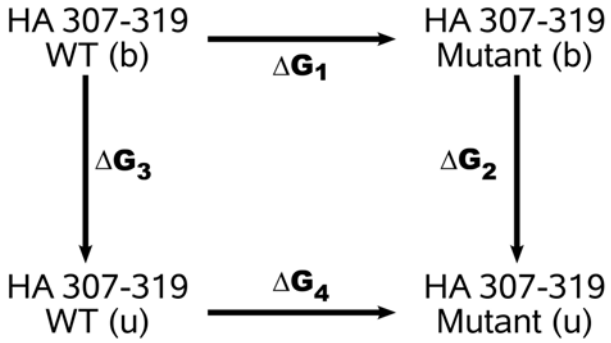

Fig. 4.3: Thermodynamic cycle. Switching of a single amino acid from hemagglutinin (HA 307-319) wildtype (WT) to a different amino acid (Mutant), bound to MHC Class II (b) and unbound (u) in solution.

4.3 Results & Discussion

Differences in binding affinities for eight point mutations in the influenza viral peptide HA 307-319 bound to the MHC Class II protein have been calculated.

To validate the stability of the constructed structure used as starting structure for the MD simulations, the backbone root mean square deviations (RMSD) with respect to the crystal structure of chain A and chain B in the MHC Class II protein as well as for the HA 307-319 peptide were calculated. Figure 4.4 shows the respective RMSDs over the equilibration time of 20 ns.

Despite the strong differences in amino acid composition of the triosephosphate isomerase

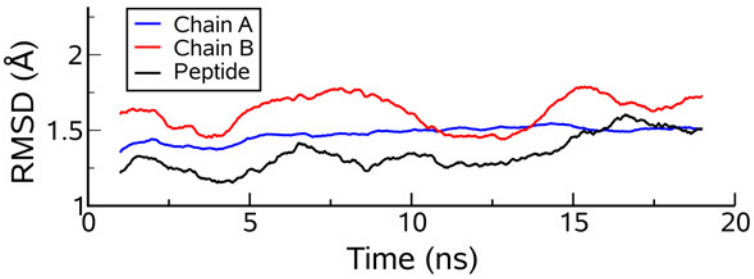

Fig. 4.4: RMSD of HA 307-319 bound to the MHC Class II protein. The backbone RMSD for chain A and B are shown in blue and red, respectively. The backbone RMSD for the peptide is shown in black. All RMSDs have been smoothed with a running average using 100 datapoints for the averaging procedure.

and hemagglutinin peptide, the constructed structure displays only little structural fluctuations around the equilibrium configuration. Moreover, the low RMSDs and the small drifts of the chains are remarkable, and therefore render the system well suited as starting point for the free energy calculations.

Tables 4.2 and 4.3 show the binding free energy differences of all simulated mutants (Y309A, Y309S, N313Q, T314A, T314Y, L315I, K316R, and L317Q), computed for the complex and the unbound peptide in solvent, respectively.

With the thermodynamic cycle shown in Fig. 4.3, the free energy double differences ($\Delta\Delta G$) were calculated from the measured, as well as the simulated free energy differences (Tab. 4.4 and Fig. 4.5). For the two mutants Y309A and Y309S, experimental measurements yielded a lower bound $\Delta\Delta G_{\mathrm{EXP}} = 17.23\,\mathrm{kJ/mol}$. No error estimates for the experimental values were available.

As can be seen in Fig. 4.5, the computed double differences ($\Delta\Delta G$) systematically deviate from the experimental values. However, the tendency of the calculated $\Delta\Delta G$ agrees very well with experimental results. Despite the differences between large numbers in some cases, the calculations are thus sufficient to predict the qualitative trend correctly.

The computed $\Delta\Delta G$ for the mutants N313Q, L315I, and K316R, however, strongly diverge from the experimental values. Although it is noticeable that these mutants are character-

Mutant	$\Delta G_{\mathrm{QQ_{off}}}$	σ		ΔG_{VdW}	σ		$\Delta G_{\mathrm{QQ_{on}}}$	σ		ΔG_{Tot}	σ
Y309A	42.25	0.37		11.51	2.70		-0.34	0.17		53.41	2.73
Y309S	41.48	0.39		19.15	3.13		42.46	0.48		103.09	3.19
N313Q	163.83	1.44		19.90	0.58		-180.93	0.60		2.80	1.66
T314A	-0.90	0.39		19.62	0.78		-3.32	0.08		15.40	0.87
T314Y	11.08	0.75		-58.36	2.80		-32.33	1.25		-79.60	3.15
L315I	-8.03	0.07		39.04	0.30		14.76	0.10		45.77	0.32
K316R	-74.22	1.76		15.27	0.70		30.21	0.67		-28.75	2.01
L317Q	-11.56	0.06		-42.23	0.46		-161.32	0.43		-215.11	0.63

Tab. 4.2: Free energy contributions of a 3-step switching process for HA 307-319 bound to MHC. "$\mathrm{QQ_{off}}$" denotes switching the charges of the perturbed atoms in the state A to zero, "VdW" the interconversion of the Lennard Jones parameters from state A to B with softcore potentials, "$\mathrm{QQ_{on}}$" the switching of the charges to their values in state B, and "Tot" the sum of these three contributions, yielding the total free energy difference. All free energies are given in kJ/mol. The estimated error for all values is σ.

Mutant	$\Delta G_{\mathrm{QQ_{off}}}$	σ		ΔG_{VdW}	σ		$\Delta G_{\mathrm{QQ_{on}}}$	σ		ΔG_{Tot}	σ
Y309A	41.04	0.22		-22.37	0.94		-0.76	0.13		17.91	0.97
Y309S	42.10	0.28		-16.04	0.94		25.68	0.49		51.74	1.10
N313Q	152.55	0.61		21.83	0.35		-178.49	0.36		-4.11	0.79
T314A	-0.67	0.34		23.04	0.55		-2.56	0.08		19.82	0.65
T314Y	1.55	0.41		-43.81	0.79		-44.06	0.30		-86.32	0.94
L315I	-5.71	0.25		26.18	0.43		15.04	0.03		35.52	0.50
K316R	-57.77	1.65		12.67	0.28		28.53	1.18		-16.57	2.05
L317Q	-10.67	0.04		-40.34	0.41		-181.28	0.45		-232.30	0.62

Tab. 4.3: Free energy contributions of a 3-step switching process for unbound HA 307-319 in solvent. Symbols are defined as in Tab. 4.2.

Mutant	$\Delta\Delta G_{\mathrm{EXP}}$	$\Delta\Delta G_{\mathrm{SIM}}$	σ
Y309A	17.23	35.50	2.90
Y309S	17.23	51.35	3.37
N313Q	0.21	6.91	1.84
T314A	-2.74	-4.42	1.09
T314Y	9.10	6.72	3.29
L315I	-0.24	10.25	0.59
K316R	-3.70	-12.18	2.87
L317Q	11.75	17.19	0.88

Tab. 4.4: Comparison of $\Delta\Delta G$ between simulations and experiments. $\Delta\Delta G_{\mathrm{EXP}}$ and $\Delta\Delta G_{\mathrm{SIM}}$ denote the binding free energy double differences from experiments and simulations, respectively. All values are given in kJ/mol. The estimated error for $\Delta\Delta G_{\mathrm{SIM}}$ is σ.

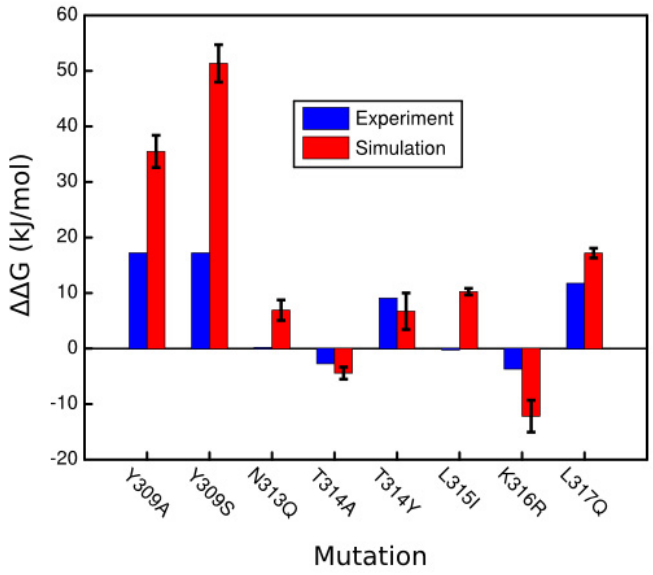

Fig. 4.5: Comparison of $\Delta\Delta G$ from experiments and simulations. The data from Tab. 4.4 is shown in a bar-plot. The error bars for the simulations denote the σ values. No error bars are available for the experimental values.

ized by a similar functional group in both states (wildtype and mutant), and an overall trend to overestimate the $\Delta\Delta G$ in the simulations, except for T314Y, is observed. This overestimation was quantified by a correlation analysis of the experimental and simulated data sets where the values for Y309A and Y309S were excluded from the analysis due to the missing accurate experimental values. The correlation coefficient was $r^2 = 0.58$ and the simulated values overestimate the experimental values by a factor of ≈ 1.2.

However, the observed deviations between the values computed by the simulations and from experiments can be attributed to the missing error estimates for the experimental values. All computed free energy difference values, except for L315I, are comparable to the experimental ones within 2σ, even if only small experimental errors would have been observed.

Our results therefore predict the binding affinity differences between the wildtype and the two mutants Y309A and Y309S which could not yet be accurately measured in experiments, and hopefully will be measured soon. In summary, these results provide the first semi-

quantitatively correct first principles calculation of peptide binding free energy differences and show that CGI free energy simulations are a valuable method to assess ligand binding affinities.

5

Snurportin 1

5.1 Introduction

Snurportin 1 (SPN) was used as a test case for the CGI method, derived in Chapter 3, with a large globular protein. The results for the computed binding affinities of the two ligands m_3GpppG and m^7GpppG to SPN lead to the investigation of the causes for this selective binding in this chapter.

Transporting macromolecules in and out of the nucleus is known to be highly important for eukaryotic cells to function properly. Hence, these transport processes have been intensively investigated and reviewed.[121,122]

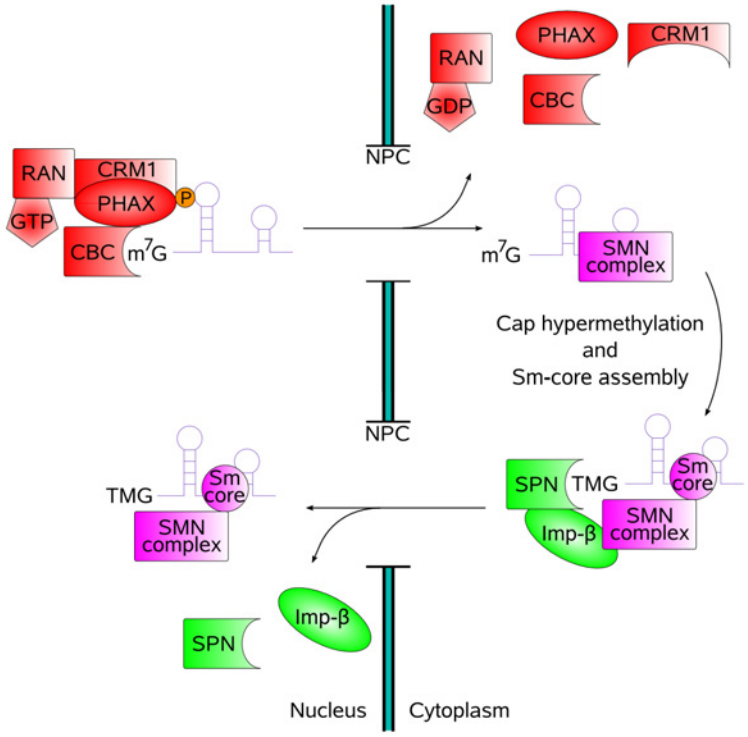

Fig. 5.1: Reduced nucleocytoplasmatic cycle of snRNA. In the nucleus, an exporting complex (red) is formed, which binds to the m^7G-capped snRNA and transports it through the nuclear pore complex (NPC). After dissociation, cap hypermethylation and Sm-core assembly, the importing complex (purple and green) is formed by aggregation of the SMN complex with importin β and snurportin 1 (SPN), which binds the m_3G-capped snRNA. After transport of the snRNA back into the nucleus, the importing complex dissociates.

The splicosome is a complex consisting out of proteins and small nuclear RNA molecules,

the snRNAs. It removes non-coding sequences, i.e. introns, from pre-mRNA. Subsequently, this processed mRNA contains only the coding sequences of a protein. The splicosome is formed by several ribonucleoprotein subunits called U snRNPs or "snurps" (uridine-rich small nuclear ribonucleoproteins). These U snRNPs have to be assembled in the cytoplasm and transported into the nucleus afterwards. Hence, a nucleocytoplasmic cycle has been postulated for these transport processes[123] (see Fig. 5.1).

The exporting complex is built up by the phosphorlyated adaptor for RNA export (PHAX), the export receptor chromosome region maintenance-1 (CRM1), the GTP-bound form of Ran GTPase and the cap-binding complex (CBC), which recognizes and binds 7-methyl-guanosine(m^7G)-labeled RNA.

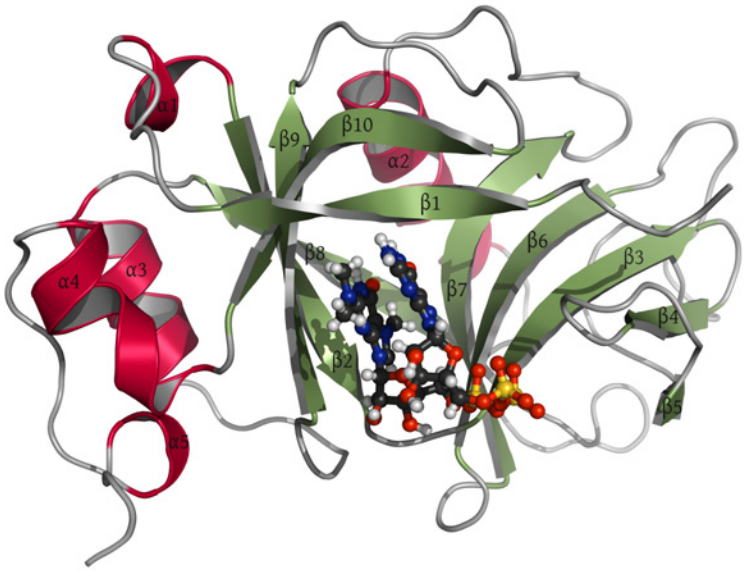

Fig. 5.2: Crystal structure of human snurportin 1 with bound m_3GpppG ligand (1XK5). The α-helices are colored in red, β-sheets in green and loop-regions are colored in grey. m_3GpppG is shown in a ball-and-stick model.

This hypermethylation is the key step to trigger the reimport of the cytosolic modified RNA and is carried out by the TGS1 protein. The importing complex consists of the survival motor neuron (SMN) complex and snurportin 1 (SPN), binding to the 2,2,7-trimethyl-guanosine-capped RNA (m_3G-RNA).[121,122] The proteins involved in binding to

these different caps are highly selective.

Recent experiments,[124] as well as the crystal structure of snurportin 1[104] (Fig. 5.2) suggest that the 2,2,7-trimethyl-guanosine-cap dinucleotide (m_3GpppG) prevents the binding of m_3G-capped U snRNA to snurportin 1 with a similar affinity and therefore is an intrinsic inhibitor candidate. Furthermore, it is a reasonable model system to investigate the effects of hypermethylation of the m^7G-capped RNA on binding to snurportin 1.

Unexpectedly, the hypermethylated cap binds better to snurportin 1 than the methylated. Despite the ability of an amino group to be both donor and acceptor for hydrogen bonding, the binding affinity of m^7GpppG (Fig. 5.3B), in contrast to m_3GpppG (Fig. 5.3A), could not be measured yet. Strasser et al. suggested the entropic penalty of the watershell in

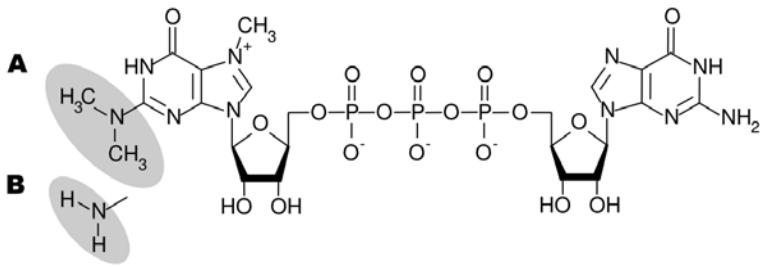

Fig. 5.3: Chemical structure of m_3GpppG (**A**) and m^7GpppG (**B**). The difference of both molecules is shown in the grey ellipsoids, where **A** represents the N2-nitrogen in the hypermethylated and **B** in the non-methylated state.

the vicinity of the free ligand to be the driving force of ligand-binding in the case of snurportin 1. The effective "shielding" of the hypermethylated guanosine-cap by a tryptophane residue of the protein therefore is a reasonable explanation for the observed behaviour of ligand binding and could be supported by mutation experiments.[104] Due to the low binding affinity of m^7GpppG to snurportin 1, a crystal structure of this complex could not be solved yet and the proof for this hypothesis on the structural level is still missing. As for the m^7GpppG/snurportin 1 complex, a crystal structure for the ligand-free SPN is still not available. The latter was suggested to be due to an effect of the dinucleotide on the structural integrity of the protein. Additionally, an unusually highly twisted conformation of

the β-strand 1 (Fig. 5.2), containing the "cap-shielding" tryptophane residue was observed, which enforces the assumption of an enhanced flexibility of snurportin 1 with no ligand bound.[104] To address these issues, insight into the dynamics of the protein upon ligand unbinding would be very interesting and helpful for the understanding of the problems occuring in the crystalization process.

In our work we investigate the dynamics of the protein upon ligand unbinding with the help of molecular dynamics simulations. To gain insight into the dynamics and overall flexibility of the ligand-free snurportin 1, we compute a trajectory of the protein without ligand. From this trajectory the globular motions as well as the dynamics of several amino acids in the binding pocket and the C-terminal region are investigated in more detail. By analyzing the watershell in the vicinity of the two methyl groups added to the methylated cap in solvent, as well as bound to the protein, we intend to gain insight in the contribution of the protein as a "shielding" factor of water from the ligand. We estimate the difference in binding free energy of m_3GpppG and m^7GpppG together with the enthalpic contributions to obtain evidence whether the binding process is driven either enthalpically or entropically.

5.2 Methods

We used the snurportin 1 structure 1XK5[104] from the PDB database[10] (Fig. 5.2) as starting structure and the AMBER99 forcefield[58,57] for our MD-simulations. The force field parameters for the ligands as in chapter 3 m_3GpppG and m^7GpppG were used.

All simulations were treated with the parameters as given in Chapter 2.2. The simulations were carried out in explicit solvent with the TIP4P water model[107] and a 150 mmol NaCl salt-concentration to mimic a physiological environment. We performed MD simulations of m_3GpppG bound to snurportin 1 with a total length of 650 ns, as well as ten 50 ns simulations of m^7GpppG. Additionally, 20 ns of each ligand in solvent and 6 trajectories with varying length (see Tab. 5.1) of the protein structure without ligand.

Trajectory #	1	2	3	4	5	6
Length (ns)	634	640	527	641	557	551

Tab. 5.1: Trajectory length of ligand-free snurportin 1 simulations

To get information about the overall stability and the changes in amino acid-mobility of the protein between the ligand-bound and -unbound systems, we calculated the root mean square deviation (RMSD) of the ligand-free protein, as well as with m_3GpppG and m^7GpppG bound to snurportin 1 along the respective trajectories. Furthermore, the backbone RMSD of every single amino acid was calculated to characterize relaxation motions upon ligand removal.

To quantify the intrinsic flexibility of the ligand-free structure of snurportin 1, we performed principal component analysis (PCA)[94] on the 650 ns equilibration trajectory of the snurportin 1-complex structure as well as on the trajectories 1, 2 and 6 of the protein without ligand. Five representative parts with a length of 5 ns each from the trajectories were taken at 50, 150, 250, 350, and 450 ns and the backbone atoms were used for the PCA. Because of the high and presumably functionally irrelevant fluctuations of the truncated termini, ten residues from both, the N- and C-terminus, were excluded from the PCA. All twenty 5 ns trajectory segments of the four simulations mentioned above were concatenated and subjected to one single PCA.

Furthermore, the distribution of water molecules in the vicinity of m_3G and m^7G, which is the only chemical difference in the ligand molecules, was analyzed in solution and bound to the protein. To this end, we extracted the water molecules from the trajectories of both ligands in pure solvent (20 ns each) and in solvated protein environment (50 ns for state A and 500 ns for state B) in a sphere with a radius of 1 nm around the N2-atom of the mono- and the trimethylated Guanine-nucleoside (see Fig. 5.3). To obtain the density distribution of water molecules, a three-dimensional grid, consisting of 100 bins in each dimension, was laid upon the spatial coordinates of the oxygen atoms of the water molecules and smoothed with a three-dimensional gaussian function of 0.01 nm width, which was chosen to trade

off resolution and statistical noise.[125]

To address the question whether the binding of m_3GpppG is mainly driven either enthalpically or entropically, we calculated the potential energies from the equilibrium simulations of m_3GpppG and m^7GpppG in pure solvent and bound to the protein. These potential energies are an estimate for the enthalpy (ΔH) of the system and were used together with the Gibbs' free energy (ΔG) from the free energy calculations in chapter 3 to estimate the entropic contribution ($T\Delta S$). To compute a reasonable error estimate for ΔH, we computed the standard error via an autocorrelation analysis of the trajectories, taking into account the underlying statistical uncertainty for a time series of correlated measurements, introduced by.[126,127]

5.3 Results

Root Mean Square Deviations To obtain quantitative informations of the difference in flexibility of the protein bound to m_3GpppG and without a ligand, the RMSD was calculated for the ligand-free and the m_3GpppG-bound trajectories. After the usual fast increase within the first few ns due to thermal fluctuations, the RMSD of the ligand-bound and of five ligand-free trajectories stays below 3 Å.

As can be seen in Fig. 5.4, in one trajectory (blue curve), the system rapidly escapes from the initial minimum towards a different minimum with an RMSD of 4 Å, whereas in another trajectory (red curve), the system stays in the first minimum for 300 ns. To identify the regions in the structure which are mainly involved in the destabilization motions upon ligand removal, the backbone RMSD for each amino acid was calculated. Figure 5.5 shows the time resolved backbone RMSD for each amino acid in the structure of snurportin 1 bound to m_3GpppG and without ligand. Since the termini exhibit an intrinsically high RMSD, 10 amino acid from each terminus were excluded from this analysis. To improve the statistics, the RMSD values from the six ligand free trajectories were averaged.

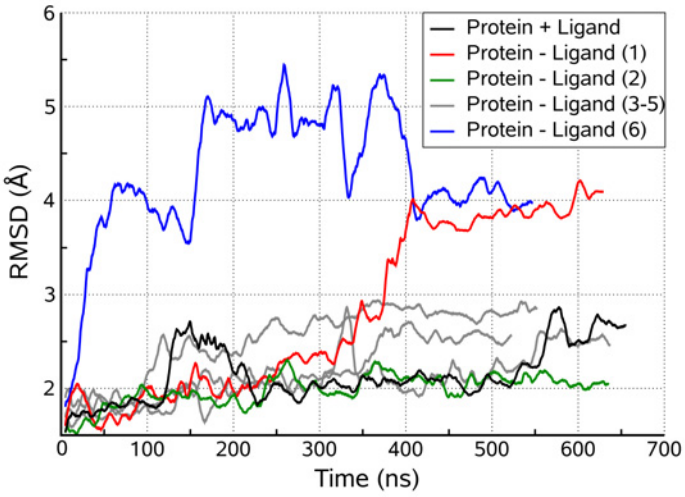

Fig. 5.4: Backbone root mean square deviations (RMSD) of snurportin 1 with and without m_3GpppG ligand. The black curve denotes the RMSD of the ligand bound protein. The remaining curves show the RMSD of six independent trajectories of snurportin 1 without a ligand. The trajectories with the highest RMSD are shown in red and blue; the most stable trajectory is shown in green.

As can be seen, only few, local regions contribute markedly to the observed flexibility. Figure 5.6 highlights in color these flexible regions. In the protein without a ligand, they are much more pronounced than in the ligand bound complex, but occur in similar regions. Accordingly, destabilization starts from enhancing fluctuations of the equilibrium motions, and not by an onset of new motions. This result is supported and quantified in Fig. 5.5 C, which shows a correlation coefficient of $r^2 = 0.54$.

The largest destabilization motions are seen in the C-terminal domain (Fig. 5.6, blue). Indeed, closer inspection of the trajectories reveals a structural rearrangement in that part of the protein upon ligand removal. Furthermore, a region of the β10-strand and adjacent loops become more flexible (Fig. 5.6, red). One further region of high flexibility is a solvent exposed loop built up from residues 161-167, shown in green in Figure 5.6. In contrast to the other flexible regions, which show enhanced flexibility upon ligand removal, this loop shows a similar flexibility in the ligand-bound structure. Furthermore, a small loop region, containing LYS144 (Fig. 5.6, yellow), exhibits a larger flexibility upon removal of the

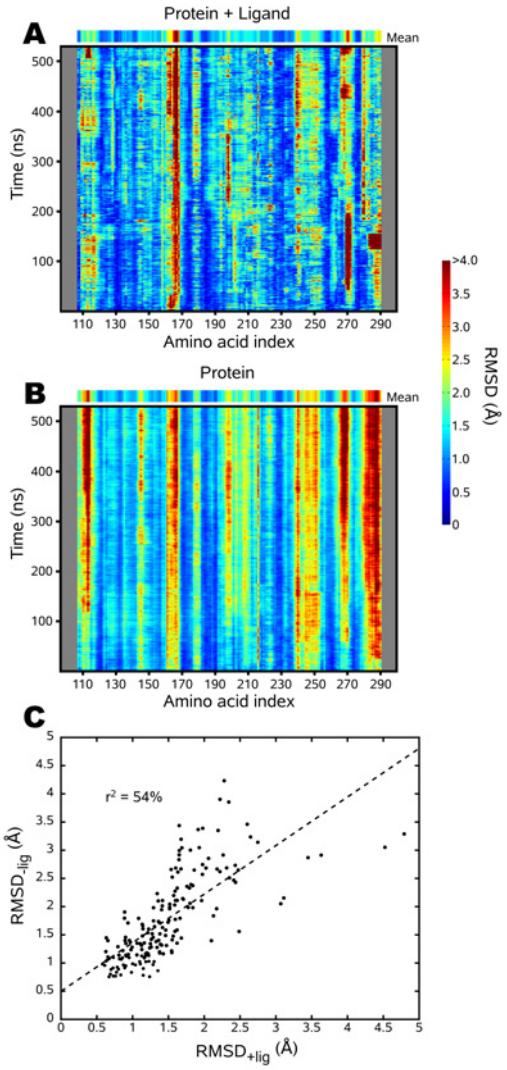

Fig. 5.5: Time resolved backbone RMSD for each amino acid. **A** shows the RMS deviations of SPN bound to m_3GpppG. **B** shows the ligand free protein, where the RMSD from all six trajectories was averaged. The first ten amino acids from each terminus have been removed in this analysis. For a better resolution in the lower RMSD regions, all values above 4 Å have been truncated to this value. **C**: Correlation of RMSD. The RMSD of each amino acid from the ligand-free trajectories (-lig) is plotted against those of the ligand-bound (+lig). The dashed line is the linear fit to the data points. The correlation coefficient from the linear regression of the data points is r^2=54%.

ligand. In the bound state, LYS144 interacts with the phosphate backbone of m_3GpppG via a saltbridge.

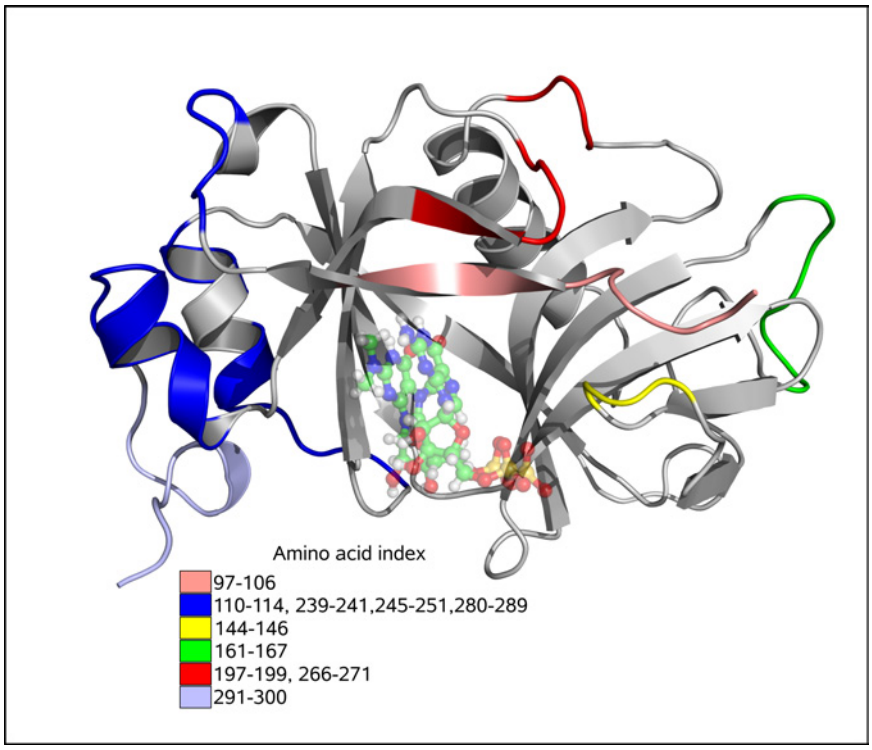

Fig. 5.6: Color coded structure of SPN. Selected high RMSD amino acids from the ligand free SPN trajectories are colored according to their position in the structure. Blue: C-terminal domain, Red: β10-related region, Green: Loop region, Yellow: LYS144-Loop, Pink: N-terminus, Lightblue: C-terminus, Tranparent: m_3GpppG (for guidance).

Principal Component Analysis Principal Component Analysis (PCA) was used to compare the global motions of ligand-free and m_3GpppG-bound snurportin 1 in a common subspace. Figure 5.7 shows the projection of four trajectories onto eigenvectors 1 and 2 of this subspace. As can be seen (Fig. 5.7, inset), these eigenvectors describe already 48% of the atomic motion.

The system with m_3GpppG bound to SPN (black cloud) remains close to the x-ray structure (yellow dot), with rare transient transitions to an adjacent shallow minimum. In contrast, removal of the ligand from the original structure leads to an extensive sampling of phase

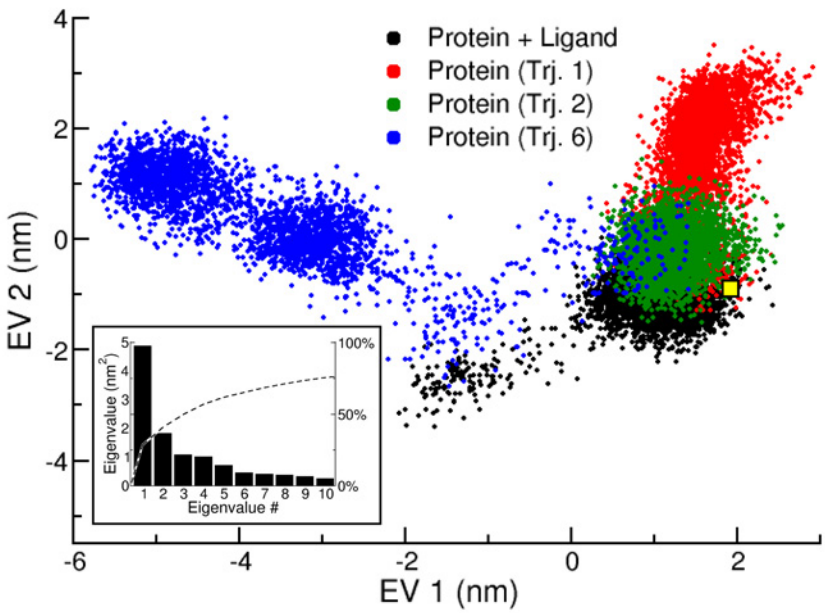

Fig. 5.7: PCA of snurportin 1. The black cloud resembles the 650 ns trajectory of the protein with m_3GpppG ligand bound, projected onto the first two eigenvectors, and the yellow square as the general starting configuration. The clouds colored in red, green and blue display the trajectories 1, 2, and 6 of the protein without ligand as in Fig. 5.4. Every 10th frame of the respective trajectories has been used in the projections. Inset: The first 10 eigenvectors of the covariance matrix. The dashed line is the cumulative sum of the contribution to the total fluctuations. The first two eigenvectors describe 48% of the main global motion.

space until the system reaches different local minima on the energy landscape (red and blue dots in Fig. 5.7). This drift motion was observed for two out of six trajectories. One trajectory of the remaining four is shown (green dots in Fig. 5.7).

The projection of the ligand-free trajectory of SPN onto its first two principal components was used to select structures for detailed analysis. The largest motions were seen for trajectory 6 (blue) cloud) which, therefore, was chosen for closer analysis. Accordingly, two further snapshots from the trajectory were chosen. These snapshots have been selected because they are close to the center of the respective substate.

We first investigated the amino acids in direct interaction with the m_3GpppG ligand. In the bound state t=0 ns, the N7-methyl-group of the m_3G-nucleobase is buried in a hydrophobic pocket, build by the residues CYS124, ILE175, LEU186 and LEU264. After removal of the

ligand at 79.7 ns, this hydrophobic pocket is exposed to the surrounding water. TRP276, moves towards the pocket, undergoing a local hydrophobic collapse. Additionally, the flexibility of LYS144 increased due to the lack of the ligand as interaction partner (Fig. 5.5 and 5.6). By moving closer to ASP173, LYS144 weakens the ionic interaction between ARG129 and ASP173. As a condequence, ARG129 can detach from ASP173 and form an new, π-stackinginteraction to TRP276. This structural rearrangement is supported by a motion of TRP276 into the binding pocket (Fig. 5.8 A).

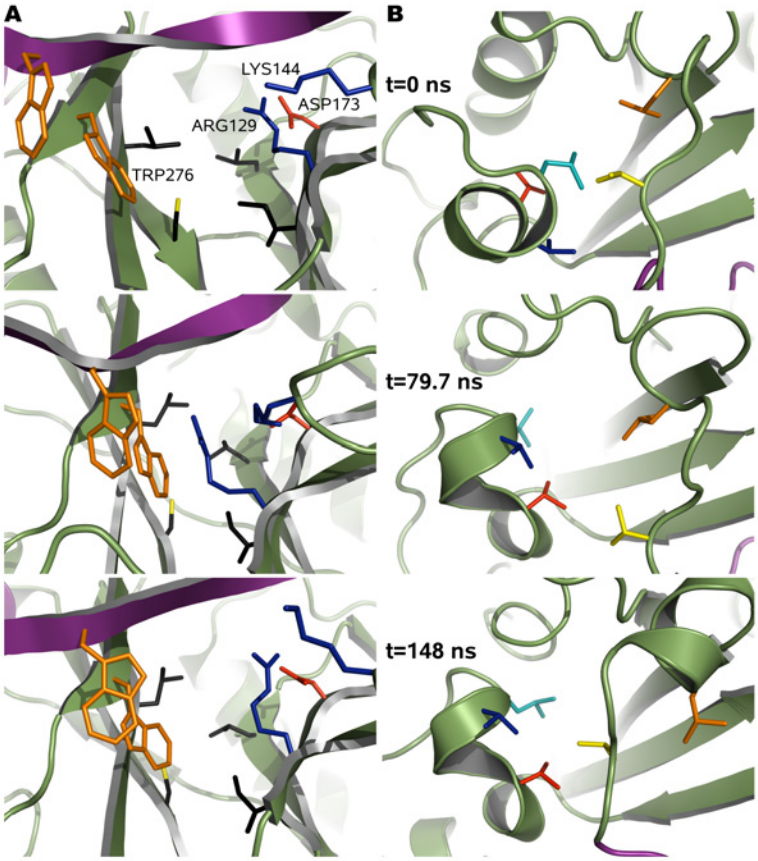

Fig. 5.8: Snapshots of the binding pocket and C-terminus of snurportin 1 at 0, 79.7 and 148 ns. **A**: TRP107 and 276 are shown in orange, the residues CYS124, ILE175, LEU186 an LEU258, building a hydrophobic pocket, in black, ARG129 and LYS144 in blue and ASP173 in red. The β-strand 1 is colored purple. **B**: VAL111(yellow), LEU115(orange), VAL282(red), VAL285(blue) and LEU286(light-blue).

Further rearrangement is seen for the N-terminal β-strand β1 (Fig. 5.8 A, purple). After

removal of the ligand from the binding pocket, a relaxation of this originally twisted β-strand is observed. TRP107 moves into the now unoccupied binding pocket, losening the strain on β-strand $\beta 1$. This movement leads to a loss of structural stability of the β-sheet (β-strands $\beta 1$ and $\beta 10$), resulting in a further distortion of $\beta 10$. This finding confirms and explains the suggestion by Strasser et al.,[104] that this β-sheet should untwist upon ligand removal.

After 148 ns no major structural changes are seen in the vicinity of the binding pocket. One exception is LYS144, which moves to a purely solvent interacting position, which leads to a loss of the saltbridge between LYS144 and ASP173. This is compensated by reformation of the ionic interaction of ARG129 and ASP173, after moving away from its former cation-π-interaction-partner TRP276. In summary, a highly dynamic structure in the vicinity of the empty binding pocket is seen.

A second quite flexible region is seen near the C-terminus (Fig. 5.6, blue and 5.8 B). The previously described movement of TRP107 and TRP276 towards the binding pocket coincides with a shift of the hydrophobic residues VAL111, LEU115, VAL282, VAL285 and LEU286. These motions destabilize the hydrophobic region, which connects the C-terminal part to the rest of the protein, resulting in a higher flexibility of the C-terminal region. Remarkably, an α-helical structure of residues 113 to 118 is formed, followed by a reorientation of LEU115, turning away from the hydrophobic interaction region. After this transient rearrangement, the initial hydrophobic cluster is reformed, albeit with reduced stability.

Water Shell We now turn to the unexpectedly strong binding of the hypermethylated m_3GpppG cap. It has been suggested[104] that the entropic penalty of the watershell in the vicinity of the free ligand is the driving force for binding to snurportin 1 in order to test this hypothesis, we compared the solvation shell around the two ligands in solvent and bound to snurportin 1.

Upon binding of m_3GpppG to SPN, the volume of the solvation shell is significantly reduced

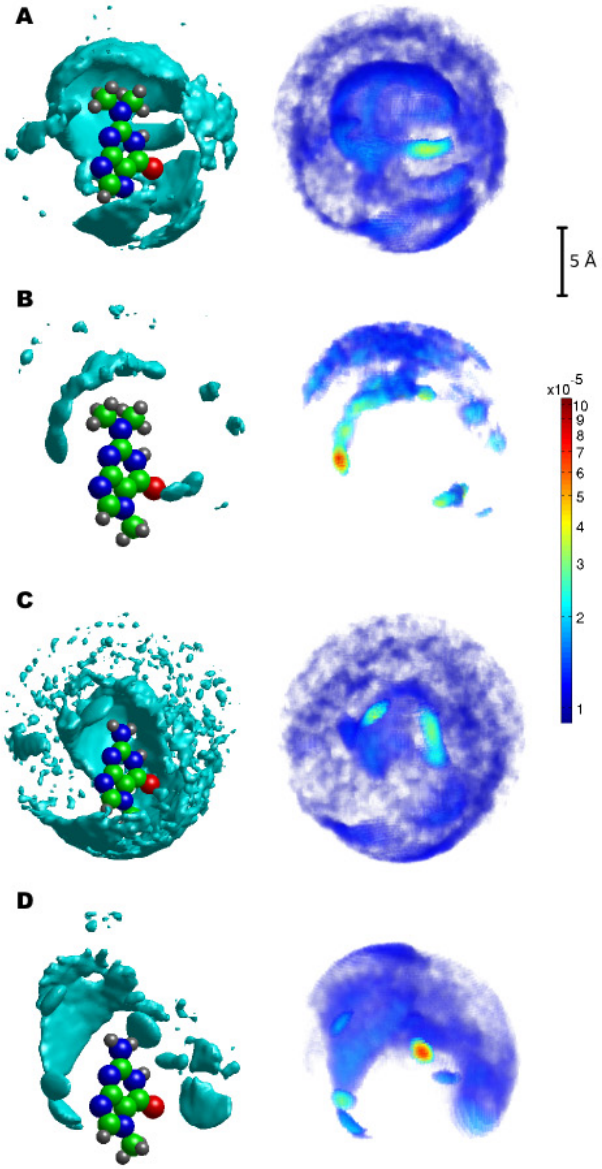

Fig. 5.9: 3-dimensional density distribution of water molecules around the m_3GpppG and m^7GpppG ligands. **A** and **B** show the m_3G-nucleoside in water and complexed to snurportin 1; **C** and **D** the m^7G-nucleoside, respectively. The left column shows 15% of the total solvation shell, starting from the highest to lower densities, as a qualitative isosurface view. The right column shows 60% of the density in a quantitative plot. For clarity of presentation, a logarithmic color scale was chosen.

(Fig. 5.9 A/B). This release of water molecules from the shell to the bulk is entropically favorable. Also for m^7GpppG (Fig. 5.9 C/D), a decrease of the volume of the solvation shell upon binding is seen. However, to a markedly smaller extent than for m_3GpppG . Accordingly, the entropy gain due to water release is larger for m_3GpppG than for m^7GpppG, rendering m_3GpppG binding favorable.

Moreover, one hydrogen bond of the N2 amino-group to water molecules is lost upon binding (Fig. 5.9 D). This is also evident from the density plots, which show two high-density peaks in Fig. 5.9 C, but only one in Fig. 5.9 D. These high-density peaks indicate the positions of the hydrogen-bonded water molecules. In contrast, bound to the protein, m^7GpppG exhibits only one stable hydrogen bond indicated by the high density peak (Fig. 5.9 D). The second donor formas hydrogen bond to the protein, which is less stable and, therefore, entails a small enthalpic loss upon ligand binding.

Both effects combine to and explain the unexpectedly strong bind of m_3GpppG. This result is further supported by recent mutation experiments,[104] which have shown that the binding affinity of SPN to m_3GpppG is significantly reduced when TRP107 is mutated into ALA107. This finding confirms the crucial role of TRP107 as key residue for the desolvation of the di-methylated N2-nitrogen of m_3GpppG. The flexibility of TRP107 should therefore be larger for bound m^7GpppG where shielding is less pronounced.

We tested this prediction by comparing the RMSD of TRP107 from ten 50 ns trajectories of m^7GpppG bound to SPN with the RMSD of the trajectory of m_3GpppG bound to SPN (Fig. 5.10). Indeed, 8 out of 10 show a significant increase of the RMSD (blue and red); four of these even very large (red). These large deviations thus indicate the loss of the hydrophobic interaction between the ligand and TRP107.

Binding thermodynamics To confirm that the observed effects in the simulations actually cause the selective SPN binding affinity, we have taken the binding free energy difference for the two ligands m_3GpppG and m^7GpppG, computed in chapter 3. We calculated via Crooks Gaussian Intersection (CGI) simulations the free energy differences

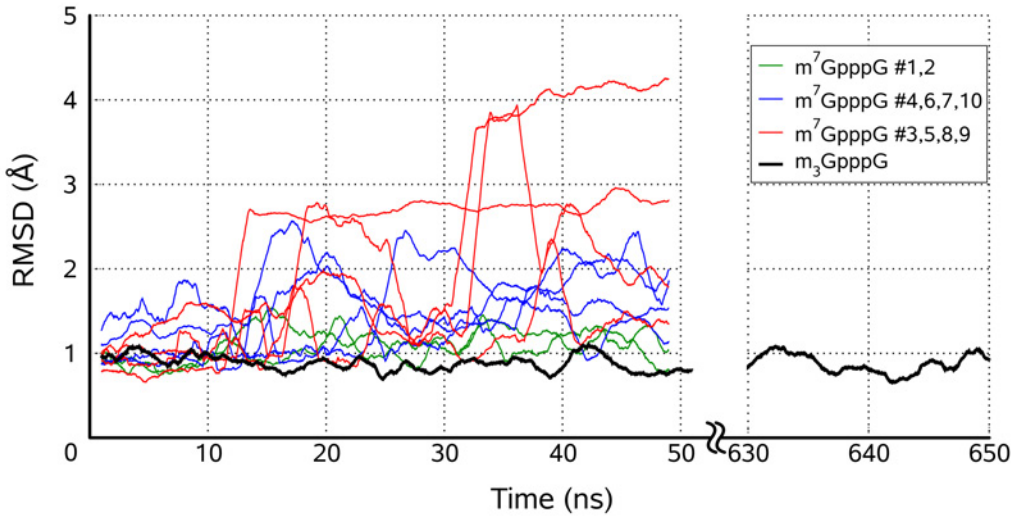

Fig. 5.10: Root mean square deviations of TRP107. The RMSD curves were smoothed with a running average, where 100 data points where used for averaging. The coloring indicates the deviation of TRP107 in trajectories with m^7GpppG, where green resembles a small, blue a medium and red a strong deviation. Black shows the deviation of TRP107 with m_3GpppG.

between m_3GpppG, bound to the protein (ΔG_b) and in solvent (ΔG_u), and m^7GpppG, respectively, as described in chapter 3, yielding ΔG_b=-430.34 ±0.86 kJ/mol and ΔG_u=-441.16 ±0.51 kJ/mol, respectively. With the thermodynamic cycle shown in Fig. 3.3 B, a binding free energy difference $\Delta\Delta G$=10.82 ±1.00 kJ/mol is obtained. Using the known dissociation constant[104] for m_3GpppG/SPN, K_D=1.00 ±0.03 μM, this free energy difference translates into K_D=72–152 μM for m^7GpppG/SPN. This value is consistent with the estimate of 100 μM–1000 μM derived from UV cross-linking studies.[109]

With this result at hand, the simulations serve to dissect the binding free energies into the corresponding enthalpic and entropic contributions. The enthalpic contributions ΔH were estimated from the averaged total energies of the respective simulations, and the entropic contributions $T\Delta S$ from $T\Delta S = \Delta H - \Delta G$. The obtained $\Delta\Delta H$=12.95 ±17.92 kJ/mol and $\Delta(T\Delta S)$=2.13 ±17.95 kJ/mol
(Tab. 5.2) suggest that the entropic and enthalpic contributions are similar within the obtained accuracy. Apparently, and in contrast to our initial expectation, the binding

selectivity (given by $\Delta\Delta G$) is not entropically driven, although the underlying free energy differences are.

System	ΔH	σ	ΔG	σ	TΔS	σ
Bound	418.74	4.50	-430.34	0.86	849.08	4.58
Unbound	405.79	17.35	-441.16	0.51	846.95	17.36
Diff.(ΔΔ)	12.95	17.92	10.82	1.00	2.13	17.95

Tab. 5.2: Enthalpy, free energy, and entropic contribution to ligand binding and respective differences. All units are in kJ/mol, temperature T is at 300 K, σ is the standard error. Differences refer to the two ligands, i.e., m_3GpppG - m^7GpppG.

5.4 Discussion & Conclusions

Several open questions concerning snurportin 1 have been addressed in this work. The m_3G-cap-binding domain of human snurportin 1 bound to the inhibitor m_3GpppG was chosen to investigate the effects of RNA-cap hypermethylation on binding to SPN. Extended molecular dynamics simulations of SPN were performed for the ligand-free protein as well as for the protein with bound ligands m_3GpppG and m^7GpppG.

Principal Component Analysis (PCA) and RMSD calculations were carried out on the trajectories with m_3GpppG bound to SPN, and with the ligand-free structure to reveal the ligand-dependent structural changes of the systems. Comparison of the complexes with simulations of the solvated two ligands showed that the solvation shell plays a crucial role for binding selectivity. Also, TRP107 was shown to be crucial for binding.

Our simulations, furthermore, served to study possible structural changes upon ligand removal for the apo protein construct, whereas the complex structure remained stable. Remarkably, a large fraction of this structural destabilization seems to be already contained in the equilibrium fluctuations of the stable complex. Indeed, in terms of RMSD values of the individual amino acids, a strong correlation of 54% is observed. The largest deviations occured in the C-terminal domain, in several residues next to the N-terminus, in a solvent

exposed loop, and in a small loop harbouring LYS144.

These large structural fluctuations also provide a likely explanation why crystallization of the apo protein was up to now unsuccessful,[104] and suggest possible constructs for further crystallization attempts. In particular, due to the observed motions of the amino acids in the binding pocket and in the C-terminal region, the K144A mutation as well as mutations of hydrophobic residues into polar ones in the C-terminal cluster, may lead to a more stable ligand-free protein construct.

PCA further served to characterize the observed structural changes. Compared to the m_3GpppG-bound dynamics, the ligand-free trajectories showed extensive sampling, additionally testifying its drastically and collective enhanced flexibility. Closer analysis of the ligand-free trajectory enabled us to characterize the amino acid rearrangements in the binding pocket and the C-terminal domain in more detail.

The destabilization of a hydrophobic cluster in the C-terminal region indicates a major conformational change upon ligand unbinding. Recent experiments have shown that the export receptor chromosome region maintenance-1 (CRM1) is highly competitive to m_3G-capped RNA in binding to SPN. Although the binding affinity of CRM1 to SPN was strongly dependent on the existence of the full N-terminal domain, the deletion of the C-terminus beyond residue 285 resulted in an affinity decrease of 60%.[128] The structural rearrangement observed here in the C-terminal domain upon ligand unbinding can explain both, the observed binding affinities of CRM1 as well as the competitive behavior to m_3G-capped RNA. Whether the observed structural rearrangements in the C-terminal domain occur in the truncated protein only or are maintained in the full length protein, can, due to lack of a suitable full length structure, not rigorously be decided.

In any case, these rearrangements provide a possible explanation for the *in vivo* binding mechanism leading to the cluster formation of an export complex consisting out of Ran-GTP, CRM1, and snurportin 1.

The role of the solvation shell in the binding thermodynamics is indeed remarkable. Both

ligands, m_3GpppG and m^7GpppG, are surrounded by an ordered watershell around the N2-nitrogen, differing in its methylation state. In contrast, when bound to SPN, the volumes of the remaining water shells are quite different for the two ligands. As a result, the protein shields m_3GpppG to a much larger extent than m^7GpppG, from the surrounding water.

As revealed by its differential flexibility, TRP107 appears to be crucial for this shielding. Indeed, the W107A mutant has been shown to bind m_3GpppG with markedly reduced binding affinity.[104] Taken together, we suggest TRP107 as the key residue for the shielding of the two N2-methyl-groups in m_3GpppG.

To validate our simulations, binding free energy differences between m_3GpppG and m^7GpppG were taken from chapter 3. From these calculations, the binding affinity of m^7GpppG to SPN, quantified by the equilibrium dissociation constant K_D, is reduced by approximately two orders of magnitude with respect to m_3GpppG. This result agrees well with estimates from UV cross-linking experiments,[109] for which a decrease by 2–3 orders of magnitude is reported. Further splitting into entropy and enthalpy shows that the desolvation of the di-methylated N2-nitrogen is indeed the main driving force for the better affinity of m_3GpppG to SPN.

In summation, the free energy calculations support the experimental findings of Huber et al.[109] as well as Strasser et al.[104] for the selective binding of m_3GpppG which mimics the m_3G-cap as an important part of the nuclear localization signal specific for UsnRNP nuclear import[129].[130]

6

Summary, Conclusion, and Outlook

Despite a considerable amount of studies in the field of free energy calculations from molecular dynamics simulations using a wide array of methods, it was still difficult to judge how these methods perform for larger and more complex systems such as biological macromolecules.

One aim of this thesis was to carry out a comprehensive evaluation of these methods for test systems at different levels of complexity, and, guided by the obtained insights, to develop an improved and efficient new method suitable to compute sufficiently accurate free energy differences for the binding of complex ligands such as peptides for qualitative prediction of binding behaviour.

To this end, a comprehensive evaluation of methods to compute free energy differences, including Slow Growth Thermodynamic Integration (SGTI), Discrete Thermodynamic Integration (DTI), Jarzynski's work averaging (EXP), and Bennett's Acceptance Ratio (BAR), has been carried out for three test systems at different levels of complexity. For this evaluation we used the interconversion of ethane to methanol, of tryptophane to glycine in a tripeptide and of m_3GpppG to m^7GpppG in the globular protein snurportin 1.

Due to the accuracy of the free energy calculations performed with the established methods, as well as due to problematic error estimates provided by some of these methods, we developed a new method to calculate free energies from non-equilibrium trajectories, Crooks Gaussian Intersection (CGI), and tested this method with the same systems. This new method was shown to accurately compute the solvation free energy difference between methanol and ethane, compared to experiments. Moreover, the calculated binding affinity difference between m_3GpppG and m^7GpppG to snurportin 1 was in very good agreement with estimates from experiments.

In general, the non-equilibrium methods were shown to outperform the traditional equilibrium methods. The best results were obtained for the newly proposed CGI method. Furthermore, the advantage of extended parallel computations, inherent to all non-equilibrium methods, is also fully exploited by CGI.

The fast and precise computation of binding free energy differences for protein-ligand complexes is an important task for the development of specific drugs or vaccines in pharmaceutical applications. To evaluate, if sufficiently accurate results by the CGI method can also be obtained for large sidechain perturbations *inside* a large globular protein, the binding affinities for eight mutants of the influenza hemagglutinin 307-319 peptide to a human MHC Class II HLA-DR1 protein were computed.

The computed binding affinity differences for these mutants to the wildtype peptide agreed semi-quantitatively with those from experiments. Moreover, we predict the binding affinities of two mutants which could not be quantified precisely in experiments. Hence, CGI free energy simulations can semi-quantitatively distinguish between better or worse binders in comparsion to the wildtype. Compared to complementary docking approaches, the CGI approach has two major advantages. First, the free energy difference is obtained by first principles free energy calculations, and, secondly, it can be further optimized via longer simulation times.

A second aim of this thesis was to investigate the dynamics of the globular protein snurportin 1 upon ligand unbinding, and to get insights into the role of the solvation shell around the hypermethylated RNA-cap upon binding to snurportin 1. Extensive simulations of the two ligands m_3GpppG and m^7GpppG bound to snurportin 1 as well as the ligand free protein were carried out. These simulations, together with the free energy differences computed in Chap. 3, allowed to conclude that the binding selectivity (given by $\Delta\Delta G$) is not entropically driven, although the underlying free energy differences are. Additionally, the globular motions as well as the motions of single amino acids in the binding pocket were investigated.

The simulations of the ligand free protein also served to study possible structural changes upon ligand removal for the apo protein construct. Remarkably, a large fraction of this structural destabilization seems to be already contained in the equilibrium fluctuations of the complex which remained stable in our simulations. These large structural fluctuations provide a likely explanation why crystallization of the apo protein was up to now unsuc-

cessful. This result enables us to suggest possible constructs for further crystallization attempts of an apo protein construct.

The role of the solvation shell in the binding thermodynamics was found to be indeed remarkable. Both ligands, m_3GpppG and m^7GpppG, are surrounded by an ordered water-shell around the N2-nitrogen, differing in its methylization state. In contrast, when bound to SPN, the volumes of the remaining water shells are quite different for the two ligands. As a result, the protein shields m_3GpppG from the surrounding water to a much larger extent than m^7GpppG. In this context, TRP107 appears to be crucial for this shielding as revealed by its differential flexibility, and can therefore be suggested as the key residue for the shielding of the two N2-methyl-groups of m_3GpppG when bound to snurportin 1. The splitting of the thermodynamic observables into entropy and enthalpy shows that the desolvation of the di-methylated N2-nitrogen is indeed the key aspect for the better affinity of m_3GpppG to SPN.

The investigation of the binding free energies, and the development of the new CGI method to compute these efficiently and reliably, significantly extends the applicability of MD simulations towards pharmaceutical applications. Moreover, extensive exploration of the dynamics of proteins with and without bound ligands by means of MD simulations yielded a microscopic picture of the underlying structural mechanisms of complex formation and therefore contributed to the understanding of regulative mechanisms in cells.

Although the applications presented in this thesis clearly represent promising first steps for sufficiently accurate free energy calculations, there are still questions which cannot be answered by all-atom MD simulations yet. One aspect is the calculation of absolute binding free energies as well as the precise entropy estimation. Another one is the well known sampling problem, which restricts the interpretations made with the help of simulations mostly to one or rarely few minima on the free energy landscape.

Therefore, a main future task is to extend this sampling to enable the calculation of quantitatively correct free energy differences.

A large variety of systems can be addressed with the CGI method. One very interesting application is, e.g., the computation of free energy differences between DNA-binding proteins, like transcription factors, to differing DNA-sequences, or the yet not properly solved calculation of free energy differences between ATP and ADP in proteins like the F1-ATPase. However, it is still an open question which perturbation size in a simulation system will limit our current free energy calculations with the CGI method, although the free energy calculations made for the MHC Class II protein in Chap. 4 hint towards such limits. However, improved sampling of the involved systems in these calculations will likely provide more accurate, and possibly quantitative results.

With the continuous increase in computational power, simulations will eventually be able to sample systems with millions of atoms on a microsecond timescale in a reasonable amount of computer time. Such extended simulations, as suggest by our evaluations in Chap. 3, would improve the computation of thermodynamic properties like the entropy and the absolute free energy, and will inspire the development of further and even more accurate methods to address these biologically most relevant properties.

7

Appendix

Parameters for m_3GpppG and m^7GpppG

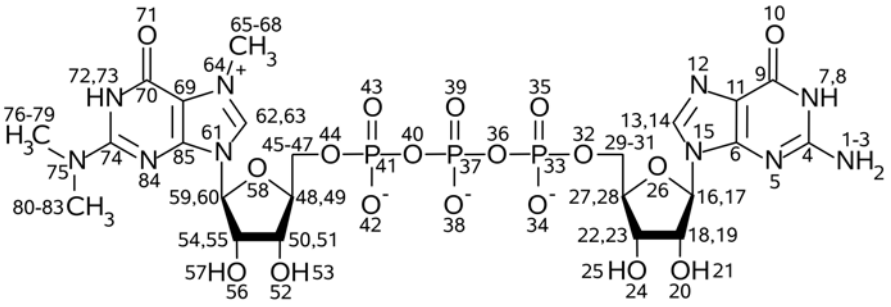

Fig. 7.1: Schematic drawing of m_3GpppG.

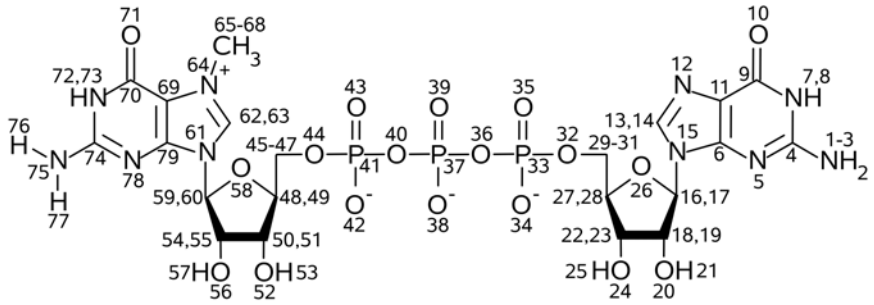

Fig. 7.2: Schematic drawing of m^7GpppG.

Element	type	mass	LJ σ	LJ ϵ
Carbon	amber99_98	12.01	3.39967e-01	4.57730e-01
Nitrogen	amber99_99	14.01	3.25000e-01	7.11280e-01

Tab. 7.1: Mass and Lennard-Jones (LJ) parameters for new amber99 force field atom types.

m_3GpppG/m^7GpppG							
Atom #	name	type	charge	Atom #	name	type	charge
1	N2	amber99_38	-0.9672	23	H3*	amber99_19	0.0615
2	H21	amber99_17	0.4364	24	O3*	amber99_43	-0.6541
3	H22	amber99_17	0.4364	25	HO*3	amber99_25	0.4376
4	C2	amber99_3	0.7657	26	O4*	amber99_44	-0.3548
5	N3	amber99_37	-0.6323	27	C4*	amber99_11	0.1065
6	C4	amber99_4	0.1222	28	H4*	amber99_19	0.1174
7	N1	amber99_35	-0.4787	29	C5*	amber99_11	0.0558
8	H1	amber99_17	0.3424	30	H5*1	amber99_19	0.0679
9	C6	amber99_2	0.4770	31	H5*2	amber99_19	0.0679
10	O6	amber99_41	-0.5597	32	O5*	amber99_44	-0.5987
11	C5	amber99_4	0.1744	33	PAT	amber99_46	1.2532
12	N7	amber99_36	-0.5709	34	OAU	amber99_45	-0.8799
13	C8	amber99_6	0.1374	35	OAS	amber99_45	-0.8799
14	H8	amber99_24	0.1640	36	OAX	amber99_44	-0.4831
15	N9	amber99_40	0.0492	37	PAZ	amber99_46	1.5851
16	C1*	amber99_11	0.0191	38	OBA	amber99_45	-0.8894
17	H1*	amber99_20	0.2006	39	OAY	amber99_45	-0.8894
18	C2*	amber99_11	0.0670	40	OBD	amber99_44	-0.4831
19	H2*1	amber99_19	0.0972	41	PBK	amber99_46	1.2532
20	O2*	amber99_43	-0.6139	42	OBJ	amber99_45	-0.8799
21	HO*2	amber99_25	0.4186	43	OBL	amber99_45	-0.8799
22	C3*	amber99_11	0.2022				

Tab. 7.2: Amber99 common force field parameters of m_3GpppG and m^7GpppG (see Fig. 7.1&7.2).

m_3GpppG							
Atom #	name	type	charge	Atom #	name	type	charge
44	OBU	amber99_44	-0.5987	65	CBC	amber99_11	-0.1996
45	CBT	amber99_11	0.0558	66	HBC1	amber99_19	0.1492
46	HBT1	amber99_19	0.0679	67	HBC2	amber99_19	0.1492
47	HBT2	amber99_19	0.0679	68	HBC3	amber99_19	0.1492
48	CBS	amber99_11	0.1065	69	CBG	amber99_4	-0.0086
49	HBS	amber99_19	0.1174	70	CBF	amber99_2	0.5158
50	CBY	amber99_11	0.2022	71	OBB	amber99_41	-0.5055
51	HBY	amber99_19	0.0615	72	NBE	amber99_35	-0.4104
52	OCB	amber99_43	-0.6541	73	HBE	amber99_17	0.3284
53	HCB	amber99_25	0.4376	74	CBM	amber99_3	0.5142
54	CBX	amber99_11	0.0670	75	NBV	amber99_38	-0.1392
55	HBX	amber99_19	0.0972	76	CBW	amber99_98	-0.1783
56	OCA	amber99_43	-0.6139	77	HBW1	amber99_19	0.1129
57	HCB	amber99_25	0.4186	78	HBW2	amber99_19	0.1129
58	OBR	amber99_44	-0.3548	79	HBW3	amber99_19	0.1129
59	CBQ	amber99_11	0.0191	80	CBZ	amber99_98	-0.1783
60	HBQ	amber99_20	0.2006	81	HBZ1	amber99_19	0.1129
61	NBP	amber99_38	0.1804	82	HBZ2	amber99_19	0.1129
62	CBI	amber99_10	-0.0231	83	HBZ3	amber99_19	0.1129
63	HBI	amber99_24	0.2450	84	NBN	amber99_37	-0.5167
64	NBH	amber99_99	-0.0480	85	CBO	amber99_4	0.1801

Tab. 7.3: Amber99 force field parameters of m_3GpppG (see Fig. 7.1).

m^7GpppG							
Atom #	name	type	charge	Atom #	name	type	charge
44	OBU	amber99_44	-0.5987	62	CBI	amber99_10	-0.0231
45	CBT	amber99_11	0.0558	63	HBI	amber99_24	0.2450
46	HBT1	amber99_19	0.0679	64	NBH	amber99_99	-0.0480
47	HBT2	amber99_19	0.0679	65	CBC	amber99_11	-0.1996
48	CBS	amber99_11	0.1065	66	HBC1	amber99_19	0.1492
49	HBS	amber99_19	0.1174	67	HBC2	amber99_19	0.1492
50	CBY	amber99_11	0.2022	68	HBC3	amber99_19	0.1492
51	HBY	amber99_19	0.0615	69	CBG	amber99_4	-0.0086
52	OCB	amber99_43	-0.6541	70	CBF	amber99_2	0.5158
53	HCB	amber99_25	0.4376	71	OBB	amber99_41	-0.5055
54	CBX	amber99_11	0.0670	72	NBE	amber99_35	-0.4104
55	HBX	amber99_19	0.0972	73	HBE	amber99_17	0.3284
56	OCA	amber99_43	-0.6139	74	CBM	amber99_3	0.7902
57	HCB	amber99_25	0.4186	75	NBV	amber99_38	-0.9672
58	OBR	amber99_44	-0.3548	76	HBV1	amber99_17	0.4364
59	CBQ	amber99_11	0.0191	77	HBV2	amber99_17	0.4364
60	HBQ	amber99_20	0.2006	78	NBN	amber99_37	-0.5167
61	NBP	amber99_38	0.1804	79	CBO	amber99_4	0.1801

Tab. 7.4: Amber99 force field parameters of m^7GpppG (see Fig. 7.2).

atom i	atom j	length (nm)	fc
amber99_10	amber99_99	0.13910	344008.5
amber99_10	amber99_38	0.13640	375723.2
amber99_4	amber99_38	0.13640	375723.2
amber99_98	amber99_19	0.10900	284512.0
amber99_98	amber99_38	0.14630	282001.6
amber99_4	amber99_99	0.13740	364844.8
amber99_11	amber99_99	0.14750	282001.6

Tab. 7.5: Amber99 force field bond parameters. fc denotes the force constant for bond stretching in kJ mol^{-1} nm^{-2}.

atom i	atom j	atom k	angle (°)	fc
amber99_24	amber99_10	amber99_99	123.050	418.400
amber99_24	amber99_10	amber99_38	123.050	418.400
amber99_38	amber99_10	amber99_99	113.900	585.760
amber99_4	amber99_99	amber99_10	103.800	585.760
amber99_4	amber99_38	amber99_10	105.400	585.760
amber99_4	amber99_4	amber99_99	110.400	585.760
amber99_4	amber99_4	amber99_38	106.200	585.760
amber99_10	amber99_99	amber99_11	128.800	585.760
amber99_10	amber99_38	amber99_11	128.800	585.760
amber99_4	amber99_99	amber99_11	125.800	585.760
amber99_4	amber99_38	amber99_11	125.800	585.760
amber99_2	amber99_4	amber99_99	130.000	585.760
amber99_38	amber99_4	amber99_37	126.200	585.760
amber99_3	amber99_38	amber99_98	120.000	418.400
amber99_19	amber99_98	amber99_19	109.500	292.880
amber99_38	amber99_11	amber99_44	109.500	418.400
amber99_20	amber99_11	amber99_38	109.500	418.400
amber99_19	amber99_11	amber99_99	109.500	418.400
amber99_98	amber99_38	amber99_98	120.000	292.880
amber99_19	amber99_98	amber99_38	109.500	418.400

Tab. 7.6: Amber99 force field angle parameters. fc denotes the force constant for angle bending in kJ mol^{-1} rad^{-2}.

atom i	atom j	atom k	atom l	C_0	C_1	C_2	C_3	C_4	C_5
X	a99_10	a99_38	X	14.22560	0.00000	-14.22560	0.00000	0.00000	0.00000
X	a99_10	a99_99	X	10.46000	0.00000	-10.46000	0.00000	0.00000	0.00000
X	a99_11	a99_99	X	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
X	a99_4	a99_99	X	21.33840	0.00000	-21.33840	0.00000	0.00000	0.00000
X	a99_98	a99_38	X	2.51040	7.53120	0.00000	-10.04160	0.00000	0.00000
a99_19	a99_11	a99_38	a99_10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
a99_19	a99_11	a99_99	a99_10	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
a99_35	a99_3	a99_99	a99_11	20.08320	0.00000	-20.08320	0.00000	0.00000	0.00000
a99_38	a99_10	a99_99	a99_11	15.48080	0.00000	-15.48080	0.00000	0.00000	0.00000
a99_10	a99_38	a99_4	X	17.57280	0.00000	-17.57280	0.00000	0.00000	0.00000
X	a99_11	a99_38	X	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
X	a99_38	a99_10	X	5.02080	0.00000	-5.02080	0.00000	0.00000	0.00000
X	a99_38	a99_4	X	5.02080	0.00000	-5.02080	0.00000	0.00000	0.00000
X	a99_38	a99_11	X	5.02080	0.00000	-5.02080	0.00000	0.00000	0.00000

Tab. 7.7: Amber99 force field Ryckaerd-Bellemans dihedral parameters. All constants C are given in kJ mol^{-1}. X denotes any atom type and amber99 is abbreviated a99. The upper part contains general parameters and the lower part parameters for specific atom combinations as given in Ref. aduri2007aff.

Danksagung

First, I want to thank Prof. Helmut Grubmüller for the opportunity to join his biophysical research group at the MPI for Biophysical Chemistry as a biologist, and for the extensive and great support during my time as a PhD student. Helmut was always there to motivate me after the huge amount of setbacks, intrinsic to scientific work, and to shed a different, positive, light one the progress I made but wasn't able to recognize. He encouraged me to work in the field of free energy calculations, and he introduced snurportin 1 and the MHC II complex as interesting target proteins for these calculations. Moreover, he provided all members of the departement with an exceptional working environment. Thank you, Helmut!

Further thanks to Eveline Heinemann for the great work and support as a secretary. Thanks to Ansgar Esztermann and Martin Fechner for the IT support. Apart from thanks to all members of the department for the great help, friendship and ambience, in particular, I deeply thank Dr. Bert de Groot, Dr. Gerrit Groenhof, Dr. Matthias Müller, Dr. Lars Schäfer, Dr. Oliver Lange, Dr. Martin Stumpe, Dr. Guillem Portella, Dr. Marcus Kubitzki, and Dr. Jochen Hub for the friendship, support, scientific discussions, and adivce given. Exceptional thanks to Dr. Daniel Seeliger for the great time in our shared office, for the patience at discussions and questions, and for the many beers and cigarettes we enjoyed together. Thanks to Prof. Ralf Ficner and his department for the fruitful discussions about snurportin 1.

Finally, and most important to me, I thank my family, my parents Hubert and Maria which always believed in me and supported me wherever possible, and my girlfriend Mayumi Cuny for building me up when I was down. I thank all of my close friends who understood that lack of time during my thesis, and spatial distance lead to individual meeting rare events.

Bibliography

[1] Darwin, C. On the Origin of Species by Means of Natural Selection (1859).

[2] Venter, J. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

[3] Collins, F. *et al.* Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

[4] Hebert, P., Cywinska, A., Ball, S. & R deWaard, J. Biological identifications through DNA barcodes. *Proceedings- Royal Society of London. Biological sciences* **270**, 313–321 (2003).

[5] Stoeckle, M. Taxonomy, DNA, and the Bar Code of Life. *BioScience* **53**, 796–797 (2003).

[6] Hoppe, W., Lohmann, W., Markl, H. & Ziegler, H. Biophysik (1982).

[7] Jung, J. & Lee, W. Structure-based Functional Discovery of Proteins: Structural Proteomics. *Journal of Biochemistry and Molecular Biology* **37**, 28–34 (2004).

[8] Brünger, A. & Nilges, M. Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Quarterly Reviews of Biophysics* **26**, 49–125 (1993).

[9] Nilges, M. Structure calculation from NMR data. *Current Opinion in Structural Biology* **6**, 617–623 (1996).

[10] Berman, H. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).

[11] Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **15**, 285–289 (2005).

[12] Frauenfelder, H., Sligar, S. & Wolynes, P. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).

[13] Moffat, K. The frontiers of time-resolved macromolecular crystallography: movies and chirped X-ray pulses. *Faraday Discussions* **122**, 65–77 (2003).

[14] Schotte, F. *et al.* Watching a Protein as it Functions with 150-ps Time-Resolved X-ray Crystallography. *Science* **300**, 1944–1947 (2003).

[15] Eisenmesser, E., Bosco, D., Akke, M. & Kern, D. Enzyme Dynamics During Catalysis. *Science* **295**, 1520–1523 (2002).

[16] Kempf, J. & Loria, J. Protein dynamics from solution NMR: theory and applications. *Cell Biochemistry and Biophysics* **37**, 187–211 (2003).

[17] Brüschweiler, R. New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins. *Current Opinion in Structural Biology* **13**, 175–183 (2003).

[18] Mittermaier, A. & Kay, L. New Tools Provide New Insights in NMR Studies of Protein Dynamics. *Science* **312**, 224–228 (2006).

[19] Lange, O. *et al.* Recognition Dynamics Up to Microseconds Revealed from an RDC-Derived Ubiquitin Ensemble in Solution. *Science* **320**, 1471 (2008).

[20] Seifert, U. Entropy Production along a Stochastic Trajectory and an Integral Fluctuation Theorem. *Physical Review Letters* **95**, 40602 (2005).

[21] van Gunsteren, W. The role of computer simulation techniques in protein engineering. *Protein Engineering Design and Selection* **2**, 5 (1988).

[22] Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **93**, 2395–2417 (1993).

[23] Rodinger, T. & Pomès, R. Enhancing the accuracy, the efficiency and the scope of free energy simulations. *Current Opinion in Structural Biology* **15**, 164–170 (2005).

[24] Shirts, M., Mobley, D. & Chodera, J. Alchemical free energy calculations: ready for prime time? *Annual Reports in Computational Chemistry* **3**, 41–59 (2007).

[25] De Feyter, S., van Stam, J., Boens, N. & C. De Schryver, F. On the use of dynamic fluorescence measurements to determine equilibrium and kinetic constants. The inclusion of pyrene in β-cyclodextrin cavities. *Chemical Physics Letters* **249**, 46–52 (1996).

[26] Malmqvist, M. BIACORE: an affinity biosensor system for characterization of biomolecular interactions. *Biochem Soc Trans* **27**, 335–40 (1999).

[27] Englebienne, P., Van Hoonacker, A. & Verhas, M. Surface plasmon resonance: principles, methods and applications in biomedical sciences. *Spectroscopy* **17**, 255–273 (2003).

[28] Morris, G. *et al.* Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *Journal of Computational Chemistry* **19**, 1639–1662 (1998).

[29] Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A Fast Flexible Docking Method using an Incremental Construction Algorithm. *Journal of Molecular Biology* **261**, 470–489 (1996).

[30] Jones, G., Willett, P., Glen, R., Leach, A. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **267**, 727–748 (1997).

[31] de Graaf, C. *et al.* Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *Journal of Medicinal Chemistry* **49**, 2417–2430 (2006).

[32] Zentgraf, M. *et al.* Wie verlässlich sind aktuelle Docking-Ansätze für strukturbasiertes Wirkstoffdesign?-Fallstudie zur Aldose-Reduktase. *Angewandte Chemie* **119**, 3645 (2007).

[33] Åqvist, J., Medina, C. & Samuelsson, J. A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering Design and Selection* **7**, 385–391 (1994).

[34] Hansen, J., McDonald, I. & Visscher, P. Theory of Simple Liquids. *American Journal of Physics* **46**, 871 (1978).

[35] Chandler, D. *Introduction to modern statistical mechanics* (Oxford University Press New York, 1987).

[36] Simonson, T. *Free Energy Calculations*, chap. Free Energy Calculations: Approximate Methods for Biological Macromolecules, 423–462. Springer Series in Chemical Physics , Vol. 86 (Springer, 2007).

[37] Srinivasan, J., Cheatham, T., Cieplak, P., Kollman, P. & Case, D. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *Journal of the American Chemical Society* **120**, 9401–9409 (1998).

[38] Massova, I. & Kollman, P. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives in Drug Discovery and Design* **18**, 113–135 (2000).

[39] Kirkwood, J. Statistical Mechanics of Fluid Mixtures. *Journal of Chemical Physics* **3**, 300–313 (1935).

[40] Zwanzig, R. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *Journal of Chemical Physics* **22**, 1420–1426 (1954).

[41] Jorgensen, W. & Ravimohan, C. Monte Carlo simulation of differences in free energies of hydration. *Journal of Chemical Physics* **83**, 3050 (1985).

[42] van Gunsteren, W., Daura, X. & Mark, A. Computation of Free Energy. *Helvetica Chimica Acta* **85**, 3113–3129 (2002).

[43] Hu, H., Yun, R. & Hermans, J. Reversibility of Free Energy Simulations: Slow Growth May Have a Unique Advantage.(With a Note on Use of Ewald Summation). *Molecular Simulation* **28**, 67–80 (2002).

[44] Lu, N., Kofke, D. & Woolf, T. Improving the efficiency and reliability of free energy perturbation calculations using overlap sampling methods. *Journal of Computational Chemistry* **25**, 28–40 (2004).

[45] Woo, H. & Roux, B. Calculation of absolute protein-ligand binding free energy from computer simulations. *Proceedings of the National Academy of Sciences* **102**, 6825–6830 (2005).

[46] Oostenbrink, C. & van Gunsteren, W. Efficient calculation of many stacking and pairing free energies in DNA from a few molecular dynamics simulations. *Chemistry(Weinheim)* **11**, 4340–4348 (2005).

[47] Oostenbrink, C. & van Gunsteren, W. Calculating zeros: Non-equilibrium free energy calculations. *Chemical Physics* **323**, 102–108 (2006).

[48] Shirts, M. & Pande, V. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *Journal of Chemical Physics* **122**, 144107 (2005).

[49] Kofke, D. On the sampling requirements for exponential-work free-energy calculations. *Molecular Physics* **104**, 3701–3708 (2006).

[50] Mobley, D., Chodera, J. & Dill, K. On the use of orientational restraints and symmetry corrections in alchemical free energy calculations. *Journal of Chemical Physics* **125**, 084902 (2006).

[51] Born, M. & Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **84**, 457–484 (1927).

[52] Ryckaert, J., Ciccotti, G. & Berendsen, H. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *Journal of Computational Physics* **23**, 327–341 (1977).

[53] Hess, B., Bekker, H., Berendsen, H. & Fraaije, J. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry* **18**, 1463–1472 (1997).

[54] Hockney, R., Goel, S. & Eastwood, J. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics* **14**, 148–158 (1974).

[55] Lennard-Jones, J. On the Forces between Atoms and Ions. *Proceedings of the Royal Society of London. Series A* **109**, 584–597 (1925).

[56] Jorgensen, W., Maxwell, D. & Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **118**, 11225 (1996).

[57] Sorin, E. & Pande, V. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophysical Journal* **88**, 2472–2493 (2005).

[58] Wang, J., Cieplak, P. & Kollman, P. How well does a restrained electrostatic potential(RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* **21**, 1049–1074 (2000).

[59] Wang, J., Wolf, R., Caldwell, J., Kollman, P. & Case, D. Development and testing of a general amber force field. *Journal of Computational Chemistry* **25**, 1157–1174 (2004).

[60] Hünenberger, P. & McCammon, J. Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study. *Biophysical Chemistry* **78**, 69–88 (1999).

[61] Weber, W., Hünenberger, P. & McCammon, J. Molecular dynamics simulations of a polyalanine octapeptide under Ewald boundary conditions: Influence of artificial periodicity on peptide conformation. *Journal of Physical Chemistry B* **104**, 3668–3675 (2000).

[62] Bergdorf, M., Peter, C. & Hünenberger, P. Influence of cut-off truncation and artificial periodicity of electrostatic interactions in molecular simulations of solvated ions: A continuum electrostatics study. *The Journal of Chemical Physics* **119**, 9129 (2003).

[63] Brooks, B. *et al.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem* **4**, 187–217 (1983).

[64] Saito, M. Molecular dynamics simulations of proteins in solution: Artifacts caused by the cutoff approximation. *The Journal of Chemical Physics* **101**, 4055 (1994).

[65] Baumketner, A. & Shea, J. The influence of different treatments of electrostatic interactions on the thermodynamics of folding of peptides. *J. Phys. Chem. B* **109**, 21322–21328 (2005).

[66] Ewald, P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **369**, 253–287 (1921).

[67] Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an N log (N) method for Ewald sums in large systems. *Journal of Chemical Physics* **98**, 10089–10092 (1993).

[68] Berendsen, H., Postma, J., van Gunsteren, W., DiNola, A. & Haak, J. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **81**, 3684 (1984).

[69] van der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701–1718 (2005).

[70] Beutler, T., Mark, A., van Schaik, R., Gerber, P. & van Gunsteren, W. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chemical physics letters* **222**, 529–539 (1994).

[71] Vriend, G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **8**, 52–6 (1990).

[72] Chipot, C. & Pohorille, A. *Free Energy Calculations*. Springer Series in Chemical Physics , Vol. 86 (Springer, 2007).

[73] Warshel, A. & Sussman, F. Toward Computer-Aided Site-Directed Mutagenesis of Enzymes. *Proceedings of the National Academy of Sciences* **83**, 3806–3810 (1986).

[74] Hwang, J. & Warshel, A. Semiquantitative calculations of catalytic free energies in genetically modified enzymes. *Biochemistry* **26**, 2669–2673 (1987).

[75] Rao, S., Singh, U., Bash, P. & Kollman, P. Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature* **328**, 551–554 (1987).

[76] Lu, N. & Kofke, D. Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling. *Journal of Chemical Physics* **114**, 7303 (2001).

[77] Lu, N. & Kofke, D. Accuracy of free-energy perturbation calculations in molecular simulation. II. Heuristics. *Journal of Chemical Physics* **115**, 6866 (2001).

[78] Lu, N., Singh, J. & Kofke, D. Appropriate methods to combine forward and reverse free-energy perturbation averages. *Journal of Chemical Physics* **118**, 2977 (2003).

[79] Pearlman, D. & Kollman, P. The lag between the Hamiltonian and the system configuration in free energy perturbation calculations. *Journal of Chemical Physics* **91**, 7831 (1989).

[80] Straatsma, T. & McCammon, J. Multiconfiguration thermodynamic integration. *Journal of Chemical Physics* **95**, 1175 (1991).

[81] Bennett, C. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **22**, 245–268 (1976).

[82] Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters* **78**, 2690–2693 (1997).

[83] Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E* **56**, 5018–5035 (1997).

[84] Beveridge, D. & DiCapua, F. Free Energy Via Molecular Simulation: Applications to Chemical and Biomolecular Systems. *Annual Review of Biophysics and Biophysical Chemistry* **18**, 431–492 (1989).

[85] Straatsma, T. & McCammon, J. Computational Alchemy. *Annual Review of Physical Chemistry* **43**, 407–435 (1992).

[86] Hendrix, D. & Jarzynski, C. A" fast growth" method of computing free energy differences. *Journal of Chemical Physics* **114**, 5974–5981 (2001).

[87] Lua, R. & Grosberg, A. Practical applicability of the Jarzynski relation in statistical mechanics: A pedagogical example. *Journal of Physical Chemistry B* **109**, 6805–6811 (2005).

[88] Jarzynski, C. Rare events and the convergence of exponentially averaged work values. *Physical Review E* **73**, 46105 (2006).

[89] Hummer, G. Fast-growth thermodynamic integration: Error and efficiency analysis. *Journal of Chemical Physics* **114**, 7330–7337 (2001).

[90] Crooks, G. Nonequilibrium measurements of free energy differences for microscopically reversible markovian systems. *Journal of Statistical Physics* **90**, 1481–1487 (1998).

[91] Shirts, M., Bair, E., Hooker, G. & Pande, V. Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. *Physical Review Letters* **91**, 140601 (2003).

[92] Shirts, M. & Pande, V. Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *Journal of Chemical Physics* **122**, 134508 (2005).

[93] Anderson, J. Separate Sample Logistic Discrimination. *Biometrika* **59**, 19–35 (1972).

[94] Amadei, A., Linssen, A. & Berendsen, H. Essential dynamics of proteins. *Proteins: Structure, Function, Genetics* **17**, 412–425 (1993).

[95] Duda, R., Hart, P. & Stork, D. *Pattern Classification* (Wiley-Interscience, 2001).

[96] Kitao, A., Hirata, F. & Go, N. The effects of solqbent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chemical physics* **158**, 447–472 (1991).

[97] García, A. Large-amplitude nonlinear motions in proteins. *Physical Review Letters* **68**, 2696–2699 (1992).

[98] van Aalten, D. *et al.* The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water. *Proteins* **22**, 45–54 (1995).

[99] van Aalten, D., Findlay, J., Amadei, A. & Berendsen, H. Essential dynamics of the cellular retinol-binding protein evidence for ligand-induced conformational changes. *Protein Engineering Design and Selection* **8**, 1129–1135 (1995).

[100] de Groot, B., Hayward, S., van Aalten, D., Amadei, A. & Berendsen, H. Domain motions in bacteriophage T 4 lysozyme: A comparison between molecular dynamics and crystallographic data. *Proteins Structure Function and Genetics* **31**, 116–127 (1998).

[101] Crooks, G. & Jarzynski, C. Work distribution for the adiabatic compression of a dilute and interacting classical gas. *Physical Review E* **75**, 21116 (2007).

[102] Massey Jr, F. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* **46**, 68–78 (1951).

[103] Kosztin, I., Barz, B. & Janosi, L. Calculating potentials of mean force and diffusion coefficients from nonequilibrium processes without Jarzynski?s equality. *Journal of Chemical Physics* **124**, 064106 (2006).

[104] Strasser, A., Dickmanns, A., Lührmann, R. & Ficner, R. Structural basis for m_3G-cap-mediated nuclear import of spliceosomal UsnRNPs by snurportin1. *The EMBO Journal* **24**, 2235–2243 (2005).

[105] Aduri, R. *et al.* AMBER Force Field Parameters for the Naturally Occurring Modified Nucleosides in RNA. *Journal of Chemical Theory and Computation* **3**, 1464–1475 (2007).

[106] Meagher, K., Redman, L. & Carlson, H. Development of polyphosphate parameters for use with the AMBER force field. *Journal of Computational Chemistry* **24**, 1016–1025 (2003).

[107] Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R. & Klein, M. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* **79**, 926 (1983).

[108] Ben-Naim, A. & Marcus, Y. Solvation thermodynamics of nonionic solutes. *Journal of Chemical Physics* **81**, 2016 (1984).

[109] Huber, J. *et al.* Snurportin1, an m_3G-cap-specific nuclear import receptor with a novel domain structure. *The EMBO Journal* **17**, 4114–4126 (1998).

[110] Berg, J., Tymoczko, J., Stryer, L. & Stryer, L. *Biochemistry* (WH Freeman San Francisco CA:, 2002).

[111] Campbell, R. & Trowsdale, J. Map of the human MHC. *Immunol Today* **14**, 349–52 (1993).

[112] Townsend, A. & Bodmer, H. Antigen Recognition by Class I-Restricted T Lymphocytes. *Annual Reviews in Immunology* **7**, 601–624 (1989).

[113] Unanue, E. & Allen, P. The basis for the immunoregulatory role of macrophages and other accessory cells. *Science* **236**, 551–557 (1987).

[114] Allen, P., Babbitt, B. & Unanue, E. T-Cell Recognition of Lysozyme: The Biochemical Basis of Presentation. *Immunological Reviews* **98**, 171–187 (1987).

[115] Puri, J. & Factorovich, Y. Selective inhibition of antigen presentation to cloned T cells by protease inhibitors. *J Immunol* **141**, 3313–7 (1988).

[116] Jones, E., Fugger, L., Strominger, J. & Siebold, C. MHC class II proteins and disease: a structural perspective. *Nature Reviews Immunology* **6**, 271–282 (2006).

[117] Madigan, M., Martinko, J. & Parker, J. *Brock biology of microorganisms* (Prentice Hall, 2000).

[118] Stern, L. *et al.* Crystal structure of the human class II MHC protein HLA-DR 1 complexed with an influenza virus peptide. *Nature* **368**, 215–221 (1994).

[119] Sundberg, E. *et al.* Minor Structural Changes in a Mutated Human Melanoma Antigen Correspond to Dramatically Enhanced Stimulation of a CD4+ Tumor-infiltrating Lymphocyte Line. *Journal of Molecular Biology* **319**, 449–461 (2002).

[120] O'Sullivan, D. *et al.* On the interaction of promiscuous antigenic peptides with different DR alleles. Identification of common structural motifs. *Journal of Immunology* **147**, 2663–2669 (1991).

[121] Matera, A., Terns, R., Terns, M. *et al.* Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews Molecular Cell Biology* **8**, 209–220 (2007).

[122] Rollenhagen, C. & Panté, N. Nuclear import of spliceosomal snRNPs. *Canadian Journal of Physiology and Pharmacology* **84**, 367–376 (2006).

[123] Will, C. & Lührmann, R. Spliceosomal UsnRNP biogenesis, structure and function. *Current Opinion in Cell Biology* **13**, 290–301 (2001).

[124] Bahia, D., Bach-Elias, M., Aviñó, A., Eritja, R. & Darzynkiewicz, E. Trimethylguanosine Nucleoside Inhibits Cross-Linking Between Snurportin 1 and m_3G-capped U1 snRNA. *Nucleosides, Nucleotides & Nucleic Acids* **25**, 909–923 (2006).

[125] Stumpe, M. & Grubmüller, H. Aqueous Urea Solutions: Structure, Energetics, and Urea Aggregation. *Journal of Physical Chemistry. B* **111**, 6220–6228 (2007).

[126] Swope, W., Andersen, H., Berens, P. & Wilson, K. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *Journal of Chemical Physics* **76**, 637 (1982).

[127] Chodera, J., Swope, W., Pitera, J., Seok, C. & Dill, K. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *Journal of Chemical Theory and Computation* (2006).

[128] Paraskeva, E. *et al.* CRM1-mediated recycling of snurportin 1 to the cytoplasm. *Journal of Cell Biology* **145**, 255–64 (1999).

[129] Fischer, U. & Lührmann, R. An essential signaling role for the m_3G cap in the transport of U1 snRNP to the nucleus. *Science* **249**, 786 (1990).

[130] Fischer, U. Diversity in the signals required for nuclear accumulation of U snRNPs and variety in the pathways of nuclear transport. *Journal of Cell Biology* **113**, 705–714 (1991).