The Ligand Binding Mechanism of the MloK1 Ion Channel

Diploma Thesis

The Ligand Binding Mechanism of the MloK1 Ion Channel

Béla Voß*

August 30, 2010

Max Planck Institute for Biophysical Chemistry†
Department of Theoretical and Computational Biophysics
Supervisor: Helmut Grubmüller

-

Georg-August-University Göttingen
Faculty of Physics‡
Supervisor: Marcus Müller

*bvoss@gwdg.de
†Am Fassberg 11, 37077 Göttingen
‡Friedrich-Hund-Platz 1, 37077 Göttingen

Contents

Abbreviations

MD molecular dynamics

COM centre of mass

RMSD root mean square deviation

PCA principal component analysis

PMF potential of mean force

WHAM weighted histogram analysis method

cAMP cyclic adeonise monophosphate

CNBD cyclic nucleotide binding domain

1 Introduction

> "The system of life on this planet is so astoundingly complex that it was a long time before man even realised that it was a system at all and that it wasn't something that was just there."
>
> *(Douglas Adams)*

The binding of molecules to proteins is an essential step in biological systems. Depending on the exact process in question, small molecules, so-called ligands, can change the behaviour of the proteins, which often act as molecular machines in a living organism. The proteins may be activated or inactivated, their activity may be accelerated, slowed down or in some other way modulated.

The antibody-antigen reaction of the immune system is a prominent example of a binding process where a protein (the antibody) binds to the antigen, which may be a protein itself, a carbohydrate or some other molecule attached to the surface of a cell or virus.

An important family of proteins whose behaviour changes upon ligand binding are ion channels. Ion channels are proteins embedded in the cell membrane enabling and controlling the flow of ions through the cell membrane. Certain ion channels are only activated upon binding of a ligand, whereas others are deactivated or blocked upon ligand binding, phenomena that play an important role in the mechanism of action of many pharmaceuticals.

The ligand may influence the function of the protein in multiple ways, for example by replacing another bound molecule. In ion channels, the ligand may block the channel, thus disabling the ion flow through the channel. Since the structure of a protein is often related to its working mechanism, a ligand may cause its effect on the protein by causing a conformational change in the protein, which then again influences its function.

Here we will focus on such conformational changes caused by ligand binding. In figure 1.1 the starting (A) and end configuration (B) of such a ligand (blue) binding process are sketched together with the associated conformational change of the protein (black lines).

For ligand binding processes connected with structural changes in a protein from state A to state B the question arises how the ligand binds, how the conformational

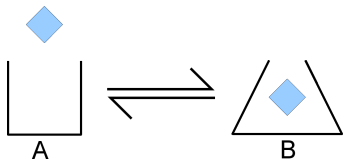


Figure 1.1: Schematic representation of a binding reaction: The ligand is symbolised by a blue square, the protein by black lines.

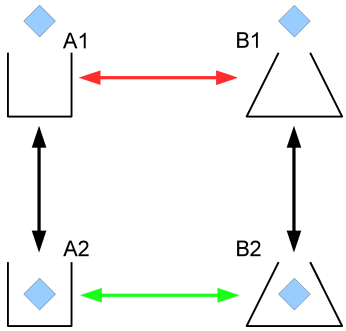


Figure 1.2: Sketch of ligand binding and conformational change

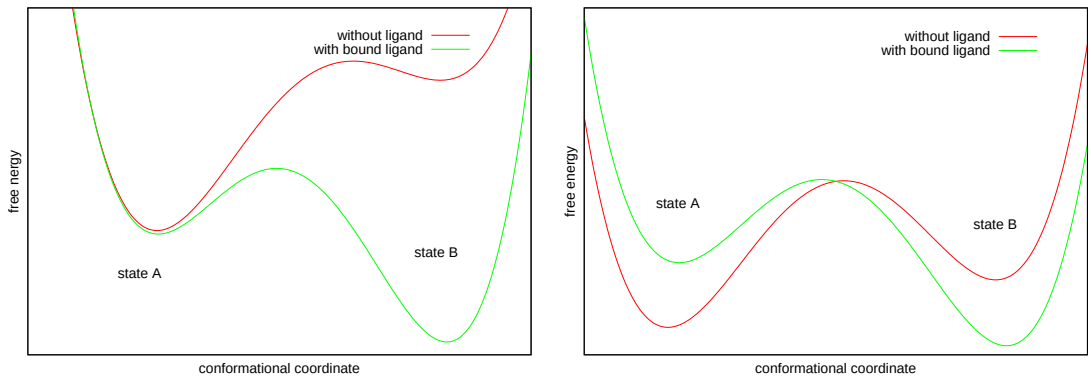
transition happens and how the structures are stabilised. Mainly two models have been suggested: *conformational selection* and *induced fit*.

The basic thermodynamic cycle is illustrated in fig. 1.2. Transitions from the one conformation (A) to the other (B) conformation may occur either during the absence of a ligand (1) or during the presence of a ligand (2). The two models differ in the free energies associated with the transitions from the open to the closed conformation (sketched in fig. 1.3).

According to the induced fit model (fig. 1.3(a)) the state A is stable in the absence of a bound ligand, that means it is energetically favoured (red line) compared to state B, thus occurrence of state B is unlikely, which means that transitions from A1 to B1 in fig. 1.2, pictured by a red arrow, are rare. Upon the binding of the ligand the free energy landscape changes (green line) in such a way that the second state becomes accessible and stable. The associated transition is pictured by a green arrow in fig. 1.2.

In the conformational selection model, for which free energy landscapes are sketched in fig. 1.3(b), both states are accessible in the absence of the ligand (red line). However, upon the binding of the ligand the energetic landscape changes in such a way that state B becomes preferred (green line).

In the case of a sufficiently large energetic barrier separating the two states that will be lowered upon the binding of the ligand, the mean transition time can become large compared to the time between binding and unbinding of the ligand. In this case transitions will only be observed after the binding of the ligand, thus



(a) Induced fit (b) Conformational selection

Figure 1.3: Simple free energy landscape models for induced fit and conformational selection

the kinetics may resemble that of an induced fit.

In summary, we define the two models by the different accessibility of state B: If this state can only occur after the ligand binds to the protein, we speak of induced fit. If both states can occur without a bound ligand and only the preference shifts after the binding, the conformational selection model applies.

In this work we focus on the binding mechanism of *cyclic adeonise monophosphate* at a specific domain of *MloK1*, a potassium channel found in the bacterium *Mesorizobium loti*. Cyclic adeonise monophosphate (cAMP) works as a *second messenger* in many biological systems, a molecule that transmits signals between different components and molecules within the cell. An extended description of the second messenger system and the role of cAMP in cell biology can be found for example in (Alberts, 2002).

The MloK1 ion channel has been shown to increase its ion conductivity upon the binding of cAMP (Clayton et al., 2004). This is assumed to be caused by a conformational change within the protein, which in turn is caused by the binding of the ligand. Whereas the largest part of in vivo MloK1 is embedded in a lipid bilayer, the cyclic nucleotide binding domain (CNBD), where a ligand can bind, sticks out of the bilayer. For this part two different structures have been observed: One structure with a bound ligand and one without a ligand. The fact that both structures as well as the position of the binding site are known, renders the system suitable for an investigation of the binding mechanism. A more extensive description of MloK1 and the CNBD will be given in section 3.2.

In this work we study the binding of cAMP to the CNBD by atomistic computer simulations. On this level protein conformation, ligand position and their temporal development can be studied. We will apply molecular dynamics simulations to study the kinetics and the thermodynamics of the binding process. This technique

and the ideas behind it will be presented in section 2.1. All applied methods are described in chapter 4.

In this work we want to answer the question if one of the described models, induced fit or conformational selection, can be applied to the binding of cAMP at the CNBD of MloK1 and if yes, which is the better suited model. To achieve this we want to find out if spontaneous ligand binding and conformational change occur in free simulations. We want to identify suiting reaction coordinates for the ligand biding and the conformational change and we want to study the thermodynamics of the systems for the reactions sketched in 1.2 along the obtained reaction coordinates to find out how the free energy landscape looks for these reactions.

2 Theoretical Background

"In theory, there is no difference between theory and practice. But, in practice, there is."

(Jan L. A. van de Snepscheut)

2.1 Molecular Dynamics

Although advances in real space microscopy such as the development of the scanning tunnelling microscope allow the resolution of single atoms, it is usually impossible to watch single atoms and molecules during chemical and biological processes. Since the basic physical laws governing atomistic systems are known (at least to a certain extend), it is in principle possible to solve the equations of motion. Because of the complexity this can only be done numerically and by using several approximations, but by doing this the temporal development of a molecular system can be simulated on an atomistic scale. The simulation method used in this work is called molecular dynamics and shall be described in the following section.

Molecular dynamics simulations describe a class of computer simulations, where a system is described by a set of small particles, usually atoms, and the time evolution of the system is determined by the forces interacting between the simulated particles. Usually the systems simulated by molecular dynamics consist of one or more interacting molecules, like proteins, other biomolecules or system from material sciences and solid state physics. Sometimes the term molecular dynamics is also used rather loosely as a synonym for *discrete element methods* in general.

2.1.1 Description of Molecular Systems

In a non-relativistic model where the substructure of an atomic nucleus is neglected, a molecule is given by a set of charged nuclei and surrounding electrons. These kind of systems are described by (non relativistic) quantum mechanics. The wave function ϕ of a many atoms system with N nuclei and K electrons in position space depends on the coordinates of the nuclei, in the following denoted by R_i, and of the electrons, r_i:

$$\phi = \phi(R_1, \ldots R_N, r_1, \ldots r_K). \tag{2.1}$$

If spin interactions are neglected (as these stem from a relativistic treatment) the Hamiltonian of the system consists of kinetic parts, a nucleus-nucleus interaction part, a electron-nucleus interaction part and an electron-electron interaction part. Since the interaction is given by Coulomb potentials, the Hamiltonian has the form

$$\hat{H}(\mathbf{R},\mathbf{r}) = \hat{T}_N + \underbrace{\hat{T}_e + \hat{V}_{NN} + \hat{V}_{eN} + \hat{V}_{ee}}_{\hat{H}_e} \tag{2.2}$$

$$\begin{aligned} &= -\sum_{k=1}^{N} \frac{\hbar^2}{2M_k} \underline{\Delta}_{R_k} - \frac{\hbar^2}{2m_e} \sum_{k=1}^{K} \underline{\Delta}_{r_k} + \frac{e^2}{4\pi\varepsilon_0} \sum_{k<j}^{N} \frac{Z_k Z_j}{||\mathbf{R}_k - \mathbf{R}_j||} \\ &\quad - \frac{e^2}{4\pi\varepsilon_0} \sum_{k=1}^{N} \sum_{j=1}^{K} \frac{Z_k}{||\mathbf{R}_k - \mathbf{r}_j||} + \frac{e^2}{4\pi\varepsilon_0} \sum_{k<j}^{K} \frac{1}{||\mathbf{r}_k - \mathbf{r}_j||} \end{aligned} \tag{2.3}$$

with e denoting electronic parts, N nuclear parts.

The time development of the system is given by the Schrödinger equation.

$$i\hbar \frac{\partial \phi(\mathbf{R},\mathbf{r},t)}{\partial t} = \hat{H}\phi(\mathbf{R},\mathbf{r},t). \tag{2.4}$$

Since the Hamiltonian is time-independent, the time-dependent solution can be written as a superposition of solutions of the stationary Schrödinger equation

$$\hat{H}\psi(\mathbf{R},\mathbf{r}) = E\psi(\mathbf{R},\mathbf{r}). \tag{2.5}$$

The total solution is then given by

$$\phi(\mathbf{R},\mathbf{r},t) = \sum_n c_n e^{-iE_n t/\hbar} \psi_n(\mathbf{R},\mathbf{r}). \tag{2.6}$$

Except for very small systems like atoms or very small molecules it is impossible to calculate the solutions of the Schrödinger equation numerically. Even with sophisticated perturbative methods, it is therefore infeasible to simulate macromolecules such as proteins over timescales of interest.

2.1.1.1 Born-Oppenheimer Approximation

The complexity of the system decreases significantly if the wave function is decomposed into a part describing the electrons and a part describing the nuclei. This is achieved by the Born-Oppenheimer approximation (Born and Oppenheimer, 1927).

The basic insight behind the Born-Oppenheimer approximation is the fact that the nuclei have a much larger mass than the electrons, their ratio is approximately $m_N/m_e = \mathcal{O}(10^3)$. Therefore it is reasonable to assume that the dynamics of the electrons takes place on a much faster timescale than that of the nuclei. This

suggests that upon a change in the configuration of the nuclei, the electrons will immediately adapt to the new ground state (provided, that the system does not occupy an excited state). This approximation lead to the following ansatz for the wave function:

$$\phi(\mathbf{R}, \mathbf{r}, t) = \chi(\mathbf{R}, t)\xi(\mathbf{r}; \mathbf{R}(t)). \tag{2.7}$$

χ represents the wave function of the nuclei, ξ the wave function of the electrons. Putting this ansatz into the stationary Schrödinger equation leads to

$$\hat{H}\phi = E\phi(R, r, t) \tag{2.8}$$

$$\hat{T}_N(\chi(\mathbf{R}, t)\xi(\mathbf{r}; \mathbf{R})) + \chi(\mathbf{R}, t)\hat{H}_e\xi(\mathbf{r}; \mathbf{R}) = E\phi(\mathbf{R}, \mathbf{r}, t). \tag{2.9}$$

With $\int \xi^*\xi dr = 1$ [1] and using

$$\hat{T}_N(\chi(\mathbf{R}, t)\xi(\mathbf{r}; \mathbf{R})) = \sum_\nu \frac{\hbar^2}{2M_\nu}\underline{\Delta}_{R_\nu}(\chi(\mathbf{R}, t)\xi(\mathbf{r}; \mathbf{R})) \tag{2.10}$$

$$= \frac{\hbar^2}{2M_\nu}\sum_\nu \chi\underline{\Delta}_{R_\nu}\xi + \xi\underline{\Delta}_{R_\nu}\chi + 2\nabla_{R_\nu}\chi\nabla_{R_\nu}\xi \tag{2.11}$$

we get after multiplying by ξ^* and integrating over r:

$$\hat{T}_N\chi + \chi\int \xi^*\hat{T}_N\xi dr + \chi\int \xi^*\hat{H}_e\xi dr = E\chi. \tag{2.12}$$

2.1.1.2 Approximation of the Electronic Wave Function

The Born-Oppenheimer approximation allows separation of the dynamics of the electrons and the nuclei. Usually two assumptions are made: The electronic wave function occupies the ground state with ground state energy E_0 and is given as an eigenfunction of the electronic Hamiltonian[2]:

$$\hat{H}_e(\mathbf{R}, \mathbf{r})\xi(\mathbf{r}; \mathbf{R}) = E_0(\mathbf{R})\xi(\mathbf{r}; \mathbf{R}). \tag{2.13}$$

The electronic wave function and thus the integrals in eq. (2.12) can now be calculated for an arbitrary number of fixed nuclei states. Together with an interpolation scheme that produces values for the integrals for non-precalculated nuclei states, eq. (2.12) is reduced to a differential equation only containing $\chi(\mathbf{R}, t)$.

[1] This means that we assume bound states.

[2] This approach might not be entirely correct. Another attempt is to describe the solution of the complete stationary Schrödinger equation as a series of solutions to the stationary Schrödinger equation: $\phi(\mathbf{R}, \mathbf{r}, t) = \sum_j \chi_j(\mathbf{R}, t)\xi_j(\mathbf{R}, \mathbf{r})$. Since ab initio molecular dynamics is not within the scope of this work we accept the inaccuracies in the deduction. See also (Gdanitz, 1999).

2.1.1.3 Classical Approximation

By constructing the limes $\frac{\hbar^2}{M} \to 0$ and replacing the wave function of the nuclei with the coordinates (i. e. treating them as classical particles) the following equations of motion are derived, assuming ground state energy for the electronic wave function (Griebel et al., 2007; Scherz, 1999):

$$\begin{aligned} M_\nu \ddot{R}_\nu &= -\nabla_{R_\nu} \int dr\, \xi_0^*(\mathbf{R}, \mathbf{r}) \hat{H}_e \xi_o(\mathbf{R}, \mathbf{r}) \\ &=: -\nabla_{R_\nu} V_e^{BO}(\mathbf{R}). \end{aligned} \tag{2.14}$$

Here we introduced the Born-Oppenheimer potential V_e^{BO}.

Solving eq. (2.14) numerically requires calculating V_e^{BO} and its derivatives with respect to R_ν. This can either be done by solving the Schrödinger equation for the electronic wave function after each integration step (see section 2.1.3.1), a procedure leading to *Born-Oppenheimer molecular dynamics*. The alternative is to approximate the Born-Oppenheimer potential as in the previous section.

2.1.1.4 Approximation of the Born-Oppenheimer Potential in the Classical Limit

In the classical approximation the Born-Oppenheimer potential only depends on the coordinates of the nuclei. To devise an approximation, it is expanded to a series:

$$V_e^{BO}(\mathbf{R}) \approx V_e^{model}(\mathbf{R}) = \sum_{\nu=1}^{N} V_1(R_\nu) + \sum_{\mu<\nu}^{N} V_2(R_\mu, R_\nu) + \sum_{\lambda<\mu<\nu}^{N} V_3(R_\lambda, R_\mu, R_\nu) \ldots \tag{2.15}$$

Note that the V_i can consist of multiple terms and can contain constants k_μ, $k_{\mu,\nu}$ etc. depending on the properties of the nuclei. The series is finite, therefore it is exact if all terms are taken into account.

With a complete approximation of the Born-Oppenheimer potential the (classical) dynamics of the nuclei can be calculated without having to solve the stationary Schrödinger equation for the electronic wave function anymore. This constitutes a completely classical treatment of a molecular system.

There are two possible ways to calculate an approximation of the Born-Oppenheimer potential. The first is to calculate it for many configurations $\{R\}$ by solving the stationary Schrödinger equation for the electronic wave function (eq. 2.13) and using inter- and extrapolation to obtain estimates for the remaining configurations. The other option is to define a set of analytical functions to approximate the V_i in eq. (2.15).

Knowing from chemistry that the strongest interactions in molecules are short ranged and act only between a few particles and that long ranged interactions can be modelled by pairwise interactions, it can be assumed that the series converges quickly and only the first terms have to be taken into account.

In molecular dynamics, a set of potentials V_i, together with all the constants contained is called a *force field*.[3] The choice of a "good" potential is not trivial. Due to the number of nuclei it will never be possible to solve eq. (2.14) analytically apart from the most trivial approximations such as $V_i = 0 \, \forall i$ which describes the ideal gas. Subsequently the differential equation (2.14) must be solved numerically which means that eq. (2.15) and all its derivatives must be calculated many times within a computer simulation. Whereas a large number of complicated V_i will raise severe computational performance issues, a too crude approximation of V_e^{BO} will lead to dynamics of the nuclei that significantly differs from the "real" solution of eq. (2.14).

2.1.2 Terms of a Common Force Field for Macromolecules

The desire of performing molecular dynamics simulations with macromolecules such as proteins will only be possible if generic force fields can be constructed which are not restricted in their usage to one specific system. The general way to reach this goal is to construct V_e^{model} from smaller building blocks by defining parameters for pair, triplet and quadruplet interactions.

In a first step one distinguishes between intramolecular short-ranged interactions between a fixed set of atoms, which will be called bonded interactions and more long-ranged electrostatic or dipole-dipole interactions which also exist between different molecules:

$$V_e^{model} = V_{ff} = V_{bond} + V_{nonbond}. \tag{2.16}$$

The exact interactions are implementation-dependent, the following sections describe the potentials that were applied for the simulations of this work.

2.1.2.1 Bonded Interactions

The bonded interactions are sketched in fig. 2.1.

Pair Bonds Interaction of atoms which are usually described by covalent bonds are modelled by a harmonic potential:

$$V(r_{ij}) = \frac{k_{ij}}{2}(r_{ij} - r_{ij,0})^2. \tag{2.17}$$

[3] Although the expression is physically completely misleading since it is a potential (a physical force field is a vectorfield containing the forces exerted by one object on another), it has been established by chemists as the standard expression in the literature.

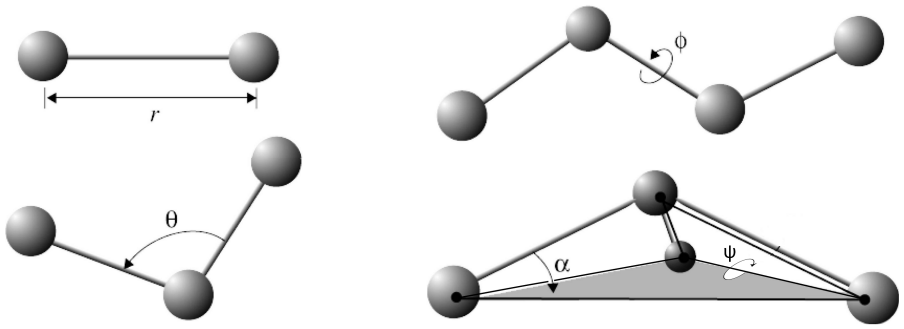


Figure 2.1: Bonded interactions

Here r_{ij} denotes the distance of the two atoms and $r_{ij,0}$ a reference length, which corresponds to the chemical bond length of the covalent bond. k_{ij} is the corresponding spring constant and describes the strength of the bond. Bonds described by a harmonic potential cannot be broken during a simulation, therefore the usage of this potential is limited to systems where no chemical reactions occur.

The spring constants of pair interactions are large, therefore the only motion along these bonds are high frequency oscillations with low amplitudes at room temperature. Since these oscillations are assumed to be of minor biological relevance for the studied systems and their contribution to free energies does not change during computer simulations, the bond lengths are usually constrained to a fixed value.

Angle Potentials In addition to the pair interaction a triplet interaction, viz. the harmonic angle potential, is defined:

$$V(\theta_{ijk}) = \frac{k_{ijk}}{2}(\theta_{ijk} - \theta_{ijk,0})^2. \tag{2.18}$$

θ_{ijk} denotes the angle spanned by the covalent bonds between the atoms i and j and j and k.

Dihedral Potentials A four-body interaction between four covalently bound atoms is defined that describes the potential energy of the dihedral angle ϕ between the plane defined by the first three atoms and the plane constructed by the last three atoms (assuming no consecutive triplet of atoms forms a straight line). There are multiple approaches to describe the general form of the potential. Either a simple periodic function is used or a series expansions of powers of cosines, so called Ryckaert-Bellemans functions. The periodic potentials are given by

$$V_d(\phi_{ijkl}) = k_\phi(1 + \cos(n\phi - \phi_{cis})). \tag{2.19}$$

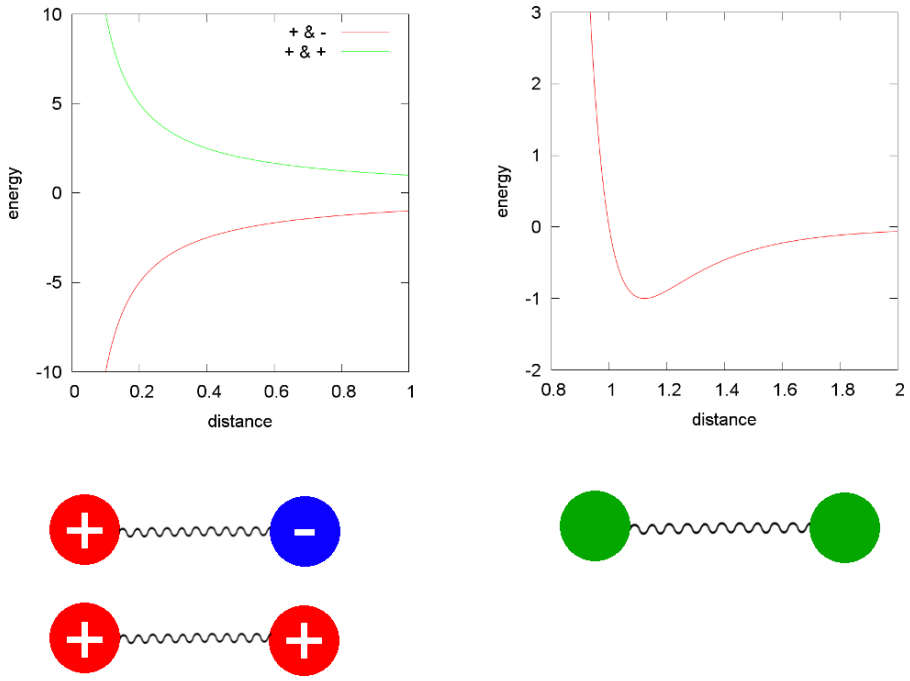


Figure 2.2: Nonbonded interactions. Left: Coulomb potential. Right: Lennard-Jones potential

The Ryckaert Bellman potentials have the form

$$V_{rb}((\phi_{ijkl}) = \sum_{n=0}^{m} c_n \cos^n(\phi), \tag{2.20}$$

where the usual choice (at least for alkanes) is $m = 5$. Note that the Ryckaert-Belleman potential is equivalent to a description with a Fourier series (only the constants are different).

Improper Dihedrals The potentials introduced so far offer no way to stabilise planar rings like benzol. Therefore a harmonic potential for the angle between the planes defined by three out of four consecutively covalently bound atoms is defined, the so called improper dihedral potential. It is given by

$$V_{id}(\psi_{ijkl}) = \frac{k_\psi}{2}(\psi_{ijkl} - \psi_0)^2. \tag{2.21}$$

2.1.2.2 Non-Bonded Interactions

The non-bonded interactions are sketched in fig. 2.2

Coulomb Interaction The Coulomb potential between two point charges is given by

$$V_C(r_{ij}) = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{q_i q_j}{r_{ij}}, \tag{2.22}$$

with the effective charges q_i. The effective charges are usually determined in such a way that the real electrostatic potential generated by the nuclei and the electronic wave function is approximated by point charges residing at the atomic positions. For true atomistic simulations the dielectric constant is $\varepsilon_r = 1$.

Ewalds Summation The Coulomb interaction is a long-range interaction, meaning that it decays only slowly over distance. Since it cannot be neglected even at large distances, it must be calculated for all pairs of particles, which means that the calculation will scale with $\mathcal{O}(N^2)$. To increase the efficiency, a number of techniques has been developed. For systems with periodic boundary conditions Ewald summation (Ewald, 1921) which scales with $\mathcal{O}(N^{3/2})$, can be used in which the electrostatic potential is split into a real space and a Fourier space part. The performance of the calculation of the Fourier part of the Ewalds summation is improved furthermore by introducing a mesh on which the charges are positioned. One algorithm that uses this strategy is the Particle Mesh Ewald method (Darden et al., 1993). Such methods scales with $\mathcal{O}(N\ln(N))$.

Lennard-Jones Interaction The Lennard-Jones Potential has the following form:

$$V_{LJ}(r_{ij}) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right). \tag{2.23}$$

It consists of an attractive part, the Van-der-Waals potential $V_{vdW} \approx 1/r^6$ which is a classical description of induced dipole-dipole interactions. The repulsive $1/r^{12}$ term describes the repulsion of the electrons surrounding the atom nuclei due to the Pauli principle. The $1/r^{12}$ dependency is chosen both empirically and out of numerical convenience. In a force field, values for ϵ and σ are specified for each atom type. To calculate the corresponding parameters for interactions between different atom types, geometric averages are used (for the σ, it is possible to use arithmetic averages, too).

Since the Lennard-Jones Potential is decaying faster than $1/r^2$, long range tails can be neglected by using either plain cutoffs or shifting or switching the potential function.

Buckingham Potential Instead of the $1/r^{12}$ repulsion term an exponential term can be used, which is more realistic but computationally more expensive. For the Buckingham potential the expression for the potential becomes

$$V_{BH}(r_{ij}) = a_{ij} \exp(-b_{ij} r_{ij}) - \frac{c_{ij}}{r_{ij}^6}. \tag{2.24}$$

2.1.2.3 Additional Notes on Potentials

In theory, an arbitrary number of potentials can be defined. Hydrogen bonds for example are of quantum mechanic nature (and not only classically electrostatic), therefore some force fields include terms to account for that.

There are two main approaches to find good parameters for all the constants in the force fields. One method is to derive them from quantum mechanical ab initio calculations. The other approach is to fit them in such a way that experimentally accessible thermodynamic quantities can be reproduced in molecular dynamics simulations using these force fields.

Force fields using point charges cannot take into account any polarisation effects apart from global orientations of dipole molecules. (The Lennard-Jones potential contains the effects of spontaneously induced dipoles though.) Therefore attempts have been made to construct polarisable force fields, which can take into account such effects (Halgren and Damm, 2001).

2.1.3 Performing a Molecular Dynamics Simulation

The basic steps of a molecular dynamics simulation are quite simple. After initialising the system by selecting the initial positions and velocities of all atoms, the forces acting on the particles are calculated by building the derivative of the potential with respect to the corresponding coordinates. Using Newtons law $F = m\ddot{x}$ and a suiting integrator algorithm, the positions (and velocities) of all atoms at the time $t = t_0 + \Delta t$ are calculated, Δt being the time step of the integrator. These two steps, force calculation and integration are repeated.

2.1.3.1 Choice of Integrator Algorithm

A common integrator is the so called *Verlet algorithm*, introduced by Carl Størmer and made popular by Verlet (1967). The positions of all atoms are calculated via

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \frac{F(t)}{m} \Delta t^2. \tag{2.25}$$

The error per step is of the order $\mathcal{O}(\Delta t^4)$. If quantities which depend on the particle velocities are of interest, the velocities can either be calculated via $v(t) =$

$\frac{r(t+\Delta t)-r(t-\Delta t)}{2\Delta t}$ up to a precision of $\mathcal{O}(\Delta t^2)$ or an equivalent formulation of the Verlet algorithm, the leapfrog algorithm, can be used. In this algorithm positions and velocities are calculated for alternating points in time:

$$v(t+\frac{\Delta t}{2}) = v(t-\frac{\Delta t}{2}) + \frac{F(t)}{m}\Delta t \tag{2.26}$$

$$r(r+\Delta t) = r(t) + v(t+\frac{\Delta t}{2})\Delta t + \frac{F(t)}{m}\Delta t^2. \tag{2.27}$$

Looking at

$$v(t-\frac{\Delta t}{2}) = \frac{r(t)-r(t-\Delta t)}{\Delta t} + \mathcal{O}(\Delta t^2) \tag{2.28}$$

reveals the equivalence.

Although the error of the positions per step is of the order $\mathcal{O}(\Delta t^4)$, the total error over long times scales with $\mathcal{O}(\Delta t^2)$. We want to point out, however, that the numerical integration of many particle molecular dynamics systems is subject to Lyapunov instability. This means that the trajectory is sensitive to the initial conditions of the system. Trajectories of systems starting with tiny differences in the starting conformation will diverge very quickly.

However, the chaotic behaviour is no significant problem: The goal of molecular dynamics is not to predict exact trajectories for all particles but to give predictions about the statistical behaviour of the entire system. Since it is believed that the simulated trajectories are representative for true trajectories[4], the statistics derived from them should be correct (Frenkel and Smit, 2001; Gillilan and Wilson, 1992).

2.1.3.2 Ensembles and Thermostats

Assuming the integrator conserves energy over long time scales, the ensemble sampled with the above method will be the microcanonical ensemble. This ensemble can only be realised approximately in experiments and does not occur in nature (apart from systems of cosmic scale and maybe the universe as a whole). Most of the systems of interest correspond either to the canonical ensemble (with constant temperature and volume) or to the isobaric-isothermal ensemble.

To mimic the impact of a heat bath which ensures a constant temperature, so called thermostats are used. A number of coupling algorithms exits for temperature coupling, e. g. the Berendsen thermostat (Berendsen et al., 1984) , which ensures strongly dampened relaxations of temperature differences but does not generate a proper canonical ensemble, the velocity rescaling thermostat (Bussi et al., 2007) which corrects for the deficiencies of the Berendsen thermostat by introducing a

[4]Note that we are talking about true trajectories in the sense of trajectories calculated for the potential given by the force field by a perfect integrator, not about trajectories occurring in nature.

stochastic term and the Nosé-Hoover algorithm (Nosé, 2002; Hoover, 1985), which leads to oscillatory relaxation. Another way to ensure a canonical ensemble is the use of Brownian dynamics. However, the trajectories from such simulations do no longer have a physical meaning for large friction constants.

For pressure coupling similar thermostat algorithms have been derived, which basically rescale the coordinates of the system and the boundaries of the simulation box, e. g. Berendsen pressure coupling (Berendsen et al., 1984) (which also does not produce a well defined ensemble) or Parrinello-Rahman pressure coupling (Parrinello and Rahman, 1981).

2.2 Umbrella Sampling

Umbrella sampling is a method introduced by Torrie and Valleau (1977) to perform free energy calculations by significantly increasing sampling.

Let us consider a classical system with an internal energy function $U(q)$ and a reference system with a different internal energy $U_0(q)$. The velocity dependent part of the Hamiltonian shall have the form $p_i^2/2m$ and thus be separable.

For the free energy difference between the systems we thus obtain:

$$F - F_0 = -kT \ln \left(\int \mathrm{d}q \, e^{-\beta U(q)} \right) + kT \ln \left(\int \mathrm{d}q \, e^{-\beta U_0(q)} \right) \tag{2.29}$$

$$\beta(F - F_0) = -\ln \frac{\int \mathrm{d}q \, e^{-\beta U(q)}}{\int \mathrm{d}q \, e^{-\beta U_0(q)}} = -\ln \frac{\int \mathrm{d}q \, e^{\beta(-U(q)+U_0(q))} e^{-\beta U_0(q)}}{\int \mathrm{d}q \, e^{-\beta U_0(q)}} \tag{2.30}$$

$$=: -\ln \left\langle e^{\beta(-U+U_0)} \right\rangle_0 \tag{2.31}$$

$$=: -\ln \left\langle e^{-\beta \Delta U} \right\rangle_0 \tag{2.32}$$

$$= -\ln \int \mathrm{d}(\Delta U) \, \beta \, \rho_0(\beta \Delta U) e^{-\beta \Delta U}. \tag{2.33}$$

Here $\langle\rangle_0$ denotes the average over the reference ensemble and $\rho_0(\beta\Delta U)$ denotes the probability density of $\beta\Delta U$ in the reference system. We want to note that in the above problem we might as well consider systems with different temperatures (this is interesting when studying free energy differences in phase transitions).

In order to get a good estimate of the free energy difference from a computer simulation, it is necessary to obtain a good estimate of $\rho(\beta\Delta U)$ especially for those values of ΔU where $\rho_0(\beta\Delta U)e^{-\beta\Delta U}$ is large and contributes to the average. In a normal Monte Carlo or molecular dynamics simulation however this region is sampled in a simulation on the first system (with energy U) and for big values of ΔU the sampling errors become significant.

As a consequence, the idea of Umbrella sampling is to introduce an artificial biasing potential that allows sampling of both the reference system and the system

of interest. If we introduce a weighting function $w(q) = W(\beta\Delta U(q))$ we obtain a new distribution function

$$p(q) = \frac{w(q) \cdot e^{-\beta U_0(q)}}{\int \mathrm{d}q\, w(q) e^{-\beta U_0(q)}}. \tag{2.34}$$

The unbiased average for any quantity $A(q)$ out of the biased simulation is obtained by

$$\langle A \rangle_0 = \frac{\int \mathrm{d}q\, \frac{A(q)}{w(q)} w(q) e^{-\beta U_0(q)}}{\int \mathrm{d}q\, \frac{1}{w(q)} w(q) e^{-\beta U_0(q)}} \tag{2.35}$$

$$= \frac{\langle A/w \rangle_w}{\langle 1/w \rangle_w}. \tag{2.36}$$

Here $\langle\ \rangle_w$ denotes averages over the distribution given by equation (2.34).

For the probability distributions $\rho_0(\Delta U)$ and $\rho_w(\Delta U)$ we have

$$\rho_0(\beta\Delta U) = \frac{e^{-\beta\Delta U} \omega(\Delta U)}{\int \mathrm{d}q\, e^{-\beta U_0}} \tag{2.37}$$

$$\rho_w(\beta\Delta U) = \frac{e^{-\beta\Delta U}\, W(\beta\Delta U)\, \omega(\Delta U)}{\int \mathrm{d}q\, w(q) e^{-\beta U_0}} \tag{2.38}$$

$$\Rightarrow \rho_0(\beta\Delta U) = \frac{\rho_w(\beta\Delta U)}{W(\beta\Delta U)} \cdot \frac{\int \mathrm{d}q\, w(q) e^{-\beta U_0(q)}}{\int \mathrm{d}q\, e^{-\beta U(q)}} \tag{2.39}$$

$$= \frac{\rho_w(\beta\Delta U)}{W(\beta\Delta U)} \cdot \frac{\int \mathrm{d}q\, w(q) e^{-\beta U_0(q)}}{\int \mathrm{d}q\, \frac{1}{w(q)} w(q) e^{-\beta U(q)}} \tag{2.40}$$

$$= \frac{\rho_w(\beta\Delta U) \cdot W(\beta\Delta U)}{\langle 1/w \rangle_w}. \tag{2.41}$$

The microcanonical partition sum $\omega(E)$ describes the degeneracy of the energy states.

If $W(\beta\Delta U)$ is chosen wisely, then $\rho_w(\beta\Delta U)$ is approximately uniform in the energy interval of interest and a good estimate for $\rho_0(\beta\Delta U)$ is obtained over the same range, too.

To gain a good sampling it is not necessary to find one $w(q)$ to sample the region of interest in a single simulation run. Instead it is usually more efficient to use multiple umbrella windows and to sample successively the relevant part of the conformational space. Using this approach, it is also easier to choose the weighting functions, because the $\rho_{w,i}(\beta\Delta U)$ only need to be approximately uniform in a much smaller region. In the following we want to reproduce a train of thoughts sketched in Frenkel and Smit (2001) that explains why the use of a bigger number of umbrella windows is more sophisticated than the use of a single weighting function.

Let us assume that in our above system we use multiple weighting functions, which successively sample parts of the conformational space corresponding to energy regions $\Delta U_i = i \cdot \Delta U/n$, with n corresponding to the number of umbrella potentials. Furthermore we assume that the associated Markov chain is that of a random walk with diffusion constant D in the energy interval with the width $\Delta U/n$. Using this model leads to an estimate of the characteristic time needed to sample one interval:

$$t_n = \frac{\Delta U^2}{n^2 \cdot D}. \tag{2.42}$$

The total simulation time would thus be

$$t_{total} = n \cdot t_n = \frac{\Delta U^2}{n \cdot D}. \tag{2.43}$$

However, for each umbrella window there will be a certain equilibration time t_{equi} that needs to be sampled regardless of the size of the window. We assume the equilibration time to be constant for all simulation windows. As a consequence the total simulation time becomes

$$t_{total} = nt_n + nt_{equi} \tag{2.44}$$

$$= \frac{\Delta U^2}{n \cdot D} + nt_{equi} \tag{2.45}$$

$$\Rightarrow \frac{\mathrm{d}t_{total}}{\mathrm{d}n} = -\frac{\Delta U^2}{n^2 D} + t_{equi} \stackrel{!}{=} 0 \tag{2.46}$$

$$\Rightarrow t_n = t_{equi}. \tag{2.47}$$

Although this result is based on some assumptions that might not be applicable to every system, the result that an optimal number of successive umbrella sampling simulations larger than one exists, remains valid.

A common choice for the biasing functions are (multidimensional) harmonic potentials along one or more coordinates in the configurational space around a reference point in the region that is to be sampled. The biasing function $w(q)$ then becomes:

$$w(q) = \exp\left(-\beta U_{umbrella}(q)\right) = \exp\left(-\beta \frac{k}{2}(q - q_{ref})^2\right) \tag{2.48}$$

with a spring constant k.

We want to emphasise that the approach is not limited to free energy differences of systems with different Hamiltonians, but for all problems where larger parts of the configurational space needs to be sampled. Another quantity (although it is strongly related to free energy differences) that can be calculated with the help of umbrella sampling is the *potential of mean force*.

2.2.1 Potential of Mean Force

The potential of mean force (short *PMF*), introduced by Kirkwood (1935) has become an important concept of statistical physics. In the following section the concept and its calculation shall be explained.

Let us consider a spatial finite system of multiple interacting particles. The configurational part of the partition sum in a canonical ensemble is then given by:

$$Z = \int \mathrm{d}q_1 \dots \mathrm{d}q_n \, e^{-\beta H(\mathbf{q})}. \tag{2.49}$$

Here H is the Hamiltonian of our system, depending on the coordinates $\mathbf{q}$. Again the impulse-dependent part shall only contribute with a constant to all quantities of interest. Furthermore we have $\beta = \frac{1}{k_B T}$.
Since our system is finite, we can define the notation

$$\int \mathrm{d}q_i = a_i. \tag{2.50}$$

The average of some observable $A(\mathbf{q})$ is then given by

$$\overline{A} = \frac{\int \mathrm{d}q_1 \dots \mathrm{d}q_n \, A \, e^{-\beta H(\mathbf{q})}}{Z} = \frac{\int \mathrm{d}q_1 \dots \mathrm{d}q_n \, A \, e^{-\beta H(\mathbf{q})}}{\int \mathrm{d}q_1 \dots \mathrm{d}q_n e^{-\beta H(\mathbf{q})}} \tag{2.51}$$

which can be written as

$$\overline{A} = \frac{1}{a_i} \int \mathrm{d}q_i \, A \, \frac{a_i \int \mathrm{d}q_1 \dots \mathrm{d}_{q-i} \mathrm{d}q_{i+1} \dots \mathrm{d}q_n \, e^{-\beta H(\mathbf{q})}}{\int \mathrm{d}q_1 \dots \mathrm{d}q_n \, e^{-\beta H(\mathbf{q})}} \tag{2.52}$$

$$= \frac{1}{a_i} \int \mathrm{d}q_i \, A \, e^{-\beta W(q_i)}. \tag{2.53}$$

$W(q_i)$ is the potential of the mean force acting on the particle(s) connected to the coordinate q_i along this coordinate in a fixed coordinate system. The approach is by no means restricted to a single one-dimensional coordinate. In fact the $\mathrm{d}q_i$ can be a quite arbitrary differential element in the configurational space.

To show that W actually describes the potential of the mean force, one first writes the (generalised) force on a particle (or group of particles) as the derivative of the potential energy. In the second step one calculates the average over this force as in eq. (2.51). The result is that the average force is indeed the derivative of W. A full deduction of this can be found in (Kirkwood, 1935).

The potential of mean force describes the probability distribution of the system along the coordinates upon which it depends. Conversely, this means that the potential of mean force along some coordinate can be calculated from the average distribution function. For the one-dimensional case we get:

$$W(q_i) = W(q_i^*) - k_B T \ln \left(\frac{\langle \rho(q_i) \rangle}{\langle \rho(q_i^*) \rangle} \right). \tag{2.54}$$

$W(q_i^*)$ and q_i^* are constants that can be chosen arbitrarily: q_i^* is an arbitrary reference point along the coordinate q_i and $W(q_i^*)$ the potential of mean force at this point. This expresses that the potential of mean force (as any potential which is defined over its resulting forces) is only fixed up to a constant. The extension to multiple dimensions is a straightforward replacement of the q_i to a set of multiple coordinates, which are of course not restricted to a specific basis $(q_1, \ldots, q_n)$ of the configurational space but can be any function of the q.

If we know the average distribution function, we can also calculate the potential of mean force. However, it is usually impossible to obtain the average distribution from computer simulations due to sampling problems, which may be caused by energy barriers along the selected coordinate(s) or simply the size of the relevant parts of the configurational space. Therefore a number of approaches exists to calculate the potentials of mean force or the distribution function, e. g. thermodynamic integration, free energy perturbation, the Jarzynski method (Jarzynski, 1997) and umbrella sampling.

In the following the calculation of the potential of mean force using umbrella sampling shall be explained. A more extensive depiction, also including a comparison of multiple methods can be found in (Roux, 1995).

As stated in section (2.2) it is usually more efficient to use multiple simulations with different biasing functions. To calculate the potential of mean force along a (one-dimensional) coordinate χ a harmonic biasing potential of the form

$$V_j(\chi) = \frac{k}{2}(\chi - \chi_j)^2 \tag{2.55}$$

is applied during the simulation.

The goal is to obtain the unbiased distribution function from the biased simulation data to calculate the potential of mean force. The biased average distribution function for a specific umbrella window is given by

$$\langle \rho(\chi) \rangle_j^{biased} = \frac{e^{-\beta V_j(\chi)} \langle \rho(\chi) \rangle}{\langle e^{-\beta V_j(\chi)} \rangle} \tag{2.56}$$

$$\langle \rho(\chi) \rangle = \langle \rho(\chi) \rangle_j^{biased} e^{\beta V_j(\chi)} \langle e^{-\beta V_j(\chi)} \rangle \tag{2.57}$$

Together with eq. (2.54) we get for the potential of mean force from the single umbrella window simulations:

$$W_j(\chi) = W(\chi^*) - k_B T \ln \left(\frac{\langle \rho(\chi)_i^{biased} \rangle}{\langle \rho(\chi^*) \rangle} \right) - V_j(\chi) - f_j \tag{2.58}$$

$$\text{with } f_j = -kT \ln \left(\langle e^{-\beta V_j(\chi)} \rangle \right). \tag{2.59}$$

The f_j cannot be directly obtained from the simulation data. As long as only one biasing potential is used, the single f_j can be absorbed in $W(\chi^*)$ and thus be

ignored. However, in the normal case of multiple biasing potentials the problem has to be tackled. The traditional approach is to fit the f_j in such a way that the PMF of neighbouring umbrella windows match in the overlapping region. A number of proposed methods are compared in (Roux, 1995). Here we want to restrict ourself on the *weighted histogram analysis method* (*WHAM*), introduced by Kumar et al. (1992).

2.2.1.1 Weighted Histogram Analysis Method

The idea of the weighted histogram analysis method is to make an estimate for the whole unbiased average distribution function by depicting it as the weighted sum of the unbiased average distribution functions of N individual umbrella windows. The weighting factor is thereby defined by the number of data points n_j in the umbrella window simulation, the biasing umbrella potential $w_j(\chi)$ and the the free energy constant f_j associated with the umbrella window:

$$\langle \rho(\chi) \rangle = \sum_{j=1}^{N} \langle \rho(\chi) \rangle_i^{unbiased} \cdot \frac{n_j \, e^{-\beta(V_j(\chi)-f_j)}}{\sum_{j=1}^{N} n_j \, e^{-\beta(V_j(\chi)-f_j)}}. \tag{2.60}$$

Together with eq. (2.57) and eq. (2.59) we can write this equation as

$$\langle \rho(\chi) \rangle = \frac{\sum_{j=1}^{N} n_j \langle \rho(\chi) \rangle_j^{biased}}{\sum_{j=1}^{N} n_j \, e^{-\beta(V_j(\chi)-f_j)}} \tag{2.61}$$

and from eq. (2.59) we get

$$e^{-\beta f_j} = \int \mathrm{d}\chi e^{-\beta V_j(\chi)} \langle \rho(\chi) \rangle. \tag{2.62}$$

Equation (2.61) and (2.62) are known as the WHAM equations. By solving them iteratively, self-consistent estimates for the f_j and $\langle \rho(\chi) \rangle$ can be obtained.

χ is not restricted to a one-dimensional coordinate but can be a multidimensional vector. In practice, however, the rapidly increasing numbers of umbrella windows limits the number of dimensions in which the umbrella sampling can be performed.

2.2.1.2 Relation to Free Energy Differences

From eq. (2.53) and (2.54) we find immediately that the partial partition sum of our system in a confined region of the configurational space is given by

$$Z_R \propto \int_R \mathrm{d}\mathbf{z} \, e^{-\beta W(\mathbf{z})}. \tag{2.63}$$

If our system occupies separated regions of the conformational space (e. g. due to large energy barriers), we can use the PMF to calculate the free energy difference between the two states.

$$\Delta F = -\frac{1}{\beta} \ln \left(\frac{Z_{R1}}{Z_{R2}} \right) \tag{2.64}$$

$$= -\frac{1}{\beta} \ln \left(\frac{\int_{R1} \mathrm{d}\mathbf{z}\, e^{-\beta W(\mathbf{z})}}{\int_{R2} \mathrm{d}\mathbf{z}\, e^{-\beta W(\mathbf{z})}} \right) \tag{2.65}$$

2.3 Principal Component Analysis

2.3.1 Motivation

When being confronted with problems in a high-dimensional space spanned by multiple variables one naturally searches for a way to reduce the dimensionality of the problem since the analysis and visualisation of such data is difficult.

Principal component analysis (short PCA) is a parameter-free method to reduce the dimensionality of such problems without any a priori hypothesis of the probability distributions of the variables simply by calculating a (usually significantly smaller) set of uncorrelated variables that consist of linear combinations of the originally correlated variables.

2.3.2 Mathematical Background

Let us consider a set of M random variables X_i, obeying some (not necessarily identical) probability distribution. For two of the variables the covariance is defined as

$$\mathrm{Cov}(X_i, X_j) = \langle (X_i - \langle X_i \rangle) \cdot (X_j - \langle X_j \rangle) \rangle \tag{2.66}$$

where the brackets denote the averages of the underlying distributions. From these covariances the covariance matrix $\mathbf{C}$ can be constructed, with entries

$$c_{ij} = \mathrm{Cov}(X_i, X_j). \tag{2.67}$$

Since the matrix is symmetric, it can be diagonalised:

$$\mathbf{D} = \mathbf{Q^T C Q}, \tag{2.68}$$

with an orthogonal matrix $\mathbf{Q}$. The entries $\lambda_1, ..\lambda_M$ of $\mathbf{D}$, which are the eigenvalues of $\mathbf{C}$, and their corresponding eigenvectors $\mathbf{v}_k$ are the covariances of the transformed random variables $Y = \mathbf{Q^T} X$.

Subsequently new "coordinates", i. e. new random variables are obtained which are pairwise and linearly uncorrelated. If we order the eigenvectors by the size of the corresponding eigenvalue, the first eigenvector corresponds to the "coordinate" in the M-dimensional space of the random variables along which the covariance of the random variables is maximised. The second eigenvector maximises the variance in the orthogonal complement of the eigenspace of the first eigenvector and so on.

In many cases (see 2.3.3) the eigenvalue spectrum is rapidly decaying and only a low number of significant uncorrelated variables remains, hence the name *principal component* analysis.

In applications, one usually deals with finite data sets instead of random variables. Subsequently the covariance and the means are replaced by estimators. For a shorter notation we transform a data vector $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,N})$ containing N entries into a new variable $\mathbf{z_i}$ with (estimated) zero mean:

$$\mathbf{z}_i = \mathbf{x}_i - \frac{1}{N}\sum_j (\mathbf{x}_i)_j. \tag{2.69}$$

For a shorter notation the $\mathbf{z_i}$ are grouped together:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_M \end{pmatrix} = \begin{pmatrix} z_{1,1} & z_{1,2} & \dots & z_{1,N} \\ z_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ z_{M,1} & \dots & \dots & z_{M,N} \end{pmatrix}. \tag{2.70}$$

The covariance can then be estimated by

$$\mathbf{C} = \frac{1}{N+1}\mathbf{Z}\mathbf{Z}^\mathbf{T}. \tag{2.71}$$

In case of large N this does not differ much from the biased estimator $\mathbf{C} = \frac{1}{N}\mathbf{Z}\mathbf{Z}^\mathbf{T}$.

2.3.2.1 Alternative Calculation for Small Datasets

If $N \ll M$, i. e. if we consider a high-dimensional problem and small data sets the above approach is computationally inefficient because of the $\mathcal{O}(M^3)$ scaling of the eigenvalue decomposition. Instead it is more efficient to use a singular value decomposition of the data matrix:

$$\mathbf{Z} = \mathbf{U}\mathbf{S}\mathbf{V}$$

with orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ and a diagonal $M \cdot N$ matrix S. We get:

$$\mathbf{C} = \frac{1}{N}\mathbf{Z}\mathbf{Z}^\mathbf{T} = \frac{1}{N}\mathbf{U}\mathbf{S}\mathbf{V}(\mathbf{U}\mathbf{S}\mathbf{V})^\mathbf{T} = \frac{1}{N}\mathbf{U}\mathbf{S}\underbrace{\mathbf{V}\mathbf{V}^\mathbf{T}}_{=\mathbf{1}}\mathbf{S}^\mathbf{T}\mathbf{U}^\mathbf{T} \tag{2.72}$$

$$= \frac{1}{N}\mathbf{U}\mathbf{S}\mathbf{S}^\mathbf{T}\mathbf{U}^\mathbf{T} \tag{2.73}$$

with $\mathbf{SS^T}$ being diagonal.

During the analysis of molecular dynamics simulations however, N, which corresponds to the number of frames used during the analysis, is usually bigger than $M = 3 \cdot P$ with P denoting the number of particles taken into account during the analysis.

2.3.3 Limitations

Principal component analysis is often a great tool to reduce the number of relevant dimensions of a distribution of data points. However, this is only possible if one looks at distributions that can be well described by a linear combination of a basis set that is (significantly) smaller than the number of dimensions of the whole conformational space. This is usually the case if the eigenvalue spectrum is rapidly decaying. However, if the data points are distributed for example in (high-dimensional) spheres or ellipsoids around the origin, it is impossible to see the structure by projecting the data on a limited set of eigenvectors.

Furthermore when analysing data via principal component analysis it is assumed that the biggest variances are actually important. If the data are subject to noise, this can only be taken for granted for sufficiently large signal-to-noise ratios.

2.3.4 Principal Component Analysis as an Analysis Tool for Molecular Dynamics Simulations

In the case of molecular dynamics simulations, our variables, in the above section denoted as x_i and z_i, correspond to the (Cartesian) coordinates of the single particles in the simulation. The principal component analysis can be used to calculate collective motions of the particles. In most cases, the large eigenvalues and their corresponding eigenvectors belong to larger conformational changes whereas the smaller eigenvalues describe thermal fluctuations and vibrations.

2.4 Parallel Tempering

Since the the absolute value of the exponential in the Boltzmann factor decreases with larger temperature, simulations carried out at higher temperature sample larger regions of the conformational space. At sufficiently high kinetic energies energetic barriers hindering sampling at lower temperatures can be overcome. The same result can be achieved by modifications of order parameters in the Hamiltonian of the system. *Parallel tempering*, often also referred to as *replica exchange*, makes use of this fact whilst still maintaining as a result a canonical ensemble at a fixed temperature/Hamiltonian. The origins of the idea stem from Swendsen and Wang (1986), later the method has been formulated almost in the current form by

Geyer (1991) under the name Metropolis-coupled Markov chain Monte Carlo. An overview can be found for example in (Earl and Deem, 2005).

The basic idea is to simulate m copies of the same system with different Hamiltonians or temperatures. (Whether molecular dynamics or the Monte Carlo methods are used is irrelevant for employing parallel tempering.) If the temperature or order parameter differences of two of the m replicas are small enough, their energy histograms will overlap - or in an equivalent formulation: There will be a none vanishing probability that a conformation sampled in one replica (at the associated temperature or with the associated Hamiltonian) would be sampled in the other replica as well. Therefore at arbitrary times temperature or Hamiltonian exchange attempts between two replicas can be performed. To ensure detailed balance and a proper canonical ensemble the Metropolis criterion (Metropolis et al., 1953) can be applied: Considering the two systems i and j with respective effective Hamiltonians $H_i^* = \frac{H_i}{k_b T_i}$ and H_j^* and the configurations q_i and q_j, the probability that the effective Hamiltonians are exchanged is given by

$$p(i \leftrightarrow j) = p_i(H_i^* \to H_j^*) \cdot p_j(H_j^* \to H_i^*) \tag{2.74}$$

$$= \min\left(1, \exp\left(H_i^*(q_i) + H_j^*(q_j) - H_i^*(q_j) - H_j^*(q_i)\right)\right). \tag{2.75}$$

For the evaluation at a specific temperature, only those configurations that are sampled with the desired effective Hamiltonian are collected.

Given effective swapping[5] the sampling can be improved by a larger degree than the computational time increases due to the additional replicas that are simulated. However, if all replicas are of interest, than any swapping should increase the efficiency.

One disadvantage of using parallel tempering in molecular dynamics simulations is that the resulting trajectories loose their physical meaning. Although collecting all configurations sampled at the desired effective Hamiltonian provides a canonical ensemble, no physical trajectory can be recovered. However, if only the thermodynamics is of interest, this does not constitute a problem.

2.4.1 Combining Replica Exchange and Umbrella Sampling

Umbrella sampling constitutes nothing more than sampling multiple copies of a system with Hamiltonians with different additional biasing potential, which often only differs in the reference point of a harmonic potential. Although sampling along

[5] As there is a lot of discussion in the literature (see for example (Predescu et al., 2005)) about the optimal number of replicas, exchange rates etc. we do not want to attempt to show in detail how effective swapping can be obtained. Instead we only specify that "effective swapping" means that the systems circulates enough between the different effective Hamiltonians that energy barriers can be overcome while a system is sampled for example at high temperatures and then returns (in an other configuration) to the effective Hamiltonian of interest.

the reaction coordinate is improved by the biasing potentials, sampling along the coordinates orthogonal to the reaction coordinate within each umbrella window may be hindered by energetic barriers. Using replica exchange and switching the biasing potentials may help to overcome these barriers if they only exist along a small range of the reaction coordinate.

An application of this can be found in (Wolf et al., 2008)[6].

[6] And, of course, in this work.

3 The Biological System

> "If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat"
>
> *(Douglas Adams)*

3.1 Ion Channels

A biological cell is surrounded by the cell membrane which separates the interior of the cell from the extracellular environment. The major constituent of a cell membrane, the lipid bilayer, is impenetrable for ions. To allow the crucial flow of ions through the membrane, e. g. in neurons, ion channels are embedded within the lipid bilayer. These ion channels consist of either single proteins or more often of an assembly of proteins. In the latter case the channel is usually made up by several identical subunits with the pore being surrounded by the monomers.

Ion channels are to a certain degree selective: Whereas the least selective ion channels are only charge selective and have similar conductivities for different cations or different anions, there are a lot of highly selective channels who only permit the passage of one ion type.

Additionally to the selectivity, ion channels may possess a gating function, i. e. a mechanism that significantly increases or decreases the conductivity. Particularly in nerve cells there are voltage-gated channels, whose opening and closing depends on the membrane potential.

A second type of gating is due to ligand binding. The associated ion channels are called ligand-gated ion channels. The idea of the ligand-induced gating is that the ligand binds at a specific site, which leads to conformational change in at least parts of the protein that closes the channel. Depending on the nature of the channel this could be realised by a plug-like mechanism, or by a simple narrowing of the pore. These kind of channels are of particular interest during drug design.

Cyclic nucleotide-gated ion channels are also controlled by the binding of a ligand, in this case a cyclic nucleotide, but are similar in structure and sequence to the voltage-gated channels. Normal cyclic nucleotide-gated channels (CNG) are absolutely dependent on the nucleotide in their gating behaviour and are impenetrable for ions unless a nucleotide is bound. However, hyperpolarisation and cyclic nucleotide-activated channels (HCN) are merely modulated by the binding of the

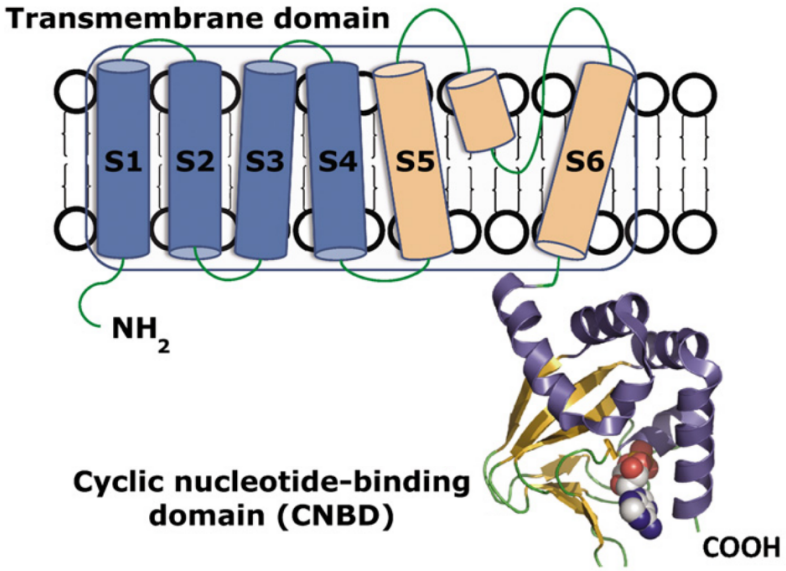


Figure 3.1: Sketch of a monomer of the MloK1 ion channel (Chiu et al., 2007)

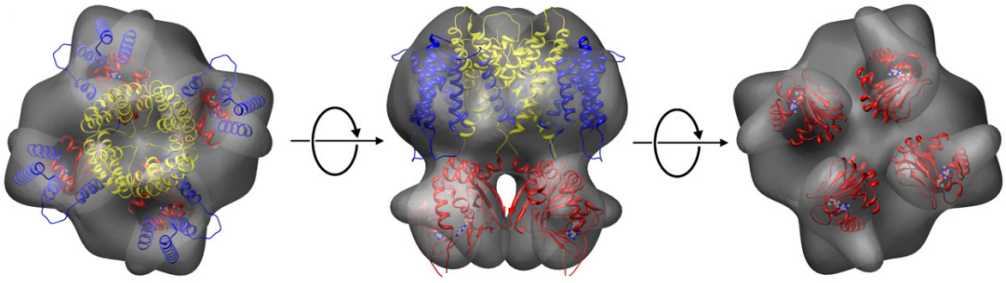
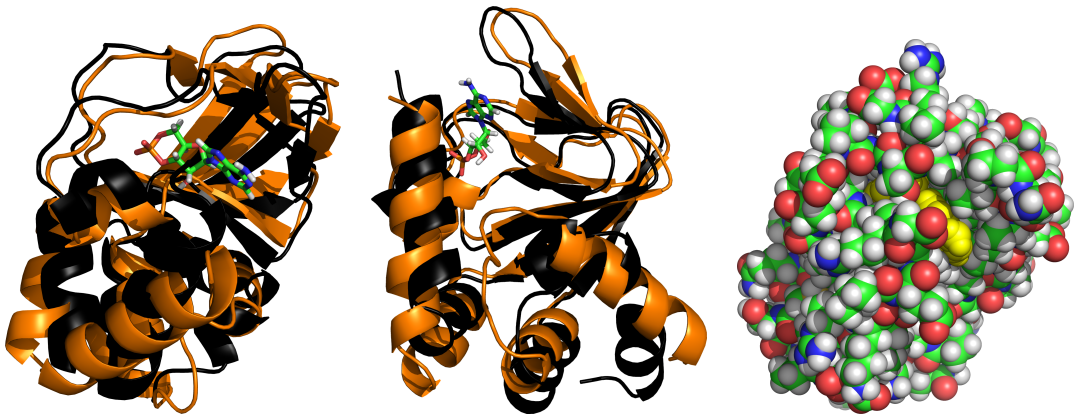


Figure 3.2: Structure of the tetrameric MloK1 (Chiu et al., 2007)

nucleotide and primarily activated by hyperpolarisation of the cell membrane, i. e. an already existing ion flux increases upon binding of a nucleotide.

3.2 MloK1

In this work we focus on a cyclic nucleotide-regulated potassium channel from the bacterium *Mesorhizobium loti*. The cyclic nucleotide binding domain (CNBD) of the channel binds both cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP). The binding of these ligands to the channel as well as their effect on the gating have been analysed in various experiments (Clayton et al., 2004; Cukkemane et al., 2007; Nimigean and Pagel, 2007). In particular, the dependency of the ion flux on the cAMP concentration has been shown (Clayton et al., 2004), and an estimate for the dissociation constant of



(a) View on C-terminus

(b) View on N-terminus

(c) Closed conformation shown in space filling model, ligand shown in yellow.

Figure 3.3: X-ray structures of the opened (orange) and closed (black) CNBD, together with the cAMP molecule from the closed conformation X-ray structure

$70\,\text{nM} < K_D < 110\,\text{nM}$ (Nimigean and Pagel, 2007) has been made.

The CNBD has been expressed separately and crystallised by Clayton et al. (2004). The binding domain has been crystallised with a bound nucleotide (cAMP) and its structure has been resolved at 1.7 Å. We will refer to this extracted X-ray structure as the bound and *closed conformation*.

Since removing the nucleotide from the binding domain without denaturating the protein proved to be infeasible, a mutated version of the binding domain was expressed where one residue (Arg348) was replaced by an alanine. This mutation was motivated by the strong interaction of the Arg348 with the bound nucleotide in the closed, bound configuration, and indeed changing the residue significantly increased the dissociation constant (Nimigean and Pagel, 2007), which allowed obtaining a crystal without a bound ligand. The obtained tertiary structure is assumed to be identical (or at least highly similar) to the structure of the wild type protein without a bound ligand. Therefore the resulting crystal structure of the binding domain will be referred to as the *open conformation*.

In fig. 3.3(a) and 3.3(b) the open (orange) and closed (black) structures of the CNBD are shown, together with the bound cAMP molecule. Figure 3.3(c) shows a calotte model of the closed conformation; the ligand is drawn in yellow. The picture allows a better perception of the ligand protein interactions and the size of the volume filled by the ligand.

The sequence of the whole channel shows homologous parts to eucaryotic K^+ channels (Nimigean et al., 2004; Chiu et al., 2007). Based on these observations,

single particle transmission electron microscopy, and electron crystallography of 2D membrane crystals, a structure has been proposed at 16 Å resolution by Chiu et al. (2007) (see fig 3.2). In the suggested model the channel consists of four monomers which form a tetramer. Each monomer consists of six helices embedded in the membrane, four of which are homologous to voltage-sensitive domains in other K^+ channels. However, whether the whole channel is actually voltage sensitive remains uncertain. The other two helices form the pore complex in the tetramer. The CNBD is connected to the last helix (see fig. 3.1).

Interestingly, the crystallised CNDB in the work of Clayton et al. (2004) forms dimers which motivated speculations about a lever-like gating mechanism. This assumption probably does not hold; in the suggested structure of the entire channel the dimerisation cannot be reproduced.

3.3 Ligand Binding and Conformational Change

3.3.1 Conformational Selection and Induced Fit

The fact that two different conformations for the CNBD are observed in the X-ray structure allows us to address the question how the transition from one conformation to the other happens and if the the induced fit or the conformational selection model, which have been presented in the introduction, are more suited for the binding process.

Regardless of which model is more applicable to the system at hand, for a protein with two states, A and B, such as depicted in fig 1.1 and 1.2) the following holds: With a bound ligand, state B should be relatively preferred compared to a system without a ligand. In terms of Gibb's free energy differences this becomes:

$$\begin{aligned} \Delta\Delta G &= \Delta G_{\text{bound}} - \Delta G_{\text{unbound}} \\ &= (G_{B,\text{bound}} - G_{A,\text{bound}}) - (G_{B,\text{unbound}} - G_{A,\text{unbound}}) \\ &< 0. \end{aligned} \tag{3.1}$$

The two models are probably oversimplified for many real cases. For example, even if the model is in general induced fit-like, the potential energy landscape might change not just after the ligand is at the actual binding site, but already at an earlier stage, when the ligand is only partly bound. The entire transition from state A to state B might thus be a stepwise process over multiple substates between which both, the protein conformation and the ligand position, changes.

3.3.2 Description of Conformational Changes

The CNBD of MloK1 contains $N = 2025$ atoms, resulting in $3N = 6075$ degrees of freedom, minus global translational and rotational degrees of freedom and

minus the bond vibrational degrees of freedom that are eliminated by the use of constraints (see section 2.1.2.1 and section 4.2). Identifying transitions, studying the thermodynamics, and visualising the energetic landscape in such a high-dimensional picture is a considerable challenge and motivates the search for a low-dimensional description.

To address the question whether the process is better described by induced fit or by conformational selection, at least two reaction coordinates, one for the progress of the binding process, and a second one describing the conformational transition from A to B are required.

It is not obvious if two reaction coordinates suffice as descriptors of the binding process and the conformational transition. If the ligand is assumed to be stiff, six dimensions describe both its position and its orientation with respect the protein and the binding site. If spherical symmetry or a confinement to one spatial coordinate can be assumed, the positional coordinates are reduced to one dimension. A measure that combines distance and orientation is the *root mean square deviation* (RMSD) of the ligand with respect to the ligand in the bound configuration. The most simplistic continuous descriptor that ignores orientational degrees of freedom is the distance of the *centre of mass* (COM) of the ligand to the binding site. The free energy landscape along this coordinate shall be analysed as well as the impact of a change in this coordinate on the protein.

The subspace and its dimensionality necessary to describe the conformational change in the protein are a priori completely unknown. However, it is unlikely that the potential energy is sufficiently flat and that the entire thermodynamics can be obtained from straightforward unbiased Boltzman sampling. Therefore a search for a reaction coordinate is motivated which can be used for sampling enhancing techniques such as umbrella sampling. Assessing the quality of a specific possible reaction coordinate has to be done by systematic testing.

It is unlikely that the motion of all atoms within the protein is relevant for a conformational change. For most of the side chains it is reasonable to assume that their relative motion does not change significantly between the two conformations. This assumption would allow a significant reduction of the dimensionality of the problem, from 2025 atoms to 399 backbone atoms.

The vector connecting the two X-ray structures or rather the projections of arbitrary configurations onto this vector provides a linear reaction coordinate where the distance between the two structures is maximised. After a full transition from state A to state B all positions between the two structures along the coordinate have to be visited. Whether the main part of a reaction pathway also lies parallel to this coordinate has to be tested.

If the system moves from one conformation to the other, it necessarily has to move along this coordinate. Therefore we will test the quality of this coordinate by performing umbrella sampling simulations and calculating the PMF (see section 2.2.1) along the coordinate.

4 Methods

"h is an abbreviation for huhohshdhjha"

(sun grid engine help (qstat -help))

4.1 Parametrisation of cAMP

Force field parameters for both protonated and unprotonated cyclic adeonise monophosphate (see fig. A.1 and A.2) were calculated using the general Amber force field (GAFF) (Wang et al., 2004) and the antechamber toolkit (Wang et al., 2006). Atomic point charges were obtained in a restrained fitting procedure in such a way that the resulting electrostatic potential fits best to the electrostatic potential generated by the electronic wave function (RESP charges) (Bayly et al., 1993). The wave function was calculated with Gaussian03 (Frisch et al.) at a Hartree Fock level using the 6-31G* basis set. A tight convergence criterion of 10^{-8} was applied for the self consistent field (SCF) calculations; 6 points per unit area were calculated in the electrostatic potential (ESP) fit. Before the calculation of the electronic wave function a geometry optimisation of the molecule was performed.

4.2 Molecular Dynamics Simulations

All simulations were carried out using Gromacs 4.0 (Van Der Spoel et al., 2005; Hess et al., 2008). Electrostatic interactions were calculated using particle-mesh Ewald (Darden et al., 1993), with a real space cut-off of 1 nm, a grid spacing of 0.13 nm and cubic interpolation. Van-der-Waals interactions were cut off at a distance of 1.6 nm. Non-bonded interactions were calculated using neighbour lists which were updated every 5 time steps. All simulations were performed in the NPT ensemble (constant particle number, pressure and temperature) using the velocity rescaling method for temperature coupling (Bussi et al., 2007) with a heat bath temperature of $T = 300\,\mathrm{K}$ and a coupling time constant of 0.1 ps and Berendsen pressure coupling (Berendsen et al., 1984) with a reference pressure of 1000 hPa and a respective coupling time constant of 1 ps (see also section 2.1.3.2). All systems were simulated in cubic boxes using periodic boundary conditions. Throughout all simulations the TIP3P water model (Jorgensen et al., 1983) was employed. All bond lengths were constrained using the LINCS algorithm (Hess et al., 1997), that means the harmonic pair bond interactions were removed from the force field

potential and replaced by constraints. The equations of motion were integrated using the Verlet algorithm (Verlet, 1967) and a time step of 2 fs was used (see section 2.1.3.1).

Prior to all simulations an energy minimisation of the force field potential was performed using a primitive steepest decent algorithm until the step size reached single point precision, thus moving the system into a local minimum.

All simulations with protonated cAMP were carried out using the amber03 force field (Duan et al., 2003), for better compatibility with the GAFF the amber99sb force field (Hornak et al., 2006) was employed in all simulations with unprotonated cAMP.

4.2.1 Preparation of the CNBD Simulation System

The starting structures were based on the X-ray structures of the wild type and the R348A mutant version of the cyclic nucleotide binding Domain (CNDB) which were determined by Clayton et al. (2004) (PDB codes 1vp6 for the wild type that constitutes the bound, closed conformation and 1u12 for the R348A mutant that is assumed to be similar if not identical to the open, unbound conformation, see section 3.2). The structures contain dimers with identical subunits, but since our interest was focused on the binding of a nucleotide to the CNBD, the second chain was removed in both cases. To obtain identical molecules for both conformations, the residue ALA348 in the mutant (1u12) was replaced with the wild type ARG348 using the YASARA software (Krieger et al.) and ensured identical molecule sizes by removing residues at the N-terminus that were only resolved in one structure. Since the N-terminus of the CNBD is not the real N-terminus of the entire channel, the cut-off point is arbitrary anyway. The modified structure from PDB 1u12 will be referred to as the open structure, the modified structure from 1vp6 will be named the closed structure.

The systems were solvated in water, ions (Cl^-) were added at random places for charge neutralisation.

4.2.2 Ligand Binding Umbrella Sampling

Umbrella sampling simulations (see section 2.2) were carried out using the distance of the COM of the unprotonated nucleotide to the binding site. The binding site was approximated using the centre of mass of three residues that surround the binding site, namely GLY297, ARG307 and SER308. Thus the reaction coordinate is defined as the distance of the ligand COM to the COM of the residues surrounding the binding site. Nine equispaced umbrella windows were employed in which harmonic biasing potentials $U_i(x) = \frac{\alpha}{2}(x - x_{0,i})^2$ were applied. The umbrella window specific reference points $x_{0,i}$ ranged from 0.4 nm to 2.0 nm in constant steps of 0.2 nm, and the spring constant was $\alpha = 1\,{}^{\mathrm{kJ}}/_{\mathrm{mol\cdot nm^2}}$. For all umbrella windows

identical starting configurations were used, obtained by placing the nucleotide at an arbitrary position in front of the binding site at a distance of 1.85 nm from the COM of the three aforementioned residues.

For the analysis, the first 50 ns of each umbrella window trajectory were discarded as equilibration time. A PMF (see section 2.2.1) was calculated using the weighted histogram analysis method (WHAM, see section 2.2.1.1). The WHAM equations were iteratively solved until the differences in the free energy constants between two iterations decreased to a value below $10^{-9}\,\mathrm{kJ/mol}$.

Furthermore, PMFs were calculated for subsets of the data, namely for increasing time intervals from 50-100 ns to 50-400 ns. Another set of PMFs were calculated for non-overlapping, consecutive 50 ns time intervals.

4.2.2.1 Error Estimation

Statistical errors for the potential of mean force were estimated using a frame wise bootstrapping method. For each umbrella frame simulation with a simulation length t and n recorded structures (meaning a data recording interval of t/n) a random sample of $t/t_{autocorr}$ points were drawn where $t_{autocorr}$ denotes the average autocorrelation time of all umbrella windows. The autocorrelation function for a single umbrella window was calculated from the values of $F_{umbrella}(t) = k(x(t) - x_{ref})$, i. e. the biasing force from the umbrella potential as a function of the simulation time, which were obtained during the simulation. The autocorrelation time of a single window was defined as the time where the autocorrelation function dropped below e^{-1}.

Using the bootstrapped data sets, a new PMF was calculated using the WHAM. This procedure was repeated 100 times. From the ensemble of via bootstrapping generated PMFs the standard deviations for each bin point were calculated; these standard deviations were used as an estimate for the statistical error of values of each point within the PMF.

4.2.3 Umbrella Sampling Simulations for Conformational Transition Along Backbone Difference Vector

Umbrella sampling simulations along the vector connecting the configuration of the backbone atoms of the closed structure and the configuration of the backbone atoms in the open structure were performed both with and without a bound nucleotide (both in unprotonated and protonated form). In the following we will refer to this vector as the *backbone difference vector*. 11 umbrella windows with equispaced reference points from $x = -4.8\,\mathrm{nm}$ to $x = 4.8\,\mathrm{nm}$ on the corresponding coordinate were used. The projections of the X-ray structures were by construction of the coordinate symmetric around $x = 0$, viz. at $x = -3.84\,\mathrm{nm}$ for the closed conformation and $x = 3.84\,\mathrm{nm}$ for the open configuration.

The starting structures for the 11 umbrella windows were obtained from inter- and extrapolation of the open and closed structure, followed by an energy minimisation and a subsequent 200 ps relaxation run in explicit water while imposing harmonic position restraints (using a spring constant of 500 $^{kJ}/_{mol \cdot nm^2}$) on all backbone atoms. Prior to any simulation runs, an additional energy minimisation was performed.

For protonated cAMP and the amber03 force field the 11 umbrella windows were simulated for 200 ns using a harmonic biasing potential $U_i = \frac{\alpha}{2}(x(t) - x_{ref,i})^2$ with $\alpha = 1 \frac{kJ}{mol \cdot nm^2}$. This was done for systems both with and without a ligand molecule at the binding site. PMFs were calculated discarding the first 10 ns as equilibration time using the WHAM. Four additional PMFs per system were calculated for data points from the time intervals 0-50 ns, 50-100 ns, 100-150 ns and 150-200 ns.

For unprotonated cAMP and the amber99sb force field Hamiltonian replica exchange umbrella sampling simulations were carried out (see section 2.4) for systems with and without a nucleotide at the binding site. Starting conformations were constructed as described above. Harmonic biasing potentials were applied, using a spring constant of $\alpha = 2\,^{kJ}/_{mol \cdot nm^2}$. Every 100 ps, exchange attempts between the umbrella windows were performed for four element-wise different pairs of umbrella windows. The probability to switch the Hamiltonians H_i, H_j of two systems was calculated via

$$P(i \leftrightarrow j) = \min\left(1, \exp(\beta \Delta_{ij})\right) \tag{4.1}$$

$$\Delta_{ij} = H_i(x) + H_j(y) - H_i(y) - H_j(x) \tag{4.2}$$

$$= \frac{\alpha}{2}\left((x - x_0)^2 + (y - y_0)^2 - (x - y_0)^2 - (y - x_0)^2\right) \tag{4.3}$$

$$= -\alpha(x - y)(x_0 - y_0) \tag{4.4}$$

with x, y being the projections of the configurations i, j onto the backbone difference vector.

For analysis the 100ps trajectories were sorted according to the applied biasing potential. Subsequently, PMFs were calculated using the WHAM.

Systems without cAMP were simulated for 70 ns, systems with cAMP were simulated for 50 ns.

4.2.3.1 Error Estimation

Error estimation was done similar to the ligand binding umbrella sampling simulations with a bootstrapping approach. The number of independent data points within each umbrella window was determined by the autocorrelation times of the projections of the trajectories on the backbone difference vector.[1]

[1] Possible correlations between the umbrella windows in Hamiltonian replica exchange simulations that further decrease the number of independent data points are subject to further studies that are not within the scope of this work.

4.2.3.2 Calculation of Free Energy Differences

The PMF was separated into two substates, an open and a closed substate. The highest local maximum in the PMF was defined to be the boundary of the two substates. The free energy difference between two substates was calculated using equation (2.64):

$$\Delta G = -\frac{1}{\beta} \ln \left(\frac{\sum_{i<j} \Delta x \cdot \exp(-\beta W(x_i))}{\sum_{i>j} \Delta x \cdot \exp(-\beta W(x_i))} \right) \tag{4.5}$$

with the PMF $W(x_i)$ depending on discrete points x_i (equispaced with distance Δx) along the backbone difference coordinate. x_j is the point that separates the substates.

Error propagation from the discretised $W(x_i)$ was done numerically: For each point x_i a $\hat{W}_i$ was drawn from a Gaussian distribution with mean $W_i = W(x_i)$ and standard deviation σ_{W_i}. From the obtained $\{\hat{W}_i\}$ ΔG was calculated. The process was repeated 1000 times; the mean of the ensemble $\{\Delta G_j\}$ was used as an estimate for ΔG, the standard deviation as an estimate for the error in ΔG.

Due to the large uncertainties in the PMF at its boundaries the outer points of the PMF were neglected for the calculation of free energy differences.

4.2.3.3 Multidimensionality

The backbone difference vector umbrella sampling simulation trajectories were used for multidimensional analysis. For this the trajectories from all umbrella windows save the first 10 ns of each trajectory (which was discarded as equilibration time) were collected. For each time frame, the biasing weight

$$w_j(q) = \exp\left(U_j(q) - f_j\right) \tag{4.6}$$

was estimated, where $U_j(q)$ denotes the biasing potential from the jth umbrella window, q the configuration at the given time frame and f_j is the free energy constant obtained from the WHAM (see section 2.2.1.1). Using these weights, a frame-weighted principal component analysis was performed (see section 2.3). The weighted covariance between the motion along two coordinates x and y was calculated via

$$COV_w(x, y) = \frac{\sum_i w_i \cdot (x_i - \overline{x_w})(y_i - \overline{y_w})}{\sum_i w_i}, \tag{4.7}$$

$$\overline{x_w} = \frac{\sum_i w_i \cdot x_i}{\sum_i w_i}. \tag{4.8}$$

The index i iterates over all time frames. For a large number of time frames this is asymptotically unbiased.

Projecting the trajectories of all umbrella windows onto the obtained eigenvectors and building a weighted histogram (with the weights form eq. (4.6)) an unbiased two-dimensional probability distribution is obtained. From this with eq. (2.54) a two-dimensional PMF was calculated. To obtain a contour map from the discrete 2D-PMF, a gridding procedure was applied to interpolate for the spaces between the data points.

4.2.4 Free Simulations

4.2.4.1 Free Binding Simulations

For free binding simulations, the unprotonated ligand was placed in a simulation box containing the CNBD. The approximate distance of the ligand's COM to the binding site was 2.4 nm. This distance ensured the ligand being well outside the binding pocket but kept the distance the ligand has to traverse via diffusion to reach the binding site at a reasonable level. 50 100 ns simulations were carried out.

4.2.4.2 Closed Conformation Simulations

Free MD simulations were carried out starting in the closed configuration taken from the X-ray structure with a bound unprotonated ligand. Four simulations with a simulation length of 50 ns each were performed.

4.2.4.3 Convergence Test of Free Binding Trajectories and Dependence on the Initial Configuration

All free binding trajectories started with the same configuration. To test whether the starting configuration, especially the ligand position and orientation, enforces a specific binding pathway all successful binding trajectories were compared. The RMSD of the ligand in a binding trajectory to a ligand in the bound state was used as measure of the binding process of the ligand. If the initial position does not significantly bias the trajectories and if the trajectories converge to the bound state, the trajectories should differ for large bound state RMSDs and become similar for small ones.

For each binding trajectory we calculated the RMSD of the cAMP molecule with respect to its position in the X-ray structure (after fitting the backbone of the CNBD onto the X-ray structure). From the 50 free binding trajectories we selected those where the ligand RMSD to the bound ligand becomes at least once smaller than 1 nm. From these trajectories we recorded structures every nanosecond and calculated the ligand RMSD to the bound configuration for each recorded structure. Afterwards the structures were sorted according to this RMSD and put into one of 30 bins with a width of 0.1 nm ranging from 0 to 3 nm.

For each pair of structures in one bin, the ligand RMSD between these two structures was calculated. The resulting values were averaged per bin.

4.2.4.4 Estimation of Barrier Height

To estimate the height of energetic barrier from barrier crossings, we modelled the conformational change as a single barrier crossing in a two-state system.

For this we assumed the barrier crossing is best modelled by a Poisson process and that the associated rate is given by

$$k = \omega \cdot \exp(-\beta \Delta G) \tag{4.9}$$

with the attempt frequency ω and the barrier height ΔG. The attempt frequency was approximated with the inverse autocorrelation time, i. e $\omega = t_{ac}^{-1}$.

Thus the probability to observe n transition events in the time span T is given by

$$P(n;k) = \frac{(kT)^n}{n!} e^{-kT} \tag{4.10}$$

$$= \frac{(\omega T \cdot \exp(-\beta \Delta G))^n}{n!} \exp\left(-\omega T \cdot e^{-\beta \Delta G}\right). \tag{4.11}$$

To derive an estimate for a lower border of ΔG, the Bayesian theorem was applied:

$$\rho(\Delta G; n) = \frac{p(n; \Delta G) \cdot \rho(\Delta G)}{p(n)}. \tag{4.12}$$

$p(n)$ and $\rho(\Delta G)$ denote a priori probability (density). Without any further knowledge $\rho(\Delta G)$ was assumed to be constant within the interval $[0, c]$. $p(n)$ had to be obtained by normalising $p(n, \Delta G) \cdot \rho(\Delta G)$.

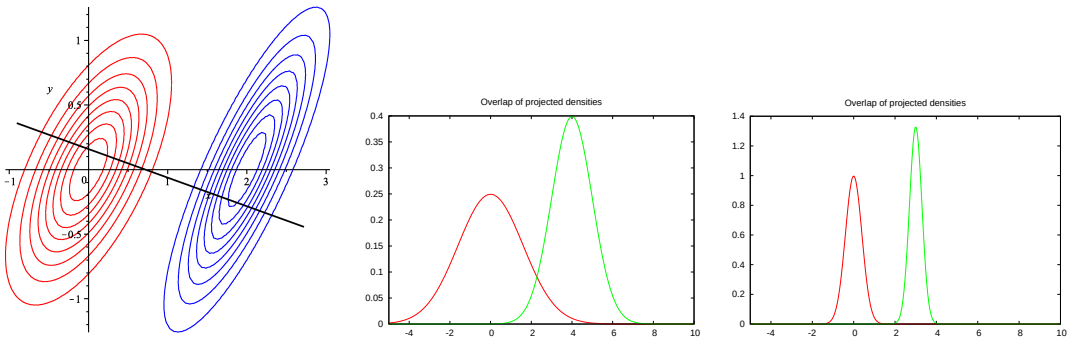
We got for sufficiently large c[2]

$$\rho(\Delta G; n) = \begin{cases} \frac{\beta \cdot \exp(-\omega T e^{-\beta \Delta G})}{E_1(\omega T e^{-\beta c}) - E_1(\omega T)} & \text{for } x = 0, \Delta G < c \\ \beta x \cdot \exp(-\omega T e^{-\beta \Delta G}) \frac{(\omega T e^{-\beta \Delta G})^n}{n!} & \text{for } n \geq 1, \Delta G < c \\ 0 & \Delta G > c. \end{cases} \tag{4.13}$$

4.2.5 Derivation of Optimised Coordinates

Assuming a system that mainly occupies two separate regions in configurational space, the highest energetic barrier in a one-dimensional PMF can be found along the coordinate along which the overlap of the projected probability densities of

[2] A full deduction can be found in A.1



(a) two-dimensional densities (b) projection on x-axis (c) projection on inclined axis

Figure 4.1: Separation of overlapping densities

two separate ensembles (one in the first the other in the second configuration) is minimised. This coordinate we refer to as an *optimal (linear) reaction coordinate*.

The concept is illustrated in figure 4.1. The two functions have their centres on the x-axes, but the maximal barrier is not found when projecting the functions on the x-axis, as sketched in fig. 4.1(b), but after a projection of the function on the diagonal axis, which results in a one-dimensional projection as in fig. 4.1(c). The overlap between the functions is much smaller along this vector, if the functions are interpreted as distribution densities this means a corresponding PMF along this coordinate has a higher local maximum.

Mathematically speaking, we are searching a normalised vector $\mathbf{v}$ in configurational space that minimises the function

$$O(\mathbf{v}) = \int \mathrm{d}x \; \rho_a(x;\mathbf{v}) \cdot \rho_b(x;\mathbf{v}) \tag{4.14}$$

where the probability densities of the separate ensembles $\rho_{a/b}(x;\mathbf{v})$ are given by

$$\rho_{a/b}(x;\mathbf{v}) = \int \mathrm{d}^n q \; \rho_{a/b}(\mathbf{q}) \cdot \delta(\mathbf{v}\cdot\mathbf{q} - x). \tag{4.15}$$

For discrete data points instead of continuous densities, the integral becomes a sum over discrete histogram bins:

$$O(\mathbf{v}) = \sum_i \mathrm{histo}_j(\mathbf{v}\cdot\mathbf{r}_j^a)(i) \cdot \mathrm{histo}_j(\mathbf{v}\cdot\mathbf{r}_j^b)(i). \tag{4.16}$$

Here, $\{\mathbf{r}_j^a\}$ denotes a sample of points in the ensemble a and histo_j a binning operation over the index j, returning a histogram with bin index i.

The search for the optimal reaction coordinate can either be done in the full $3N$-dimensional configurational space or can be restricted to a smaller subspace.

Since most of a protein's dynamics is captured within the first eigenvectors of a PCA on a trajectory generated by computer simulations, a search for the optimal vector in the space spanned by the eigenvectors of a PCA whose corresponding eigenvalues differ significantly from zero is motivated.

For finding the $\mathbf{v}$ that minimises $O(\mathbf{v})$ in the d-dimensional subspace spanned by the first d eigenvectors of a principal component analysis on all backbone atoms of the collected trajectories of the free binding simulations, projections of the free binding trajectories and the trajectories of the closed conformation simulations on the first 20 eigenvectors obtained by the aforementioned PCA were calculated.

We employed an algorithm based on the downhill simplex method by Nelder and Mead (1965). In the following we will sketch the algorithm.

In general, the algorithm finds a (local) minimum of a nonlinear function $f : M \to \mathbb{R}$ where M is subset of a n-dimensional space. Out of $n+1$ points of the original set a so called simplex is constructed. The idea of the algorithm is to replace individual points of the simplex while ensuring that the function values of the simplex points decrease, thus "moving" the simplex towards the minimum. This process does not require any derivatives of the function f. From a starting simplex, a set of rules for the construction of new points is applied, until a minimum is found. In our case the function f is the overlap function described above which works on points on the surface of a d-dimensional unit sphere, or, in an equivalent formulation, on the unit vectors of an d-dimensional space.

The steps of our implementation are as follows:

1. A starting simplex is build up by d randomly chosen points on a $(d-1)$-dimensional unit sphere i. e. d unit vectors. The pairwise scalar products between the starting vectors are constrained within the specific but arbitrarily chosen interval $[0.8, 0.9]$. A "change counter" c is initialised with $c = 0$; a "maximum change counter" c_{max} is initialised with $c_{max} = 8$.

2. For all vectors v_i the overlap function $O(v_i)$ is calculated, the vectors that produce the largest and smallest values for O are identified as v_h and v_l.

3. For v_h we calculate the *reflection* $v_r = \frac{u_r}{|u_r|}$, $u_r = (1+\alpha)\overline{v} - \alpha v_h$ at the mean $\overline{v} = \sum_{i \neq h} v_i/(d-1)$ of all the remaining points in the simplex using a reflection coefficient $\alpha = 1/2$.

4. If $O(v_r) < O(v_l)$, an *expansion* $v_e = \frac{u_e}{|u_e|}$, $u_e = \gamma v_r + (1-\gamma)\overline{v}$ with the expansion coefficient $\gamma = 2$ is calculated. Otherwise we proceed at step 6.

5. If $O(v_e) \leq O(v_l)$, the expansion is accepted and v_h is replaced by v_e. Otherwise, if $O(v_e) > O(v_l)$, the reflection is accepted by replacing v_h with v_r. In both cases the algorithm is continued at step 10.

6. If there is an $i \neq h : O(v_i) > O(v_r)$, the reflection is accepted by replacing v_h with v_r. The algorithm is continued at step 10.

7. If reflections would not produce improvements, contractions are tested: First we define $v_t = \begin{cases} v_r & \text{if } O(v_r) < O(v_h), \\ v_h & \text{otherwise.} \end{cases}$
Using this, the contraction $v_c = \frac{u_c}{|u_c|}$, $u_c = \beta v_t + (1-\beta)\overline{v}$ is calculated.

8. If $O(v_c) < O(v_h)$, the contraction is accepted by replacing v_h with v_c and continuation takes place at step 10.

9. Otherwise a compression is performed by replacing all points by $v_i \leftarrow v_i + \delta \cdot (v_l - v_i)$ using a compression coefficient of $\delta = 1/2$.

10. If there is an $i : O(v_i) = 0$, the algorithm is aborted and the v_i is the desired vector. Otherwise if we have $O(v_n) < O(v_l)$ for any vector v_n that has been added during the last step to the simplex, the "change counter" is set to $c \leftarrow 0$. Otherwise c is increased by 1.

11. If $c > c_{max}$ or if the vectors of the simplex have become linear dependent[3], a new simplex is created by keeping the best vector and drawing $d-1$ new random vector whose scalar product with the kept vector is within the interval $[0.9, 0.99]$.
We set $c \leftarrow 0$, $c_{max} \leftarrow c_{max} + 1$ and continue at step 2.

12. Otherwise the algorithm is repeated at step 2 with the new simplex.

For the construction of the histograms, we chose the interval $[-10\,\text{nm}, 10\,\text{nm}]$ for the x-axes and 40 bins for the histogram. The method was applied for $d = 3, 4, 5, 6, 7$ and 8. ρ_a was approximated by the data obtained during the collected trajectories of the "closed conformation runs", the free binding runs gave an estimate for ρ_b.

4.2.6 Umbrella Sampling Along Optimised Reaction Coordinate

Umbrella sampling simulations were performed along the optimised reaction coordinate calculated in the three-dimensional subspace spanned by the first three eigenvectors from the PCA on the collected trajectories of the free binding simulations.

A set of 25 starting structures was generated by selecting a snapshot of a trajectory of the bound conformation simulations and using essential dynamics sampling (Amadei et al., 1996) to drive the system, i. e. the backbone atoms, along the new

[3] Checking for linear dependency is done by testing if the smallest value of a singular value decomposition of the matrix build from the unit vectors is below a certain threshold.

reaction coordinate both towards the open configuration as well as in the opposite direction. For the former a 20 ps simulation was performed with linear expansion of 0.001 nm per simulation step along the new reaction coordinate. From the obtained trajectory 20 structures were recorded in intervals of 1 ps, yielding configurations with almost equispaced projections along the new reaction coordinate. For the latter case, a 4 ps simulation was performed with a linear expansion of -0.001 nm per step along the new coordinate. Snapshots taken every ps yield another 4 configurations, making 25 structures in total (including the starting conformation)

The obtained structures were – after energy minimisation – used as starting structures and their projection on the optimised reaction coordinate were used as reference points for biasing potentials in the umbrella sampling simulations.

Another set of starting and reference structures was obtained by selection a structure taken from the free binding simulations trajectories where binding of the ligand to the binding domain had already occurred. Similar to the method sketched above, structures were generated by using essential dynamics to drive the systems along the new reaction coordinate both towards the closed configuration as well as to the other directions. 22 structures with projections from -6.7 nm to 3.74 nm on the new reaction coordinate were generated this way.

For the first set of simulations, all 25 umbrella windows were simulated for 420 ns. Biasing potentials $U_i(x) = \frac{\alpha}{2}(x - x_{0,i})^2$ with $\alpha = 10\,{}^{\mathrm{kJ}}/_{\mathrm{mol\cdot nm^2}}$ were applied within the individual umbrella windows. The equispaced reference points $x_{0,i}$ had a distance of 0.5 nm, thus covering a range from $x_{0,0} = -6.517$ nm to $x_{0,24} = 5.451$ nm along the optimised reaction coordinate.

PMFs for consecutive 50 ns time windows were calculated using the WHAM. An additional overall PMF over the entire time span was calculated discarding the first 300 ns as equilibration time.

For the second the set of starting structures, 22 umbrella windows were simulated for 300 ns using biasing potentials with identical spring constants and reference points from $x_{0,0} = -6.726$ nm to $x_{0,24} = 3.74$ nm along the optimised reaction coordinate.

Analog to the former simulation set, PMFs for consecutive 50 ns time frames were calculated together with an overall PMF taking into account the entire simulation apart from the first 200 ns using the WHAM.

Error estimation for the PMFs which were based on larger time windows was done exactly as in the former simulations via a bootstrapping procedure using only data points assumed to be independent. The autocorrelation time along the optimised reaction coordinate of the individual umbrella windows was chosen as the time where the autocorrelation function had dropped below $e^{-1} \approx 0.37$. The number of independent data points was determined by the average autocorrelation time of all umbrella windows.

5 Results & Discussion

"But it works - if you press start, it starts to wiggle."

(Timo Graen)

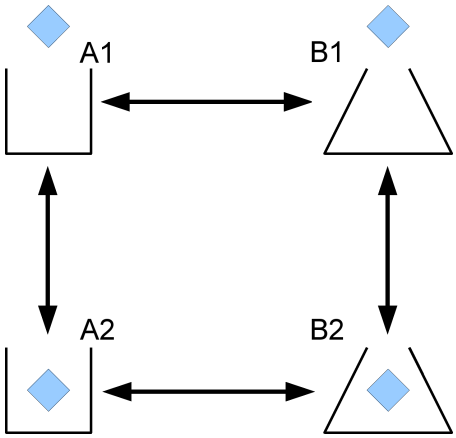


Figure 5.1: Substates in a ligand (blue) binding process with conformational change

To find a good model for the binding process computer simulations have been performed to identify descriptors for the transitions between the substates sketched in fig. 5.1 and to calculate the potential of mean force along these coordinates.

5.1 Ligand Binding Umbrella Sampling

To describe the binding of the ligand to the protein in the open conformation – which corresponds to the vertical transitions in fig. 5.1 – the distance of the ligand to the binding site was chosen as a continuous descriptor of the ligand binding.

Using the distance of the ligand to the binding site, approximated by the centre of mass of GLY297, ARG307 and SER308, as a reaction coordinate for the binding of the ligand, a PMF has been calculated along this coordinate (see section 4.2.2), shown in fig. 5.2. The average autocorrelation time that was used for error estimation was 18 ns. The PMF shows a minimum at $x \approx 0.6\,\mathrm{nm}$ with a depth of $\sim 5\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$. This minimum corresponds to the ligand being in the bound state. Its position on the x-axes differs from $x = 0\,\mathrm{nm}$ due to the fact that the COM of the

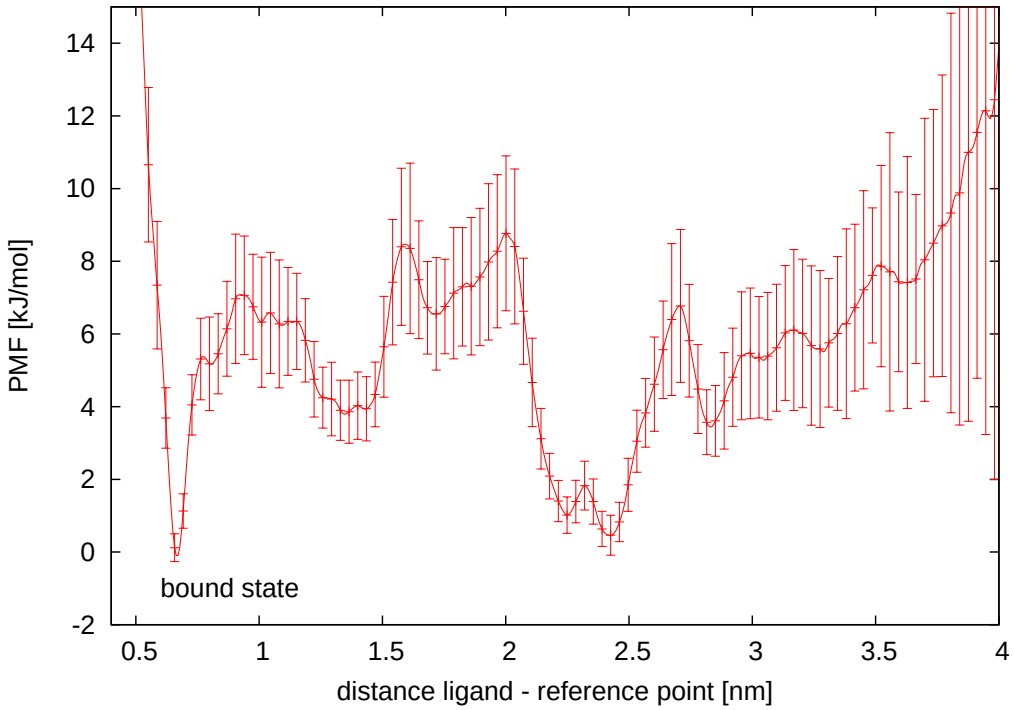


Figure 5.2: PMF along the distance of the ligand to the binding site.

aforementioned residues is not exactly identical to the site occupied by the ligand COM in the bound state.

Although the distance coordinate is one-dimensional, a biasing potential that fixes the ligand COM to a certain distance does not fix it to a specific point in the three-dimensional space, but only to a sphere whose radius is given by the COM distance. Without any protein ligand interactions we thus have a purely entropic contribution to the PMF, which is given by $W_{rad}(x) = -2kT \ln(x)$. Due to the shape of the protein this sphere is only partially accessible. For larger distances, however, larger parts of the sphere become accessible which introduces a drop in the PMF. The minima around $x = 2.3\,\text{nm}$ in fig. 5.2 can be attributed to this effect: At this point the ligand is well outside the binding site, can move freely in water and can interact with other parts of the protein surface. For regions $x > 2.5\,\text{nm}$, however, the PMF becomes unreliable since all biasing potentials had reference values $x_0 \leq 2.0\,\text{nm}$.

To test whether the PMF is converged, additional PMFs over shorter time windows have been calculated with the goal to observe possible trends and changes in the PMF. Figure 5.3 and 5.4 shows the temporal development of the calculated PMF over simulation time. Figure 5.3 shows PMFs over increasing time spans, whereas fig. 5.4 depicts PMFs calculated over consecutive and non-overlapping time periods. Since the PMF is only fixed to an additive constant, the offset on the y-

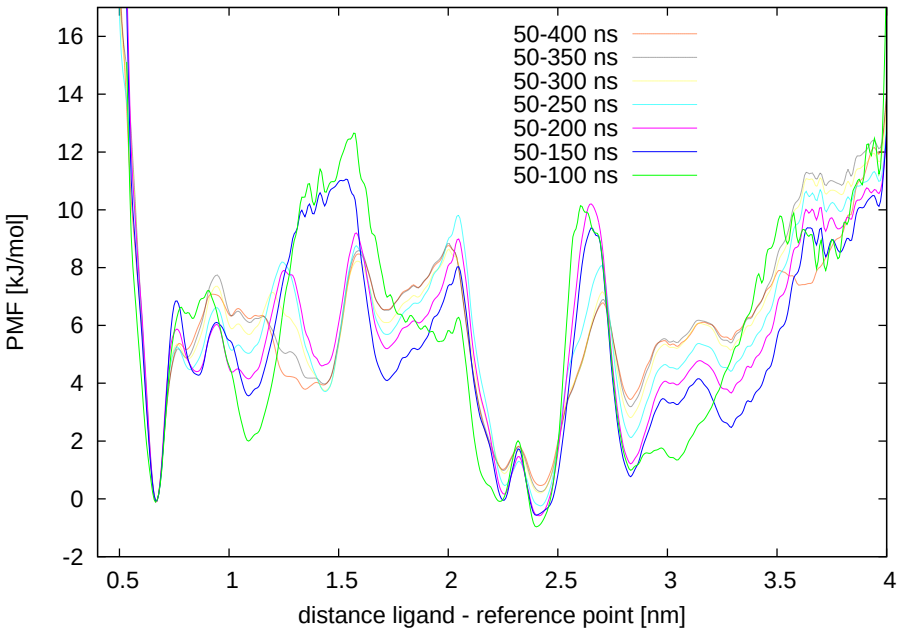


Figure 5.3: PMFs along the ligand distance coordinate calculated using increasing simulation time windows

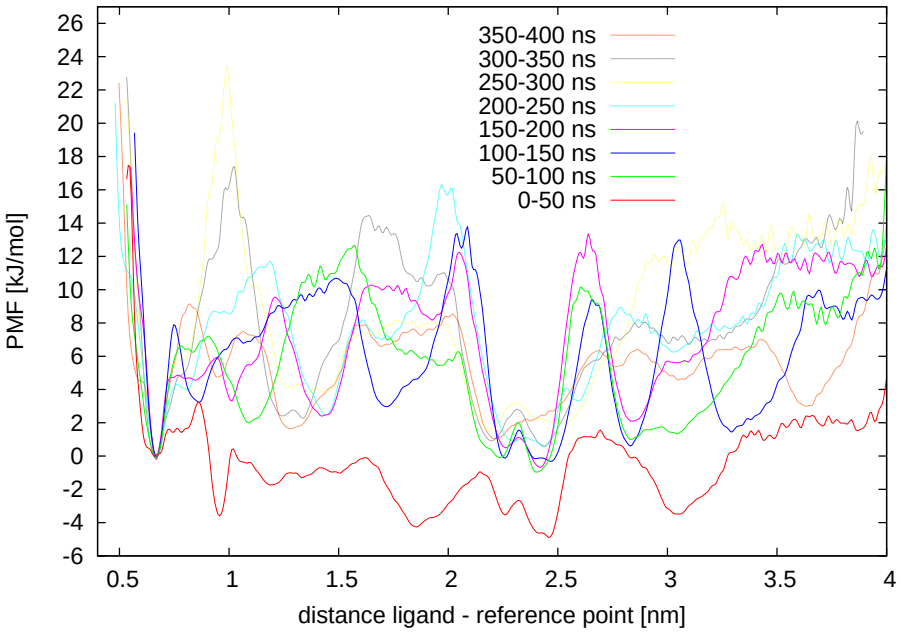


Figure 5.4: PMFs along the ligand distance coordinate calculated using consecutive simulation time windows

axes is arbitrary. It was chosen in such a way that both offset and absolute value in the minimum near $x = 0.6\,\text{nm}$ are zero.

The PMF in fig. 5.3 over 350 ns differs at most points only little from the PMFs over 300 ns or 250 ns and even at points where differences can be noted (between $x = 1\,\text{nm}$ and $x = 1.5\,\text{nm}$) these are within the error bars in figure 5.2. Although the fluctuations in fig. 5.4 are very large, no global trend can be observed apart from the PMF of the first 50 ns, whose underlying data points have therefore been discarded as equilibration time for the overall PMF.

The lack of an overall trend over simulation time and apparent convergence indicate that the PMF is reasonably converged in the sense that the error bars in 5.2 are reliable estimates for the uncertainty in the PMF. Furthermore an estimate for the equilibration time is obtained.

A potential issue with the reaction coordinate is the fact the positions of the reference residues are also subject to fluctuations relative to the rest of the protein. If the position of the actual binding site is constant during these fluctuations, the measured distribution of ligand distances along the distance coordinate is artificially broadened, which again influences the PMF.

The results show that the bound state of the ligand is indeed associated with a minimum in the PMF. The analysis so far does not take into account any conformational changes, which correspond to horizontal changes in fig. 1.2. The free energies associated with this transition are presented in the next section.

5.2 Backbone Difference Vector Umbrella Sampling

For transitions from the open to the closed conformation of the protein, meaning transitions from A1 to B1 and A2 to B2 in the picture of fig. 5.1, the backbone difference vector between the X-ray structures is tested as a reaction coordinate (see section 4.2.3). To calculate the free energy differences between open and closed conformation and to test whether the backbone difference vector is a good choice for a one-dimensional reaction coordinate, PMFs along this coordinate were calculated. This was done for systems both with and without a bound ligand. For the ligand, both the protonated and the unprotonated version were employed (see section 4.1 and 4.2.3).

Figure 5.5 shows the PMFs for systems with and without unprotonated cAMP. Also drawn are the projections of the X-ray structures on the backbone difference vector, as well as the reference points of the 11 umbrella windows used during the umbrella sampling simulations for the calculation of the PMF. The left X-ray structure, labelled by the circle at $x \approx -4\,\text{nm}$ in fig. 5.5, is the closed conformation, whereas the right one, labelled by the circle at $x \approx 4\,\text{nm}$ in the same figure, corresponds to the open conformation. The PMFs are only fixed up to a constant, the free gauge parameters are chosen in such a way that the minima of both

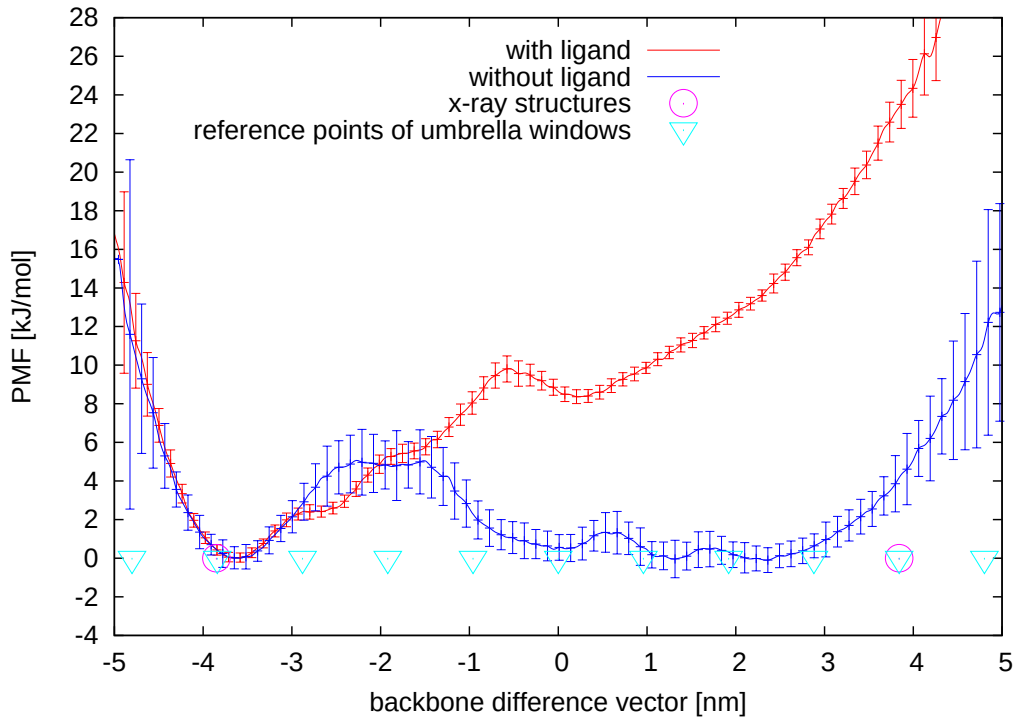


Figure 5.5: PMF along backbone difference vector with unprotonated cAMP

PMFs near the closed conformation X-ray structure have a value of $0\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$. The autocorrelation times, which were used for error estimation (see sections 4.2.2.1 and 4.2.3) were $t_{ac} = 800\,\mathrm{ps}$ for the system with ligand and $t_{ac} = 2800\,\mathrm{ps}$ for the system without a ligand. These short autocorrelation times and the resulting small error bars in PMFs are to be taken with a grain of salt: As explained in section 4.2.2.1, the simulation data was obtained in Hamiltonian replica exchange umbrella sampling simulations (see section 2.4), where possible correlations between the umbrella windows were not taken into account.

Both PMFs show a minimum on the left side between $x = -3.6\,\mathrm{nm}$ and $x = -3.5\,\mathrm{nm}$. For the system without a bound ligand (blue) a barrier can be noted around $x = -2\,\mathrm{nm}$ on the backbone difference vector coordinate with a height of $\Delta E \approx 5\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$ with reference to the minimum near the closed conformation. For higher values ($x > -1\,\mathrm{nm}$) a broad "minimum region" is observed, that extends to $x = 3\,\mathrm{nm}$ on the coordinate. For the system with a ligand (red) a local maximum at $x = -0.5\,\mathrm{nm}$ was obtained. However, only a small local minimum at $x = 0.5\,\mathrm{nm}$ with an energy difference of $\Delta E < 2\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$ is obtained in region where $x > 0\,\mathrm{nm}$.

For a second simulation set where protonated cAMP and a different force field, namely the amber03 force field were emplyed, (see section 4.2.3) PMFs were obtained for systems both with and without a bound ligand. The result is shown in fig. 5.6.

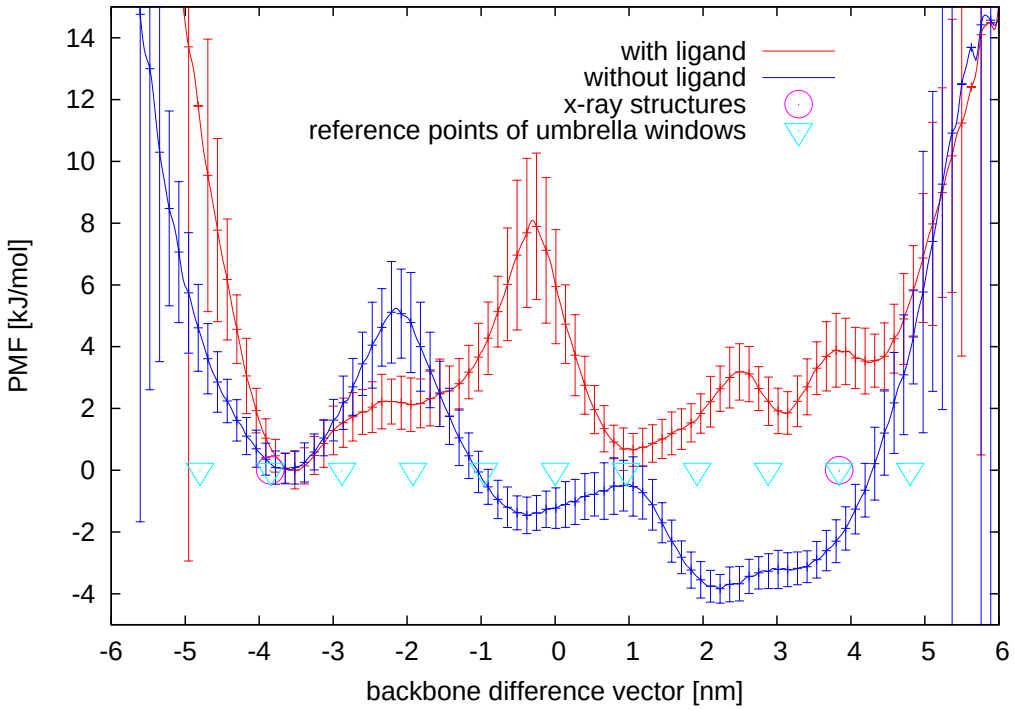


Figure 5.6: PMF along backbone difference vector with protonated cAMP

Both PMFs show a minimum near $x = -3.5\,\text{nm}$ similar to the former simulations. The PMF for the system without ligand (blue) shows a local maximum around $x = -2\,\text{nm}$ with $\Delta E \approx 5\,\text{kJ/mol}$. The PMF in the region $-1\text{nm} < x < 4\,\text{nm}$ looks similar to the corresponding region in fig. 5.5, however, the minimum in the interval $[-2\text{nm}, 2\text{nm}]$ is more distinct and with -3.8 kJ/mol deeper.

For the PMF with a bound ligand (red) we again obtain a local maximum at $x \approx -0.3\,\text{nm}$, similar to fig 5.5, but its value is with 8 kJ/mol lower than for the PMF for unprotonated cAMP (10 kJ/mol). The most important difference however is the fact the entire "open conformation" region between $x = 0\,\text{nm}$ and $x = 4.5\,\text{nm}$ shows three minima with values for the PMF comparable to the minima in the closed conformation (1, 3.1 and 4.4 kJ/mol).

If we define the region left to the highest local maximum as the closed conformation region and right to the maximum as the open conformation region and use eq. (2.64) to calculate the free energy differences between the two substates, we obtain for the Hamiltonian replica exchange umbrella sampling simulations with unprotonated cAMP:

$$\Delta G_{\text{no ligand}} = G_{\text{closed conf.}} - G_{\text{open conf.}} = (2.94 \pm 0.19)\,\text{kJ/mol} \tag{5.1}$$

$$\Delta G_{\text{ligand}} = (-7.81 \pm 0.10)\,\text{kJ/mol}, \tag{5.2}$$

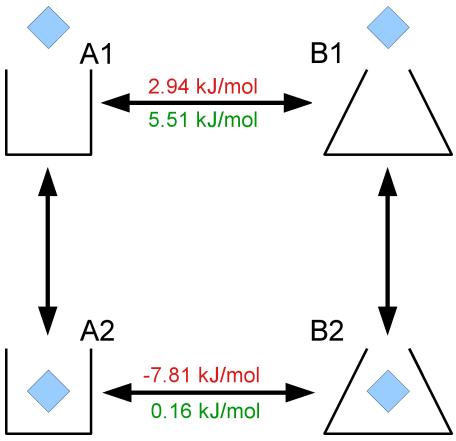


Figure 5.7: Substates in the binding process with free energy differences along backbone difference vector from simulations with unprotonated cAMP (red) and protonated cAMP (green)

and for the simulations with protonated cAMP:

$$\Delta G_{\text{no ligand}} = (5.51 \pm 0.24)\,\text{kJ/mol} \tag{5.3}$$
$$\Delta G_{\text{ligand}} = (0.16 \pm 0.15)\,\text{kJ/mol}. \tag{5.4}$$

We get:

$$\begin{aligned}\Delta\Delta G &= \Delta G_{\text{ligand}} - \Delta G_{\text{no ligand}} \\ &= (-5.35 \pm 0.32)\,\text{kJ/mol} \qquad \text{for the unprotonated cAMP,} \qquad (5.5)\\ \Delta\Delta G &= (-10.75 \pm 0.39)\,\text{kJ/mol} \qquad \text{for protonated cAMP.} \qquad (5.6)\end{aligned}$$

In fig. 5.7 these numbers are associated with the corresponding reactions. Negative numbers mean the right side is favourable in terms of free energies.

The PMFs presented in this section show that the projection of the closed conformation as obtained from the X-ray structure coincides with a minimum in all PMFs of fig. 5.5 and 5.6. This means that the closed conformation is also stable in the MD simulations. On the other hand, the projection of the open conformation X-ray structure does not coincide with a minimum in any of the PMFs. However, all PMFs show an energetic barrier that separates the minima around the closed conformation from a region closer to the open conformation X-ray structure. Therefore it makes sense to interpret the region right of the barrier as open conformations. Although the free energy differences in eq. (5.1) - (5.6) strongly differ, eq. (3.1) still holds. Furthermore it should be noted that the estimates for the free energy difference between open and closed state are also influenced by the shape of the PMF at its ends, especially beyond the last umbrella potential reference point, where it cannot be assumed to be exact.

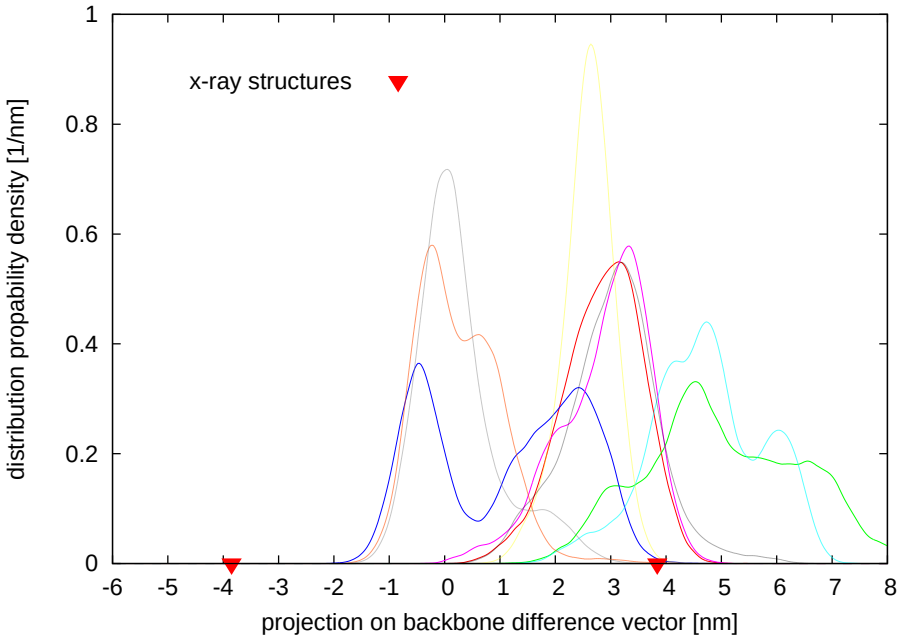


Figure 5.8: Projections of umbrella sampling trajectories onto the backbone difference vector.

The main features of the PMFs for the system without a ligand a preserved in both simulations. The slight bias to the open conformation is either due to the force field or it is an convergence issue. The simulations with a ligand however show a significant difference. A force field induced trend to the closed conformation for the simulations with unprotonated cAMP alone hardly justifies the difference and it is not obvious why the protonation state of the ligand alone should destabilise the open conformation to such an extent. Since PMFs for consecutive time frames (not shown) do not reveal a trend of parts of the PMF to higher or lower energy values, the possibility of sampling issues in the subspace orthogonal to the backbone difference vector has to be studied.

If the reaction coordinate was well suited in the sense that it contains the maximal energetic barrier that has to be crossed during a conformational change of the protein (5-10 $^{\mathrm{kJ}}/_{\mathrm{mol}}$ for the ligandless system, 2-10 $^{\mathrm{kJ}}/_{\mathrm{mol}}$ for the systems with ligand) and given a autocorrelation time of 10 ns, it should be possible to observe transitions in simulation time.

5.2.1 Projections of Ligand Binding Umbrella Sampling Simulations on Backbone Difference Vector

The results of the previous section show energetically separated open and closed conformations. The sampling along the entire range of the reaction coordinate was

made possible by umbrella sampling simulations. During the simulations with a bound ligand the ligand position does not change significantly, thus we are certain that the reactions depicted by the horizontal arrows in fig. 5.1 were analysed. During the ligand binding umbrella simulations (section 4.2.2 and 5.1) however, the system was only constrained along the ligand distance coordinate, whereas not constraints were imposed upon the conformation of the protein.

To find out whether the conformation does change from the open, starting conformation during the ligand binding umbrella sampling simulations the trajectories obtained for each umbrella window are projected on the backbone difference vector and the resulting probability distributions along the coordinate are calculated (figure 5.8).

The included positions of the X-ray structures and fig. 5.5 and 5.6 show that the closed conformation is never reached within the ligand binding simulations. Therefore the PMF in fig. 5.2 shows only the contributions of the position of the ligand and no (significant) contribution of the conformational change of the protein. The obtained PMF (fig. 5.2) therefore corresponds to the left arrow in fig. 5.1, meaning the transition form A1 to A2.

5.2.2 Multidimensionality

In a one-dimensional PMF all degrees of freedom along coordinates orthogonal to the selected reaction coordinate are integrated out. Multiple local minima in the high-dimensional potential energy landscape that have the same value projected on the reaction coordinate are thus no longer separable in the one-dimensional picture.

In section 5.2 the conformational change is only described by the change along a one-dimensional coordinate, although the backbone of the protein has many more degrees of freedom. By projecting the trajectories on the first eigenvectors of a frame weighted PCA and performing a reweighted binning along this eigenvectors, a multidimensional picture is regained (see section 4.2.3.3). This has been done for the trajectories obtained during the backbone umbrella sampling simulations with protonated cAMP for which the corresponding one-dimensional PMF is shown in fig. 5.6.

The first eigenvector obtained from the frame weighted PCAs, both with and without cAMP, is very similar to the backbone difference vector, the scalar products are $\mathbf{e}_{1,\text{with ligand}} \cdot \mathbf{v}_{bb} = 0.893$ and $\mathbf{e}_{1,\text{without ligand}} \cdot \mathbf{v}_{bb} = 0.898$, respectively.

Figure 5.9 depicts a contour map of the 2-dimensional PMF along the backbone difference vector and the 2$^{\text{nd}}$ eigenvector of the frame weighted PCA on the respective simulations. Figure 5.10 shows the same for the system without a ligand. It should be noted that the 2$^{\text{nd}}$ eigenvectors are not identical and with a scalar product of 0.43 only slightly similar. The absolute offset of the PMFs is not set to a specific value, which means that the absolute values of the PMF in one figure

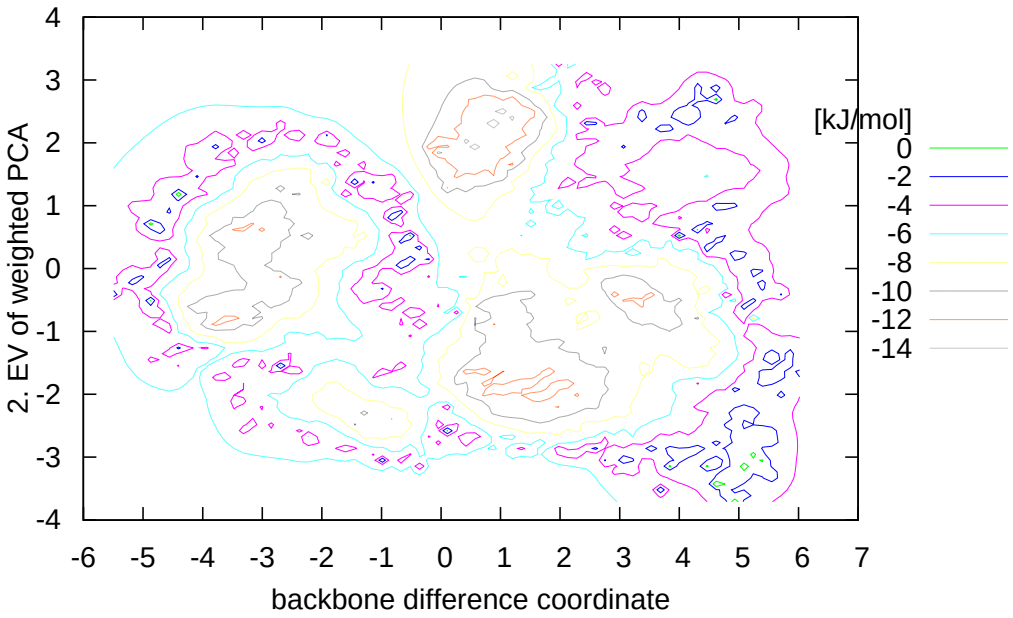


Figure 5.9: 2D-PMF from backbone difference vector umbrella sampling simulations for systems with a bound ligand

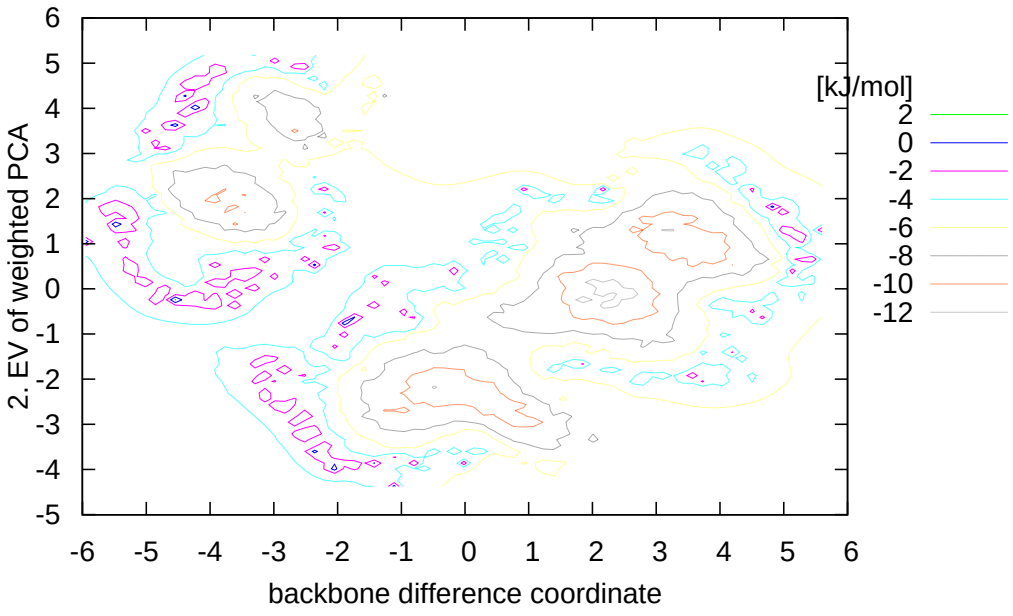


Figure 5.10: 2D-PMF from backbone difference vector umbrella sampling simulations for systems without a ligand

cannot be compared to the values in the other figure.

Although there is no reason to assume that the PMF is fully converged in any direction orthogonal to the reaction coordinate and although it is possible that there exist energetic minima beyond the boundaries not visible in the PMF because the separating barriers have not been crossed, the 2-dimensional PMFs show that the closed and open conformations occupy larger regions in the conformational space that are not best separated by a energetic barrier strictly orthogonal to the backbone difference vector.

The gridding performed to obtain a contour map allows no exact predictions in little sampled regions. Nevertheless we identify in figure 5.10 a diagonal barrier from the lower left corner of the figure to the upper middle that separates the closed conformation minima in the upper left corner and the open conformation minima in the lower right corner. This barrier has to be higher in terms of energy values than a barrier along the backbone difference vector because, though being separated in the two-dimensional space, the closed conformation minima and open conformation minima partly overlap when being projected on the backbone difference vector.

The results show that both open and closed conformations occupy extended regions in conformational space that are to a certain degree separated even in a projection onto the backbone difference vector, but that the highest barrier cannot be expected to be found along this coordinate.

The question of finding a better reaction coordinate along which a higher barrier can be found and which provides better separation of the bound state will be addressed in section 5.3.2.4 using an extensive sampling especially of the open conformation.

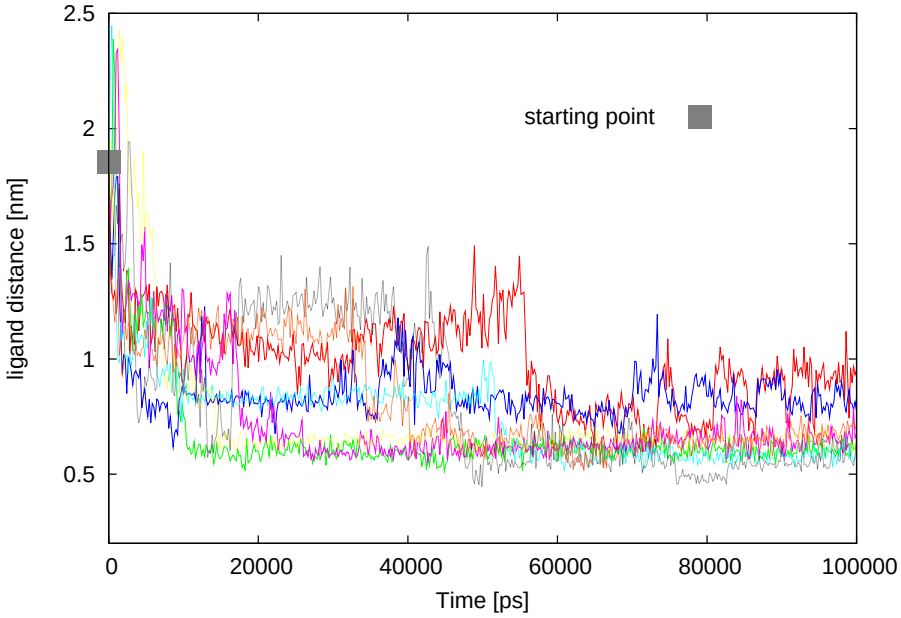
First the question is addressed if spontaneous ligand binding and conformational transitions can be observed in free MD simulations.

5.3 Free Binding Simulations

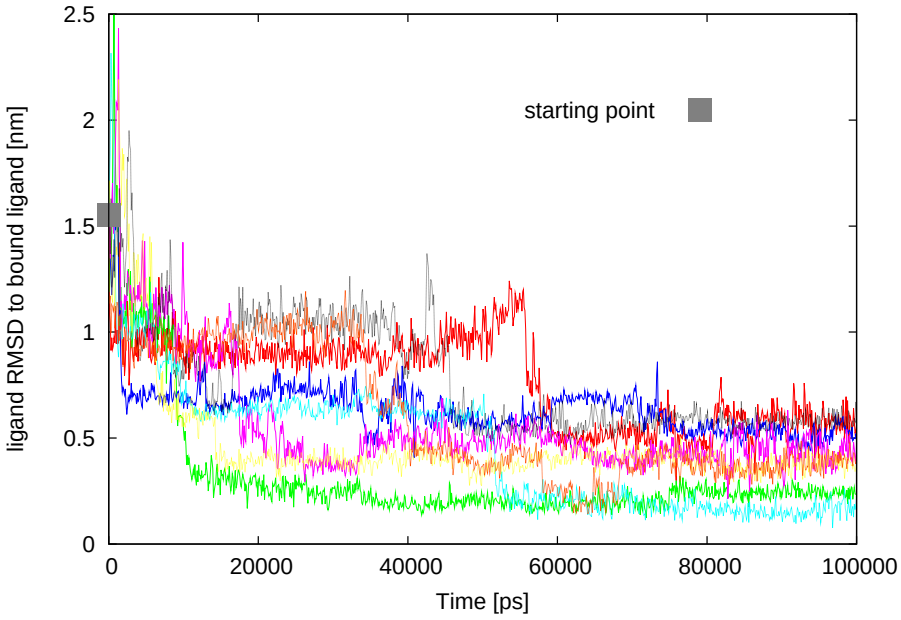
The starting point of 50 free binding simulations (see section 4.2.4) is the open protein conformation with an unbound ligand, i. e. state A1 in fig. 5.1.

In ten out of fifty free MD trajectories the ligand COM distance to the COM of the residues that represent the binding site becomes smaller than 0.76 nm for at least 2 ns. In seven further trajectories we observed a weaker binding with a COM distance of less than 0.94 nm for at least 2 ns.

Figure 5.11 shows examples of trajectories where the ligand successfully binds at the binding pocket. In fig. 5.11(a) the distance of the ligand to the COM of the residues surrounding the binding site is plotted (i. e. the position of the system along the coordinate used for fig. 5.2). Figure 5.11(b) shows the RMSD of the ligand from the free binding trajectory to the ligand in the bound configuration (taken from the X-ray structure) after fitting the backbone atoms of the protein



(a) Ligand distance over simulation time



(b) Ligand RMSD to bound state over simulation time

Figure 5.11: Free binding trajectories showing successful binding

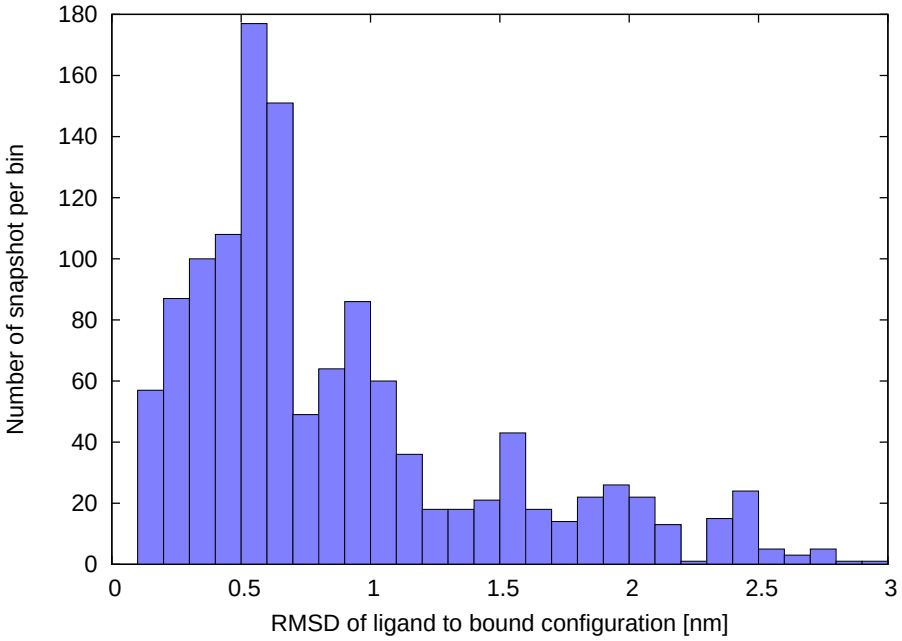


Figure 5.12: Histogram of ligand RMSDs to bound configuration

to the backbone atoms of the protein in the X-ray structure.

The two plots show high similarities which is due to the fact that the value for the RMSD is mainly dominated by the distance of the ligand to binding site. Conversely, since the RMSD of the ligand to the bound state is in principal a better measure of the binding progress of the ligand, the high similarities show that the ligand distance is actually a suited descriptor for the ligand binding process. This justifies the choice of the ligand distance as a reaction coordinate for the PMFs presented in section 5.1.

5.3.1 Convergence of Binding Trajectories and Independence of Starting Values.

Multiple trajectories where successful binding occurs show that the binding of a ligand can be simulated using free MD simulations. This supports the ligand binding umbrella sampling simulations that yield a PMF with a minimum in the bound configuration. The time until a binding occurs depends of course on the concentration of cAMP, which is equivalent to the size of the simulation box of the MD simulation.

For all 50 free binding trajectories the initial conditions are – apart from the initial velocities – identical. Therefore it has to be checked whether the initial conditions of the MD simulations, especially the ligand position, have an influence on the ligand binding.

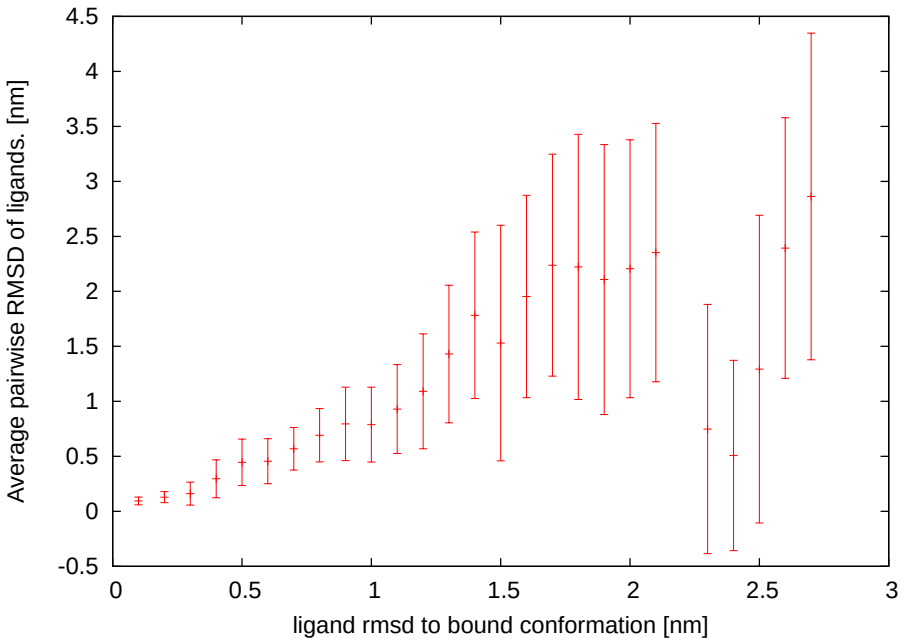


Figure 5.13: Average pairwise ligand RMSD of two different structures vs. ligand RMSD to bound state

This is done by using the RMSD of the ligand to the bound configuration as a measure of the binding process. A histogram of those RMSDs from recorded structures of successful binding trajectories is shown in fig. 5.12. The average pairwise RMSD from the ligand in one structure within a bin of fig. 5.12 to the ligand in an other structure within the same bin is plotted in fig. 5.13 for each bin. Additionally, the standard deviation of the pairwise RMSDs ensemble of each bin is plotted for each bin.

The average pairwise ligand RMSD decreases with decreasing ligand RMSD to the bound state with an exception for values around 2.4 nm almost monotonic.

This shows that there are multiple ligand binding pathways always converging in one state. Therefore a possible hypothesis that the starting position of the ligand enforces a specific binding pathway can be rejected. On the contrary the results strongly support that the starting conditions have a negligible influence on the observed thermodynamics and kinetics.

The standard deviations of the ensemble of pairwise RMSDs should not be confused with error bars in the traditional sense. The pairwise RMSD ensembles are not Gaussian distributed around the mean which becomes especially apparent for bins where the standard deviations lap into regions of negative RMSDs which are of course by definition not possible. Small pairwise ligand RMSDs even for large ligand RMSD to the bound state are most probably due to consecutive structures from the same trajectory where the ligand position does not change significantly.

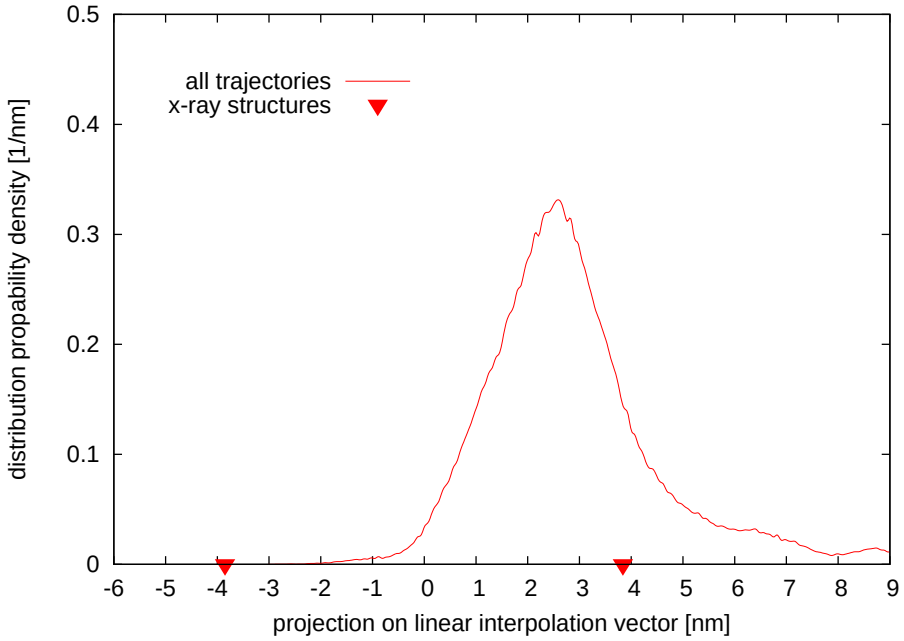


Figure 5.14: Histogram of projections of free binding trajectories on backbone difference vector

5.3.2 Protein Conformation

In the previous section it is shown that the binding of the ligand to the protein can be simulated in free MD simulations. In a second step, the question whether a conformational change in the backbone occurs during free simulation has to be addressed.

If the maximum energetic barrier between open and closed conformation is lower than $\Delta G = 10\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$ and the barrier crossing attempt frequency is given reasonably estimated by $\omega = 1/t_{ac} \approx 5 \cdot 10^7/s$, t_{ac} being the autocorrelation time, then, assuming the barrier crossings are modelled by a Poisson process, a transition should be observable within a total simulation time of $T = 5\,\mu$s with a probability of

$$p > 1 - \exp\left(-\omega T \cdot e^{-\beta \Delta G}\right) = 98.9\%. \tag{5.7}$$

Figure 5.14 shows a histogram of projections of the free binding trajectories (including those where no actual binding occurs) on the backbone difference vector, comparable to fig. 5.8. It shows that the closed conformation is not reached in any binding trajectory.

This result cannot be explained by a PMF as in figures 5.5 and 5.6 where only relative low energetic barriers can be found. With this result it has to be assumed that there exists at least one higher energetic barrier as was already conjectured

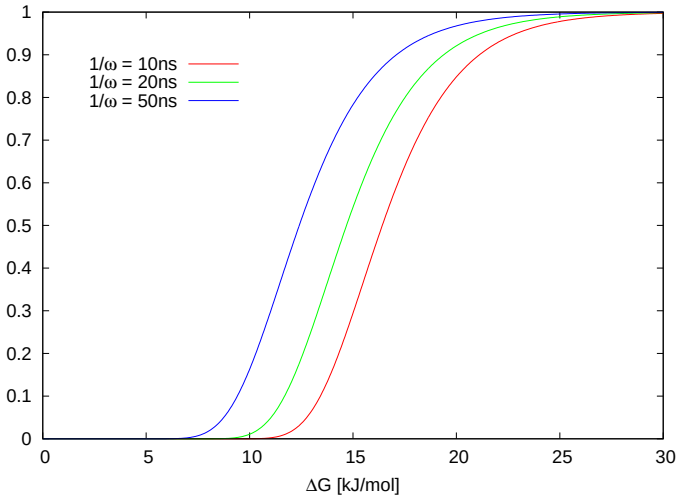


Figure 5.15: $\exp\left(-\omega T \cdot e^{-\beta \Delta G}\right)$ for different values of ω

from the two-dimensional PMF in fig. 5.10. Therefore a better suited reaction coordinate has to be searched.

5.3.2.1 Estimation of the Barrier Height

Although no conformational transition could be observed, from the extensive sampling of the open conformation and using a Poisson model an estimate for a lower limit of the barrier separating open and closed conformation, i. e. state A1 and B1 and A2 and B2 in fig. 5.1 is made (see section 4.2.4.4).

For zero transitions from the bound to the closed state the probability density for the barrier height is given by eq. (4.13) (see section 4.2.4.4):

$$\rho(\Delta G; n = 0) = \frac{\beta \cdot \exp(-\omega T e^{-\beta \Delta G})}{E_1(\omega T e^{-\beta c}) - E_1(\omega T)}. \tag{5.8}$$

An estimate for a lower barrier is given by

$$\int_0^{G_0} \mathrm{d}(\Delta G)\ \rho(\Delta G; n = 0) \stackrel{!}{=} \alpha. \tag{5.9}$$

with a small α, defining the error of the estimate. c is the upper limit in a uniform a priori probability distribution of ΔG. Both equations still depend on the arbitrary choice of c, therefore the estimate is chosen by the value for ΔG where $\exp\left(-\omega T \cdot e^{-\beta \Delta G}\right)$ starts to differ significantly from 0.

Figure 5.15 shows this expression for different values for ω. With an autocorrelation time of 10 ns the estimate $\Delta G > G_0 = 10\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$ is made.

The total simulation time of 5 μs contains all 50 trajectories and is not limited to those where binding occurs. Thus the estimate for ΔG is not valid for a specific

state with or without a bound ligand. Therefore this estimate can only be safely applied to the larger barrier. Furthermore it has to be noted that the system is more complex than a simple two state system. Since the determination of the autocorrelation time from autocorrelation functions that do not decay exponentially always bears some arbitrariness, the actual attempt frequency might differ from the assumed $\omega = 10^8/\mathrm{s}$. Nevertheless the estimate seems, as shown in fig. 5.15, also reasonable for smaller attempt frequencies.

5.3.2.2 Principal Component Analysis

Since the backbone difference vector does not provide the best separation of closed and open conformation a better reaction coordinate is searched that separates the A and B states in fig. 5.1. As explained in section 4.2.5 the subspace in which a better reaction coordinate is embedded is likely to be spanned by the eigenvectors of principal component analysis of the the trajectories of open conformation simulations.

For this reason and to allow visualisation of the sampled conformational space of protein in free binding simulations, a principal component analysis is performed on the the trajectories of the free binding simulations.

The principal component analysis on the collected data of all 50 trajectories yields a set of eigenvectors $\{\mathbf{e}_i\}$. The first normalised eigenvector $\mathbf{e}_1$ is similar to the backbone difference vector $\mathbf{v}$, with a scalar product of $\mathbf{v} \cdot \mathbf{e}_1 = 0.84$. The similarity is also visualised in a projection of the free binding and closed conformation simulation data on the first eigenvector vs. the backbone difference vector, see fig. 5.17.

Figure 5.16 shows isosurface plots of the projections of all trajectories from the free binding simulations as well as the trajectories of the closed conformation simulations on the first three eigenvectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$.

Both simulation sets occupy separated regions in the three-dimensional subspace spanned by the first three eigenvectors. Between the region occupied by the free binding simulations and the closed conformation simulations there exits a region where almost no simulation points can be found which thus constitutes a large energetic barrier.

Prior to using this separation for the derivation of a better coordinate in section 5.3.2.4 the open conformation sampling is tested for convergence.

5.3.2.3 Convergence of Protein Sampling

The method to derive an optimised reaction coordinate is explained in section 4.2.5. It makes use of the good sampling of closed and open conformation. Although full convergence of sampling of a substate is difficult to prove, we want to rule out an

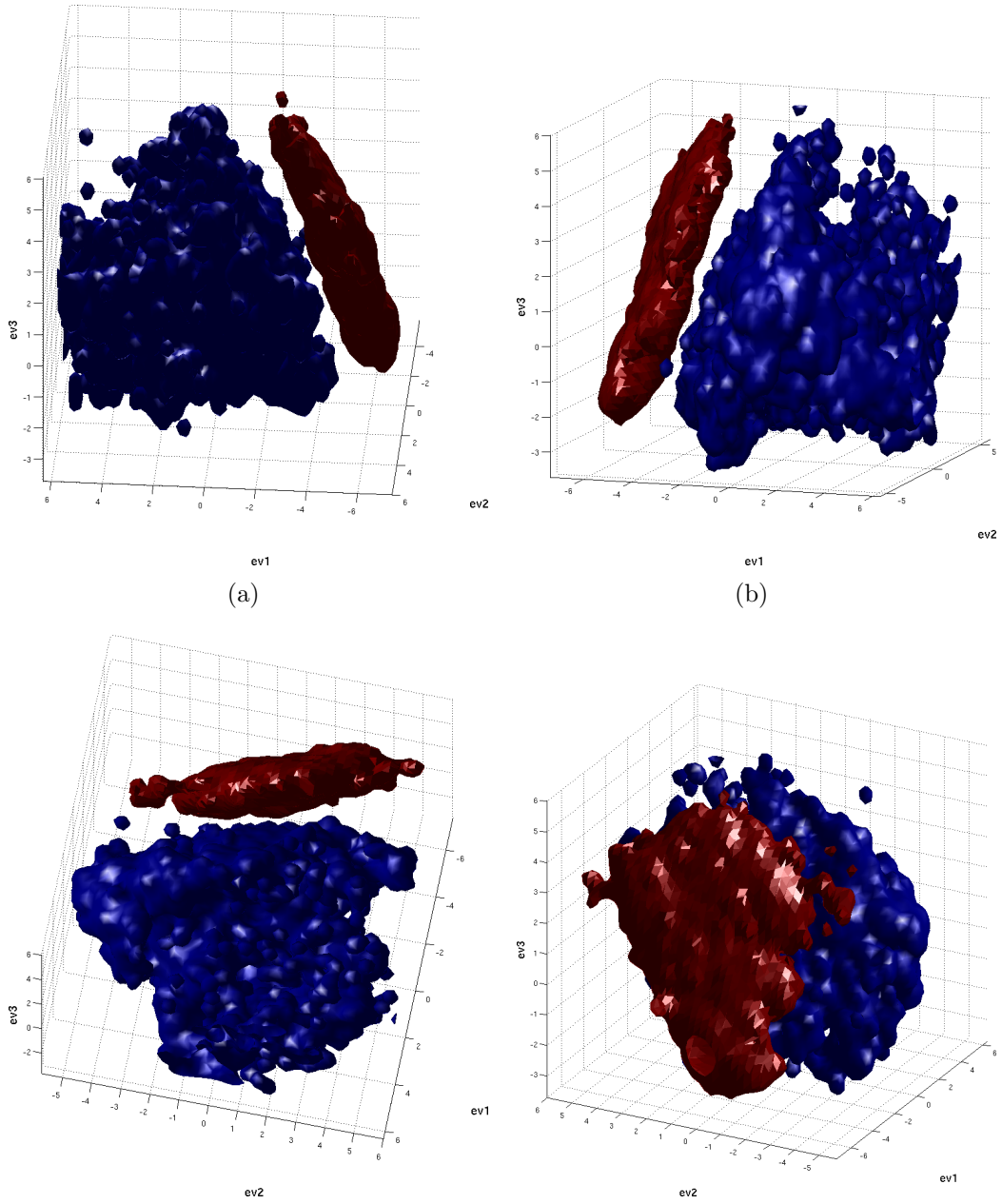


Figure 5.16: Open and closed conformation projected on the first three eigenvectors. Red: projections of closed conformation simulations; blue: projections of free binding simulations

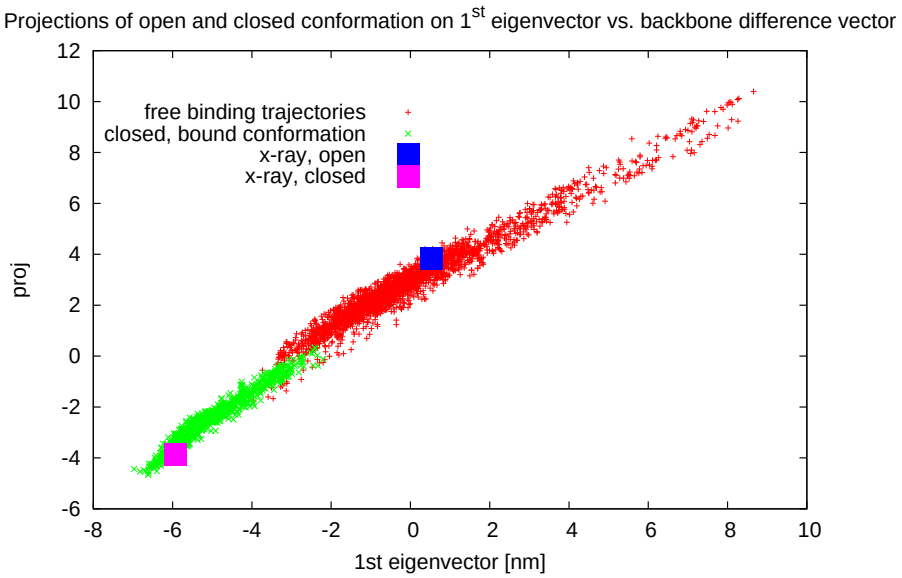


Figure 5.17: Projections on 1st eigenvector vs. projections on the backbone difference vector

obvious lack of sampling. The cosine content (Hess, 2000, 2002) method offers a way to check for sampling issues.

Therefore PCAs have also been performed on the individual free binding simulations. By determination of the cosine content of the projection of each trajectory on the first eigenvector of the corresponding PCA the similarity of the trajectory to random diffusion is obtained. A histogram of the obtained cosine contents along the first eigenvectors is shown in fig. 5.18. For a large number of free binding simulations the cosine content has values above 0.5, which means that convergence within the individual simulation runs is hardly reached. However, the projection of all collected trajectories on the first eigenvector from the PCA on the collected trajectories of all free binding simulations yields a cosine content of 0.012.

Although concatenated trajectories of random diffusion cannot necessarily be expected to be identical to one large diffusion trajectory, the results support the hypothesis that the collection of all 50 free binding simulation trajectories provide sufficient sampling of the open conformation of the CNBD.

5.3.2.4 Derivation of an Optimised Reaction Coordinate

Assuming that the sampling of the open conformation is sufficient, an optimised reaction coordinate is calculated that minimises the overlap the projections of open and closed conformation on the corresponding vector (see section 4.2.5).

By applying this method, new vectors in a d-dimensional subspace spanned by the eigenvectors of the PCA on the open conformation trajectories have been ob-

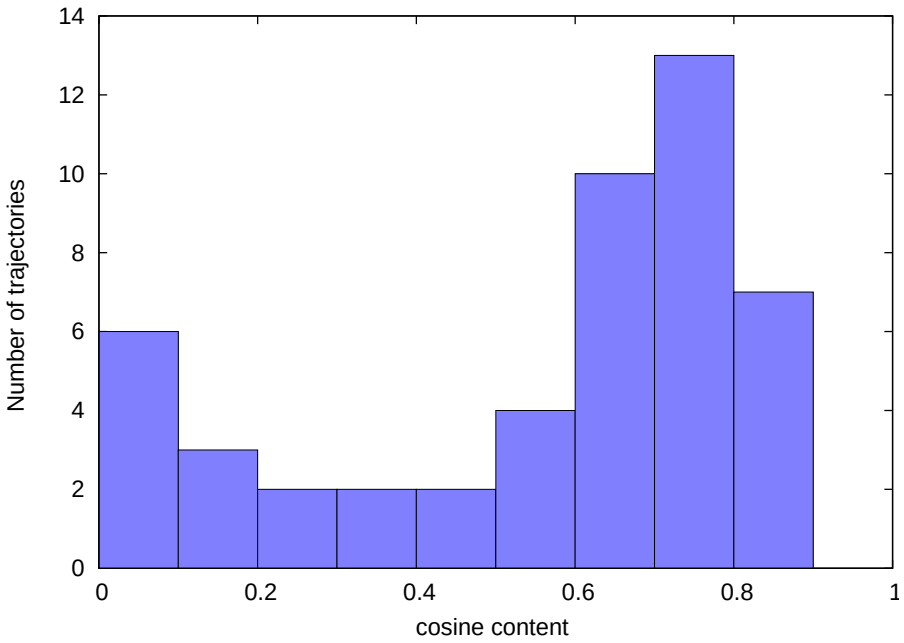


Figure 5.18: Histogram of the cosine content of the projections of the free binding trajectories on the first eigenvector of trajectory wise PCA

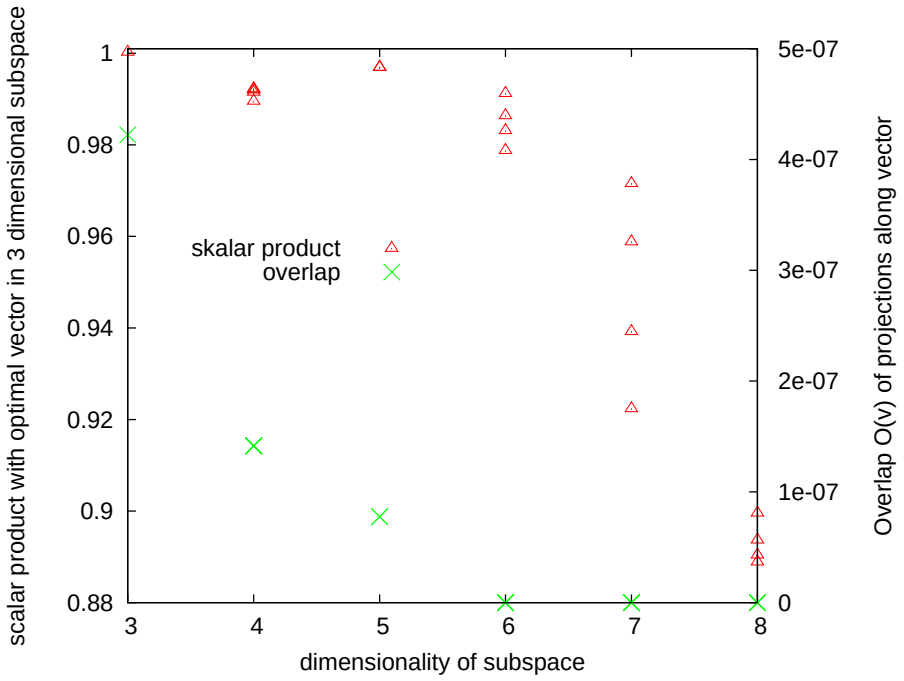


Figure 5.19: Overlap along optimised coordinates in different subspaces and scalar product with optimised coordinate in three-dimensional subspace

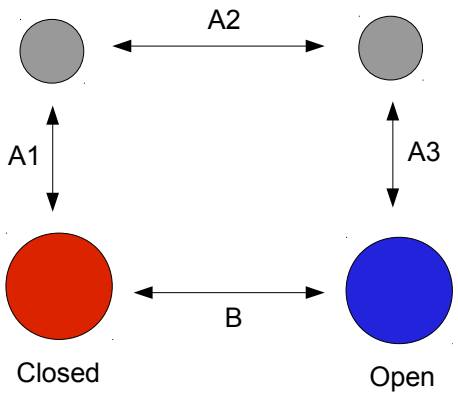


Figure 5.20: Hypothetical reaction pathways and energetic barriers

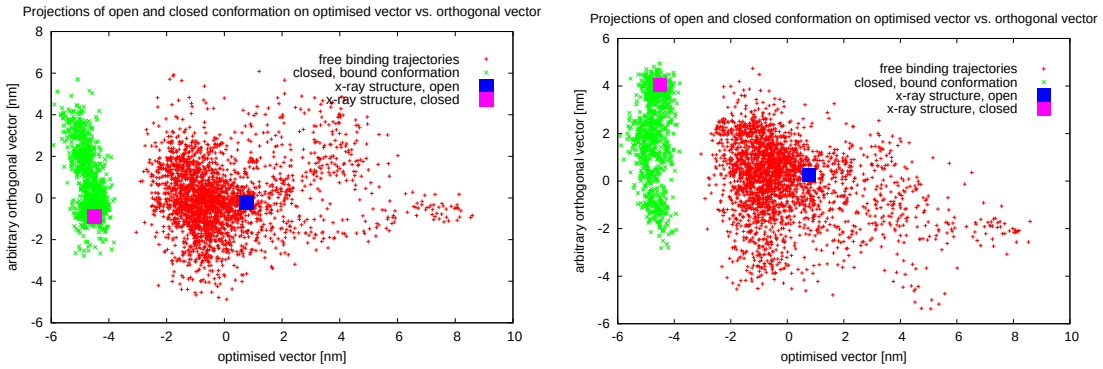
tained for $d \in [3, 8]$. The vector from a d-dimensional subspace that minimises the overlap function is denoted as $\mathbf{v}_d$, if multiple vectors produce the same numerical value we name them $\mathbf{v}_{d,i}$. For $d \geq 6$ vectors were found along which the projections of closed conformation and free binding simulations no longer produces any overlap. Since the overlap function has by construction a lower bound of 0, no further optimisation for those vectors is possible.

To compare the obtained vectors, the scalar products with the optimal vector found in the three-dimensional subspace, embedded in the d-dimensional space, are calculated. Both scalar products and the corresponding overlap function value are plotted in fig. 5.19.

The results show that the vectors obtained in 4, 5 and 6-dimensional subspace are very similar (scalar products $\mathbf{v}_3 \cdot \mathbf{v}_{d,i} > 0.97$) to the optimal vector found in three dimensions. Even in higher-dimensional subspaces the similarity remains large. The drop in the scalar product with higher-dimensional optimal vectors can be explained by the fact that in a high-dimensional space, the size of the high-dimensional cone containing all vectors along which the overlap is zero increases. The scalar product between the optimal vector in three dimensions with the backbone difference vector is $\mathbf{v}_3 \cdot \mathbf{v}_{bb} = 0.685$. $\mathbf{v}_3$ is given by $\mathbf{v}_3 = 0.862 \cdot \mathbf{e}_1 + 0.230 \cdot \mathbf{e}_2 - 0.452 \cdot \mathbf{e}_3$.

Due to the high similarity of the three-dimensional vector with the optimal six-dimensional vectors the optimal three-dimensional vector is used as an improved reaction coordinate, along which the highest energetic barrier is assumed to be found.

The assumption is valid as long as there are not any reaction pathways over multiple yet unidentified orthogonal substates which are separated by barriers with similar heights. This is illustrated in fig. 5.20: If there unidentified substates (illustrated in grey) separated by barriers A1 and A3 along coordinates orthogonal to the optimised reaction coordinate, then the actual barrier in a PMF along the opti-



(a) $\mathbf{v}_3$ vs. $\mathbf{v}_{o,1}$ (b) $\mathbf{v}_3$ vs. $\mathbf{v}_{o,2}$

Figure 5.21: Open and closed conformation projected on optimal vector $\mathbf{v}_3$ vs. $\mathbf{v}_{o,1}$ and $\mathbf{v}_{o,2}$

mised reaction coordinate might be very low due to the possible existence of a "low energy pathway" along A2. A fully converged PMF along the optimised reaction coordinate would then not show a high barrier, because all direct "high-energy-pathways" (exemplary illustrated by arrow B) do not contribute significantly to this PMF.

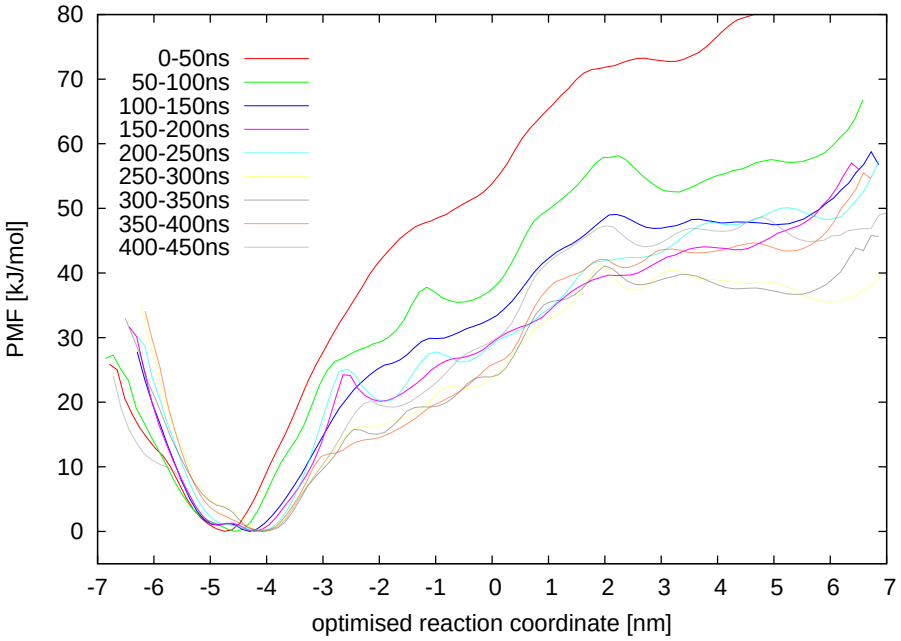
Projections of closed and open conformation simulation data on the optimised vector are plotted against projections on two (arbitrarily chosen) orthogonal vectors $\mathbf{v}_{o,1}$ and $\mathbf{v}_{o,2}$ (these vectors fulfil $\mathbf{v}_3 \perp \mathbf{v}_{o,1} \perp \mathbf{v}_{o,2}$) are plotted in fig. 5.21.

With this reaction coordinate, we want to improve the estimates for the barrier separating state A1 and B1 and A2 and B2 respectively in figure 5.1. Furthermore, by using sampling improving techniques (see section 4.2.6) along this coordinate, the estimates for the free energy differences calculated in section 5.2 between the substates should be improved.

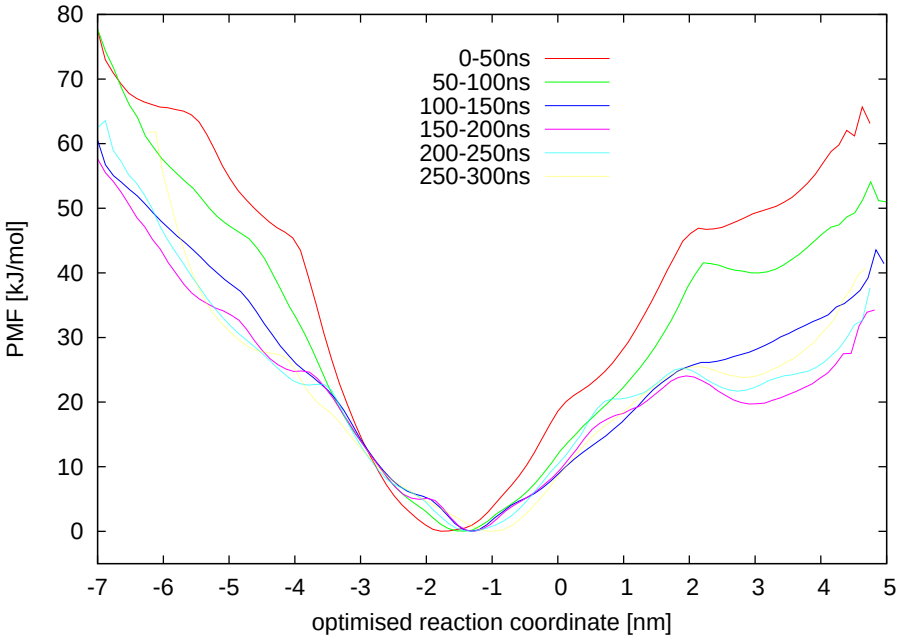
5.4 Umbrella Sampling Along Optimised Reaction Coordinate

The optimised reaction coordinate has been used for umbrella sampling simulations. The construction of the starting configurations are described in section 4.2.6.

PMFs along the optimised reaction coordinate over consecutive time windows for the simulation set for which the starting configuration have been derived using essential dynamics starting at the closed conformation are plotted in fig. 5.22(a). Figure 5.22(b) contains the analog PMFs for the set of simulations for which the starting structures were obtained by using essential dynamics starting from an open conformation.

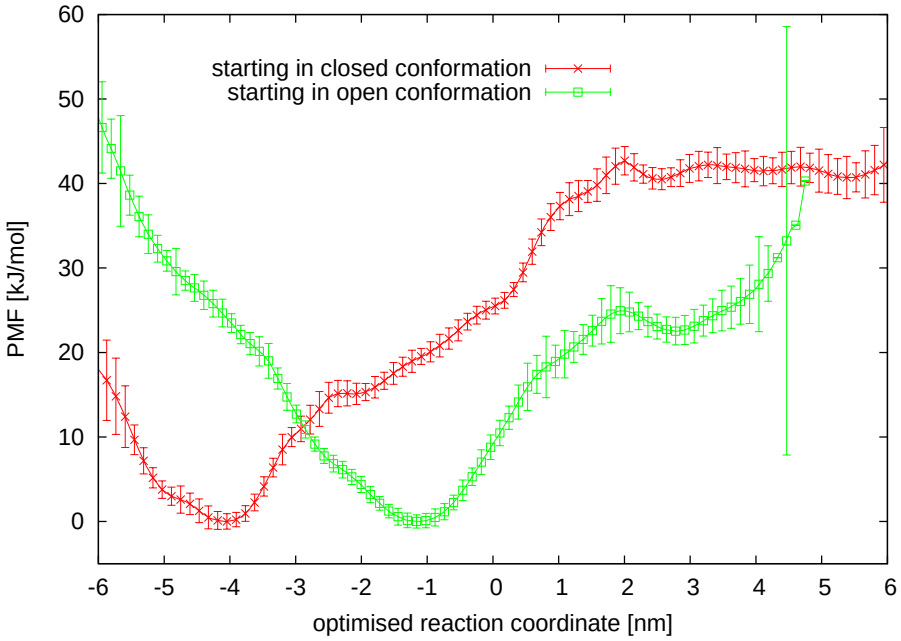


(a) using starting configurations obtained using the closed conformation

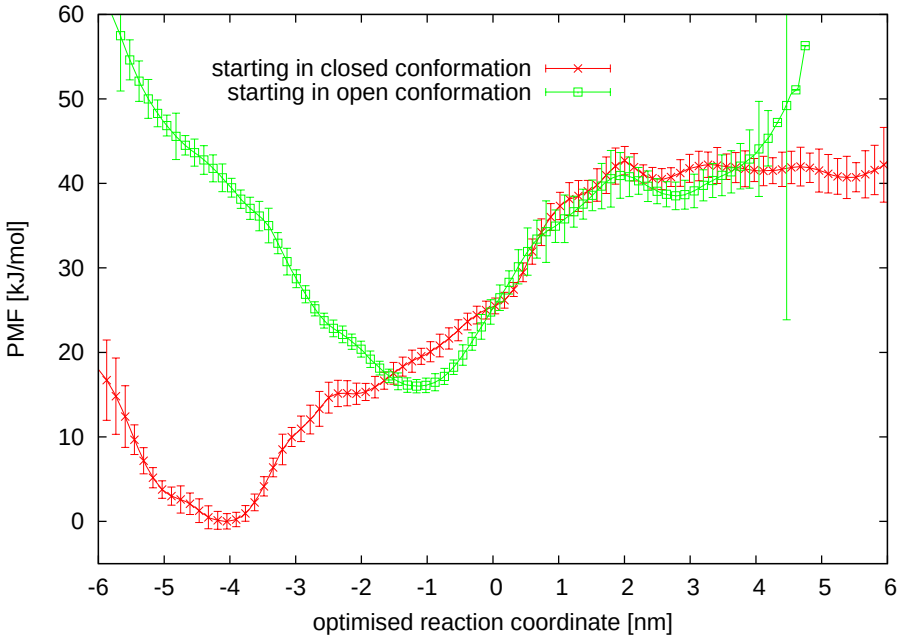


(b) using starting configurations obtained using the open conformation

Figure 5.22: Consecutive PMFs along optimised reaction coordinate



(a)



(b)

Figure 5.23: (Unequilibrated) PMFs along optimised reaction coordinate

Figure 5.23 shows PMFs for both simulation sets. Since both PMFs were calculated independently, their relative vertical offset is not fixed. In fig. 5.23(a) the minimum in each PMF is set to zero, in fig. 5.23(b) both PMFs are overlayed in such a way that the right part overlap.

The PMFs in fig. 5.22 show a large trend over a large range of the reaction coordinate from high energy values for early time windows to lower energy values for later time windows. Since both PMFs are for the same reaction coordinate they should be identical if the simulations were equilibrated and the PMFs converged. This equilibrium has not been reached in at least one simulation set, much more likely however both PMFs are far from convergence. The minima of both PMFs are at the position of the structure that was used as the starting point for the preparatory essential dynamics runs. This suggest that the starting structures obtained by the essential dynamics procedure are very far from the equilibrium of their respective umbrella window. The error bars do not show this systematic error since they only capture the stochastic errors in finite but equilibrated umbrella window simulations.

The unbiased simulations of open and closed conformation show that the minima at $x = -4\,\mathrm{nm}$ and $x = -1.5\,\mathrm{nm}$ do exist. We assume that the inaccuracies increase for umbrella windows whose starting structure has a longer distance along the optimised reaction coordinate from the starting structure of the preparatory essential dynamics simulations, which in turn means that the PMF is more accurate close to the minima. With this assumption and the knowledge that both minima have to exist we estimate for the upper limit of the barrier along the optimised reaction coordinate between the two minima $\Delta G < 20\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$.

Although we are confident that the optimised reaction coordinate does separate open and closed conformation in the subspace spanned by the the degrees of freedom of the backbone atoms well enough, the possibility remains that there is at least one additional energetic barrier in the motion of the side chains that has to be overcome during a transition from the closed to the bound conformation. The motion of the side chains takes by definition place in the space orthogonal to the reaction coordinate, sampling along this coordinate is thus not improved by the umbrella sampling simulations.

6 Conclusion & Outlook

> Diese Arbeit wurde mit Hilfe von Computersimulationen erstellt und ist ohne Experimente gültig.
>
> *(Béla Voß)*

In this work we analysed free energy differences and free energy barriers for the transitions between the substates of the binding process sketched in fig. 1.2. The intention was to find out whether the binding process of cAMP at the CNBD of MloK1 is better described by an induced fit or by a conformational selection model.

First we studied ligand binding, that is transitions from state A1 to state A2 in fig. 1.2. We intended to find out whether spontaneous ligand binding occurs in unbiased simulations starting in the unbound state (state A1 in fig. 1.2). We observed successful ligand bindings in multiple 100 ns free binding simulations. This shows that the bound state must be related to a free energy minimum and that the binding is not hindered by a large energetic barrier which again means the related on-rate is sufficiently large.

To quantify these findings we wanted to calculate a PMF for the binding process along a suited reaction coordinate. Comparison of the ligand RMSD to the bound state during a binding trajectory with the distance of the ligand COM to the binding site showed that the distance of the ligand COM to the COM of the binding site-surrounding residues does provide a suitable reaction coordinate.

A PMF along this coordinate showed that the bound state, state A2 in fig. 1.2, corresponds to a minimum in the PMF. The PMF furthermore indicates that the ligand entering the vicinity of the protein and the binding site itself introduces a free energy barrier probably due to the confined accessible space of roughly $6\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$. The ligand protein interactions then create a minimum in the bound state. This is in good agreement with the fact that free ligand binding events were observed.

Next we addressed the question whether conformational changes from the open to the closed conformation (a transition from A2 to B2 or A1 to B1 in fig. 1.2) happen within the free binding simulations. We did not observe such transitions during the unbiased simulations. This shows that open and closed conformation are separated by an energetic barrier. Together with Kramers barrier crossing model a lower estimate of $10\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$ for the barrier that separates open and closed conformation was obtained.

In a following step, we wanted to quantify this barrier for systems with and without a bound ligand, meaning for the transitions from A1 to B1 and A2 to B2 in fig. 1.2 and we wanted to calculate the free energy differences between these states. The means for this were the calculation of PMFs along a suitable reaction coordinate describing the transition from the open to the closed conformation.

To find such a suiting reaction coordinate, we first tested whether the backbone difference vector connecting the two X-ray structures constitutes an applicable reaction coordinate for the conformational change. Subsequently we calculated PMFs along this coordinate for systems with and without a bound ligand. In the obtained PMFs we observed minima for the open and closed conformation, but only a small separating energetic barrier that was not compatible with our lower estimate for the barrier. To test whether the low barrier was caused by an overlap of the projections of open and closed conformation on the backbone difference vector, we displayed the underlying simulation data using more than one dimension. A two-dimensional PMFs along the backbone difference vector and an orthogonal coordinate showed that the open and closed conformation occupy separate regions in the configurational space, but that their projections on the backbone difference vector partly overlap. This means that the highest energetic barrier in a one-dimensional PMF is not found along the backbone difference vector. This result was supported by extensive sampling of the open and closed conformation, which shows that both conformations occupy regions in configurational space that are not well separated by the backbone difference vector.

In a next step we searched for a vector for which the overlap of the projected distribution densities of open and closed conformation is minimal. By minimising the overlap we thus obtained an optimised reaction coordinate along which the maximal energetic barrier is expected. The obtained reaction coordinate is likely to be a suited reaction coordinate for the conformational change in the protein, as long as the assumption holds that the side chain motions have a negligible influence on the free energy differences between the two conformations. Future analysis of the side chain motion will be needed to verify or falsify this assumption.

PMFs derived from umbrella sampling simulations along the optimised reaction coordinate gave an upper limit for the energetic barrier between open and closed conformation of $20\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$. Together with the lower boundary, this limits the height of the energetic barrier between the two configurations to an interval $10\,{}^{\mathrm{kJ}}/_{\mathrm{mol}} \leq \Delta G \leq 20\,{}^{\mathrm{kJ}}/_{\mathrm{mol}}$.

Apart from free energy barriers, the PMFs allowed an estimation of the free energy differences between the open and closed conformation, i. e. between state A1 and B1 and between A2 and B2. This was done using the PMF along the backbone difference vector. The obtained values show that the closed conformation is preferred to the open conformation while a ligand is bound relative to the case where no ligand is bound. However, the exact numbers for the free energy differences between open and closed conformation obtained from different simulation sets differ

strongly. This was not due to the overlap of closed and open conformation, which we had already observed. Instead it was due to underlying sampling problems in the umbrella sampling simulations used for the calculation of the PMFs. Fully converged PMFs along the optimised reaction coordinate for systems with and without a bound ligand which are not available yet will most likely provide better estimates for the free energy differences.

Using the obtained PMFs and the free energy profiles, we finally want to give an educated guess concerning the question whether the reaction is better described by conformational selection or induced fit. The on-rate of the ligand-protein binding reaction is dependent on the concentration of the ligand. Any conformational change in the protein on the other hand only depends on the free energy profile. Even without any exact knowledge of the separating barrier, the PMFs along the backbone difference vector show that in the case of an unbound ligand, both conformations have comparable free energies. This means that - given a sufficiently low concentration of cAMP - the average transition time of the conformational change will always be lower than the time between an unbinding and a binding event. In these low concentration cases the binding process will always be conformational selection-like. However, since the transition time is at least in the microsecond regime, for higher concentrations of cAMP, these transitions will not occur between two binding events but only while a ligand is bound. In this case the kinetics of the binding process including the conformational change will not differ from the induced fit model.

Possible ligand binding events while the protein is already in the closed conformation, i. e. transitions between state B1 and B2 in fig. 1.2 and the associated free energy landscape have not been analysed in this work. The reaction coordinate is, however, of high interest in the light of the conformational selection model, because such binding events would not occur in an induced fit model. From the structure, depicted in fig. 3.3(c) it becomes obvious that the binding might be hindered in the closed conformation and we except a larger free energy barrier in a PMF along the ligand distance reaction coordinate than in the PMF for the ligand binding in the open conformation (section 5.1). Preliminary simulations not included in this work starting in the bound closed state (B2 in fig. 1.2) where a pulling force was applied to the ligand indicate that a conformational change in the protein is likely, but not necessarily connected with an enforced unbinding of the ligand. Further studies and free energy calculations for a transition from state B1 to B2 in fig. 1.2 or reverse will hopefully provide insight whether this conformational selection-like pathway is accessible.

It has to be noted that the computational effort for the free energy calculations is very large. This means that binding processes associated with conformational changes are unlikely to be studied for a large number of systems, especially if those are of a bigger size in terms of atom numbers. Future work which may focus on the working mechanism of an entire ion channel should therefore make use of the

results of free energy calculations for a subsystem like the CNBD as a basis for the study of opening and closing events of the entire ion channel.

Appendix

A.1 Estimation of the Barrier Height

In this section the full derivation of the probability distribution of the barrier height, $\rho(\Delta G, n)$, is written down. n denotes the number of transition, ΔG the barrier height, ω the attempt frequency in a transition rate model and T the simulation time. In the following we will use for shorter notation $g := \Delta G$ and $\alpha := \omega T$.

The primitive of

$$f(g) = \frac{\exp(-\alpha \cdot e^{-\beta g})}{c} \tag{A.1}$$

is given by

$$F(g) = \int_1^\infty \frac{\exp(-t\alpha \cdot e^{-\beta g})}{tc\beta} \mathrm{d}t \tag{A.2}$$

$$=: \frac{E_1\left(\alpha \cdot e^{-\beta g}\right)}{\beta c} \tag{A.3}$$

with the exponential integral E_1 because

$$\frac{\mathrm{d}F(g)}{\mathrm{d}g} = \int_1^\infty \frac{\exp(-t\alpha \cdot e^{-\beta g}) \cdot (-t\alpha) \cdot e^{-\beta g}(-\beta)}{tc\beta} \mathrm{d}t \tag{A.4}$$

$$= -\frac{1}{c}\left[\exp\left(-t\alpha \cdot e^{-\beta g}\right)\right]_{t=1}^{\infty} \tag{A.5}$$

$$= \frac{1}{c}\exp\left(-\alpha \cdot e^{-\beta g}\right). \tag{A.6}$$

Since

$$1 \stackrel{!}{=} \int \mathrm{d}g\ \rho(g, n) = \int \mathrm{d}g\ \frac{p(n, g) \cdot \rho(g)}{p(n)} \tag{A.7}$$

we get for $p(x)$:

$$p(x) = \int \mathrm{d}g\ p(n, g) \cdot \rho(g). \tag{A.8}$$

With

$$p(n;g) = \frac{(\alpha e^{-\beta g})^n}{n!} \cdot \exp(-\alpha e^{-\beta g}) \tag{A.9}$$

$$\rho(g) = \begin{cases} \frac{1}{c} & 0 < g < c \ [c] = \frac{kJ}{mol} \\ 0 & else \end{cases} \tag{A.10}$$

this becomes

$$p(x) = \int_{-\infty}^{\infty} \mathrm{d}g \ p(n;g) \cdot \rho(g) \tag{A.11}$$

$$= \int_0^c \mathrm{d}g \ \frac{\rho(n,g)}{c}. \tag{A.12}$$

The cases $n = 0$ and $n > 0$ are handled separately. For sufficiently large c and α, i. e. if

$$\alpha \cdot \exp(-\beta c) \ll \exp(-\alpha e^{-\beta c}), \tag{A.13}$$

$p(n)$ becomes

$$p(n) = \frac{1}{\beta c n} \ \forall \, n \geq 1. \tag{A.14}$$

For $n = 0$ we get

$$p(n) = \frac{E_1(\alpha e^{-\beta c}) - E_1(\alpha)}{\beta c}. \tag{A.15}$$

Thus the overall result is

$$\rho(g;n) = \begin{cases} \frac{\beta \cdot \exp(-\alpha e^{-\beta g})}{E_1(\alpha e^{-\beta c}) - E_1(\alpha)} & \text{for } n = 0, g < c \\ \beta n \cdot \exp(-\alpha e^{-\beta g}) \frac{(\alpha e^{-\beta g})^n}{n!} & \text{for } n \geq 1, g < c \\ 0 & g > c. \end{cases} \tag{A.16}$$

A.2 Parameters for cAMP

Figure A.1 shows a two-dimensional sketch of unprotonated cAMP. The green numbers indicate atom numbers, the corresponding charges and Lennard-Jones parameters which were derived from the GAFF/antechamber are listed in table A.1.

Figure A.2 shows the 2D-structre of protonated cAMP, the force field values can be found in table A.2

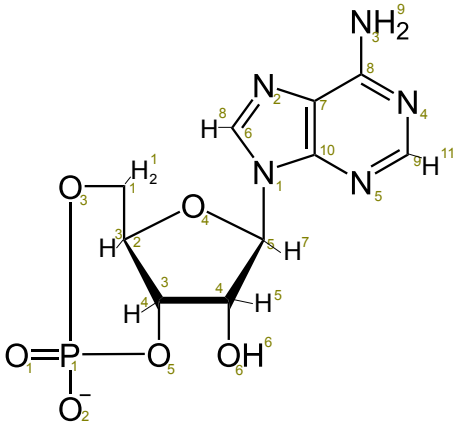


Figure A.1: 2D-structure of unprotonated cyclic adeonise monophosphate

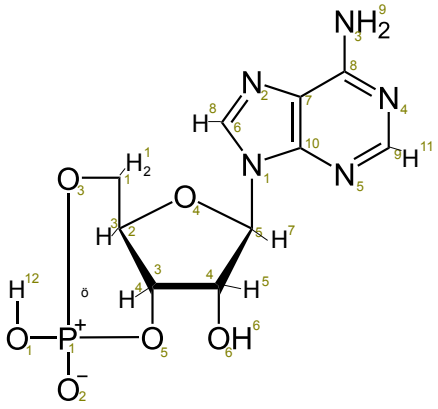


Figure A.2: 2D-structure of protonated cyclic adeonise monophosphate

Atom	Charge [e]	σ (10^{-1}nm)	ϵ (kJ/mol)
P_1	1.28090	3.74177	8.36800e-01
O_1/O_2	-0.77690	2.95992	8.78640e-01
O_3	-0.53100	3.00001	7.11280e-01
O_4	-0.52930	3.00001	7.11280e-01
O_5	-0.51800	3.00001	7.11280e-01
O_6	-0.73990	3.06647	8.80314e-01
C_1	0.13820	3.39967	4.57730e-01
C_2	0.17570	3.39967	4.57730e-01
C_3	0.07460	3.39967	4.57730e-01
C_4	0.33990	3.39967	4.57730e-01
C_5	0.26780	3.39967	4.57730e-01
C_6	0.18890	3.39967	3.59824e-01
C_7	-0.03900	3.39967	3.59824e-01
C_8	0.75250	3.39967	3.59824e-01
C_9	0.62020	3.39967	3.59824e-01
C_{10}	0.50810	3.39967	3.59824e-01
N_1	-0.15770	3.25000	7.11280e-01
N_2	-0.59380	3.25000	7.11280e-01
N_3	-0.91890	3.25000	7.11280e-01
N_4	-0.81560	3.25000	7.11280e-01
N_5	-0.78400	3.25000	7.11280e-01
H_1/H_2	0.05170	2.47135	6.56888e-02
H_3	0.07660	2.47135	6.56888e-02
H_4	0.04880	2.47135	6.56888e-02
H_5	0.00530	2.47135	6.56888e-02
H_6	0.45920	0	0
H_7	0.12300	2.29317	6.56888e-02
H_8	0.17870	2.42146	6.27600e-02
H_9/H_{10}	0.39720	1.06908	6.56888e-02
H_{11}	0.04480	2.42146	6.27600e-02

Table A.1: Charges and Lennard-Jones parameters for cAMP.

Atom	Charge [e]	σ [10^{-1}nm]	ϵ [kJ/mol]
P_1	1.21250	3.74177	8.36800e-01
O_1	-0.66290	2.95992	8.78640e-01
O_2	-0.69220	2.95992	8.78640e-01
O_3	-0.45780	3.00001	7.11280e-01
O_4	-0.56510	3.00001	7.11280e-01
O_5	-0.46300	3.00001	7.11280e-01
O_6	-0.69180	3.06647	8.80314e-01
C_1	0.04940	3.39967	4.57730e-01
C_2	0.33800	3.39967	4.57730e-01
C_3	0.17810	3.39967	4.57730e-01
C_4	0.05750	3.39967	4.57730e-01
C_5	0.35340	3.39967	4.57730e-01
C_6	0.32390	3.39967	3.59824e-01
C_7	-0.00590	3.39967	3.59824e-01
C_8	0.81540	3.39967	3.59824e-01
C_9	0.63380	3.39967	3.59824e-01
C_{10}	0.44070	3.39967	3.59824e-01
N_1	-0.26720	3.25000	7.11280e-01
N_2	0.62810	3.25000	7.11280e-01
N_3	-0.99130	3.25000	7.11280e-01
N_4	-0.81900	3.25000	7.11280e-01
N_5	-0.76890	3.25000	7.11280e-01
H_1/H_2	0.05170	2.47135	6.56888e-02
H_3	0.07660	2.47135	6.56888e-02
H_4	0.04880	2.47135	6.56888e-02
H_5	0.00530	2.47135	6.56888e-02
H_6	0.45920	0	0
H_7	0.12300	2.29317	6.56888e-02
H_8	0.17870	2.42146	6.27600e-02
H_9/H_{10}	0.39720	1.06908	6.56888e-02
H_{11}	0.04480	2.42146	6.27600e-02
H_{12}	0.49390	0	0

Table A.2: Charges and Lennard-Jones parameters for protonated cAMP.

Bibliography

B. Alberts. *Molecular biology of the cell.* Garland Science, 2002. 1

A. Amadei, A. B. Linssen, B. L. de Groot, D. M. van Aalten, and H. J. Berendsen. An efficient method for sampling the essential subspace of proteins. *Journal of biomolecular structure & dynamics*, 13(4):615–625, February 1996. 4.2.6

C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97(40):10269–10280, October 1993. 4.1

H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984. 2.1.3.2, 4.2

M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 389(20):457–484, 1927. 2.1.1.1

G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007. 2.1.3.2, 4.2

P.-L. Chiu, M. D. Pagel, J. Evans, H.-T. Chou, X. Zeng, B. Gipson, H. Stahlberg, and C. M. Nimigean. The structure of the prokaryotic cyclic nucleotide-modulated potassium channel mlok1 at 16 Å resolution. *Structure*, 15(9):1053–1064, 2007. 3.1, 3.2, 3.2

G. M. Clayton, W. R. Silverman, L. Heginbotham, and J. H. Morais-Cabral. Structural basis of ligand activation in a cyclic nucleotide regulated potassium channel. *Cell*, 119(5):615–627, 2004. 1, 3.2, 3.2, 4.2.1

A. Cukkemane, B. Gruter, K. Novak, T. Gensch, W. Bonigk, T. Gerharz, B. U. Kaupp, and R. Seifert. Subunits act independently in a cyclic nucleotide-activated k+ channel. *EMBO Reports*, aop(current), July 2007. 3.2

T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An $n \cdot log(n)$ method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993. 2.1.2.2, 4.2

Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, December 2003. 4.2

D. J. Earl and M. W. Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005. 2.4

P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921. 2.1.2.2

D. Frenkel and B. Smit. *Understanding Molecular Simulation, Second Edition: From Algorithms to Applications (Computational Science Series, Vol 1)*. Academic Press, 2nd edition, November 2001. 2.1.3.1, 2.2

M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004. 4.1

R. J. Gdanitz. Methoden der Quantenchemie, Jul 1999. 2

C. J. Geyer. Markov chain monte carlo maximum likelihood. *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pages 156–163, 1991. 2.4

R. E. Gillilan and K. R. Wilson. Shadowing, rare events, and rubber bands. a variational verlet algorithm for molecular dynamics. *The Journal of Chemical Physics*, 97(3):1757–1772, 1992. 2.1.3.1

M. Griebel, S. Knapek, and G. Zumbusch. *Numerical Simulation in Molecular Dynamics: Numerics, Algorithms, Parallelization, Applications (Texts in Computational Science and Engineering)*. Springer, September 2007. 2.1.1.3

T. A. Halgren and W. Damm. Polarizable force fields. *Current Opinion in Structural Biology*, 11(2):236–242, 2001. 2.1.2.3

B. Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62(6):8438–8448, Dec 2000. 5.3.2.3

B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65(3): 031910, Mar 2002. 5.3.2.3

B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997. 4.2

B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3):435–447, March 2008. 4.2

W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, Mar 1985. 2.1.3.2

V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, November 2006. 4.2

C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2693, Apr 1997. 2.2.1

W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983. 4.2

J. G. Kirkwood. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313, 1935. 2.2.1, 2.2.1

E. Krieger, J. van Meel, G. Vriend, and chris Spronk. Yasara. 4.2.1

S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992. 2.2.1

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. 2.4

J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965. 4.2.5

C. M. Nimigean and M. D. Pagel. Ligand binding and activation in a prokaryotic cyclic nucleotide-modulated channel. *Journal of Molecular Biology*, 371(5):1325–1337, 2007. 3.2

C. M. Nimigean, T. Shane, and C. Miller. A Cyclic Nucleotide Modulated Prokaryotic K+ Channel. *J. Gen. Physiol.*, 124(3):203–210, 2004. 3.2

S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 100:191–198(8), Jan 2002. 2.1.3.2

M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 1981. 2.1.3.2

C. Predescu, M. Predescu, and C. V. Ciobanu. On the efficiency of exchange in parallel tempering monte carlo simulations. *J. Phys. Chem. B*, 109(9):4189–4196, March 2005. 5

B. Roux. The calculation of the potential of mean force using computer simulations. *Computer Physics Communications*, 91(1-3):275–282, 1995. 2.2.1, 2.2.1

U. Scherz. *Quantenmechanik: Eine Einführung mit Anwengun auf Atome, Moleküle und Festkörper*. Teubner Verlag, 1999. 2.1.1.3

R. H. Swendsen and J.-S. Wang. Replica monte carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57(21):2607–2609, Nov 1986. 2.4

G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977. 2.2

D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, December 2005. 4.2

L. Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Phys. Rev.*, 159(1):98, Jul 1967. 2.1.3.1, 4.2

J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, July 2004. 4.1

J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247 – 260, 2006. 4.1

M. G. Wolf, J. A. Jongejan, J. D. Laman, and S. W. de Leeuw. Rapid free energy calculation of peptide self-assembly by remd umbrella sampling. *J. Phys. Chem. B*, 112(43):13493–13498, October 2008. 2.4.1

Acknowledgement

"Gratitude is best and most effective when it does not evaporate itself in empty phrases."

(Isaac Asimov)

I would like to thank Helmut Grubmüller for supervising the research behind this thesis, introducing me to many interesting topics and questions, providing a tremendous amount of valuable feedback and for having many illuminating discussions with me. I would like to thank all members of the Department of Theoretical and Computational Biophysics for very helpful discussions and providing answers to all kind of scientific and technical questions. I would like to thank the Max Planck Institute for Biophysical Chemistry for funding and providing me with all necessary resources to do the research behind this thesis, especially the computational resources.

I would also like to thank Marcus Müller from the Institute for Theoretical Physics at the Georg-August-University Göttingen for agreeing to become my official supervisor at the Faculty of Physics and thus enabling me to write an external diploma thesis.