Master's Thesis

Allosterie

Allosteric Transitions

prepared by

Julian Tim Brennecke

from Hildesheim

at the Max-Planck-Institute for Biophysical Chemistry

Thesis period:	1st October 2013 until 30th September 2014
First Rreferee:	Prof. Dr. Bert de Groot
Second referee:	Prof. Dr. Helmut Grubmüller

Abstract

More than 50 years after the initial model of cooperative binding was proposed by Monod, Wyman and Changeux a general theoretical approach to allostery remains controversial. The understanding of allostery as a biological phenomenon of fundamental importance would be a crucial step for the manipulation of such proteins for example by drugs. In this study we focused on Hemoglobin (Hb), an oxygen-binding protein composed of four subunits, the prime example of an allosteric protein. We used the previously described method of tertiary - quaternary disentanglement to propose a new model of the sequence of events (i. e. simultaneous quaternary - tertiary transition) in Hb. Furthermore we employ Essential Dynamics to find a causal coupling of the quaternary motion to the tertiary motion while other couplings could not be found. In addition a Functional Mode Analysis (FMA) was used to identify sidechain correlation to the quaternary transition. This revealed important residues at the α_1/β_2 and α_2/β_1 interfaces as well as a single residue in the α_1 subunit previously found by others. In the last part of our work we applied (FMA) to develop a systematic approach to find correlations between the structural transitions and the breaking and formation of contacts. This approach was applied to the intersubunit contacts to identify 25 majorly correlated contacts that dominate in the global allosteric transition.

Contents

Nomenclature

Abbreviations

Abbreviation	Meaning
e. g.	exempli gratia
FMA	Functional Mode Analysis
Hb	Hemoglobin
i. e.	in explicit
PCA	Principal Component Analysis
PDB	Protein Data Bank
PLS	Functional Mode Analysis based on Partial Least Squares
QM	Quantum Mechanics
Qv	Quaternary transition vector, first eigenvector of PCA analysis on quaternary motion
R	Quaternary R state
r	Tertiary R state
T	Quaternary T state
t	Tertiary T state
Tv	Tertiary transition vector, first eigenvector of PCA analysis on PLS motion.

1. Introduction

The thesis is a study of allostery in Hemoglobin. A theoretical introduction of allostery as well as to Hemoglobin is given in chapter 2. Here, also the technique of Molecular Dynamics simulation is introduced. This is the method providing us with trajectories to analyze allosteric transitions and underlying structural mechanisms.
In chapter 3 the methods used in this study are reviewed briefly. These are the Principal Component Analysis, used for dimensionality reduction and the Functional Mode Analysis, to find collective vectors in datasets that correlate most to a given functional property. Furthermore, a method to separate tertiary and quaternary motions in the trajectory of a multimeric protein is reviewed. At the end of this chapter the previous work done in our department, that is important for this study, is presented.
Chapter 4 presents the progress of the last year of work. This starts with a study on coupling between internal motion of different subunits to their relative motion and of the time-sequence of events taking place in the conformational change of Hemoglobin. Here the influence of the subunits onto each other as well as on their relative motion is studied beyond correlation by manipulating one and recording the effect on the other. Furthermore, an analysis of the sidechain rearrangements during the conformational change is performed to find communication networks in Hemoglobin. In the last part of this chapter a new application of Functional Mode Analysis is described. Here the contacts at the subunit interface are taken as input data to describe the transition.
Conclusions drawn from the results found are summarized in chapter 5. As a final remark a possible further progression of the project is suggested in the outlook section.

2. Theoretical Background

In this chapter an introduction to allostery and the ways how to study this concept are reviewed. Subsequently, the exemplary allosteric system of Hemoglobin is described, followed by a more extended introduction to Molecular Dynamics simulations, the main technique used in this study.

2.1. Allostery

In the following section I will give a brief review on allostery in general making clear why it is of importance. Subsequently, I will present the most popular models used to describe allostery and current experimental as well as computational methods to study it.

Allostery is the regulation of an active site of a protein by the binding of an effector molecule at a distant binding site. This is in contrast to the orthostery which is regulation by binding to the same site. Allostery must not only be positively coupled to the binding of the ligand but can also be negatively coupled. The allosteric regulation is a fundamental property of proteins and a malfunction is often related to diseases[1–3]. Therefore, it is of great importance to study and understand the fundamental mechanisms at play to be able to predict regulatory effects. Current concepts of allostery remain unquantifiable and are not able to predict allostery on the atomistic level[4, 5]. But there are two main models known to describe allosteric transitions well on a coarse level. These are the MWC (Monod-Wyman-Changeux) or conformational selection model[6] and the KNF (Koshland-Nemethy-Filmer) or induced fit model[7, 8]. Both will be described in greater detail in the following. Allostery is nowadays often thought of as consisting of two states with one being active (often relaxed or R) and the other inactive (tense or T) and the possibility to undergo a transition between the two of them.

2.1.1. Induced Fit and Conformational Selection Model

The early models of Monod, Wyman and Changeux (MWC, also called conformational selection) and Koshland, Nemethy and Filmer (KNF, also called induced fit) are still the basis of the understanding of allostery[9]. Both of these models provide a phenomenological description of the sequence of events an allosteric protein undergoes in its transition from the T to the R state. A schematic representation of the two models is presented in figure 2.1.

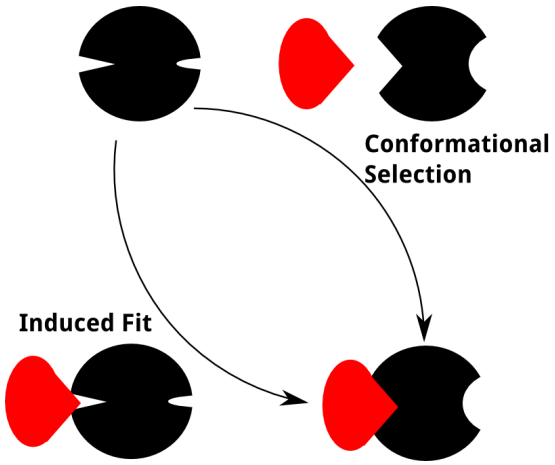


Figure 2.1.: Pathways of allosteric transitions Two major models for allostery are the induced fit and the conformational selection model sketched here. A protein with two distant binding sites is shown in black. A ligand fitting to one of the binding sites is shown in red.

The MWC model describes the transition from the T to the R state by a shift of the equilibrium between the two states. Without any ligand present the dominant state is the T state. If the protein is by chance in the R state the ligand can bind. On binding the equilibrium shifts towards the R state[6]. The name conformational selection for the MWC model arises from the selection of the R state over the T state by the binding of the ligand. Monod, Wyman and Chageux proposed that the maintenance of the symmetry in the protein enforces a concerted change of a symmetric protein. Due to required symmetry negative cooperativity is excluded. Extensions of this model are able to describe non symmetric proteins and and negative cooperativity[10].

Independently Koshland proposed a general version for induced fit in the late 60's in which a ligand, binding to a protein, changes the binding site to activate the protein for catalysis[11]. Koshland, Nemethy and Filmer expanded the model to the KNF model of induced fit in response to ligand binding[7, 8]. This model suggests that in the absence of the ligand only the T state exists. On binding of a ligand to the protein a transition of primarily the binding site but also of the entire protein towards the R or active state is enforced. As this results in a sequence of events,

this model is also known as the sequential model.

The current view on allostery suggests that conformational selection answers the question *which of the conformations of the protein will bind* while induced fit answers *how a protein binds its ligand if their shapes do not match well*[12]. Therefore the two models should not thought of to be mutually exclusive and are not applicable to different proteins. But, rather, the models are too strict and represent the extremes in a spectrum[13, 14].

2.1.2. Expanded View on Allostery

As the research on allostery progresses newer findings come up which have to be incorporated and must be explained by a general model of allostery. In the following some of them are reviewed briefly, and primarily for the interest of the reader. But it also shows the limitations current models face.

Roughly 15 years ago proteins such as pyruvate kinase M1[15] and phosphofructokinase[16] have been found to be convertible from nonallosteric to allosteric proteins by single point mutations[17–19]. Along the same lines e.g. Frauenfelder et al. demonstrated the ability of typical nonallosteric proteins such as myoglobin to show allosteric behavior[20]. A variety of other studies suggests that nonallosteric proteins can indeed be regulated by binding of a ligand to a distant binding site. These binding sites can also be on the surface of a protein and can only be identified by extensive studies[21–24]. Those findings led to the extreme proposal that all proteins might be allosteric[25]. Cooper and Dryden found that changes in the entropic contribution can lead to an activation of the protein even without having a conformational change[26]. This finding contradicts in some ways the general idea of allostery being encoded in the structure and poses limitations for simple techniques such as gaussian network models, which are reviewed later.

2.1.3. Experimental Characterization of Allostery

In vitro experiments are still the most believed in standard to gain new insights and to verify theories. Therefore it is of importance to understand what is actually measurable in experiment to understand what a theory should be able to predict to be verified by experiments.

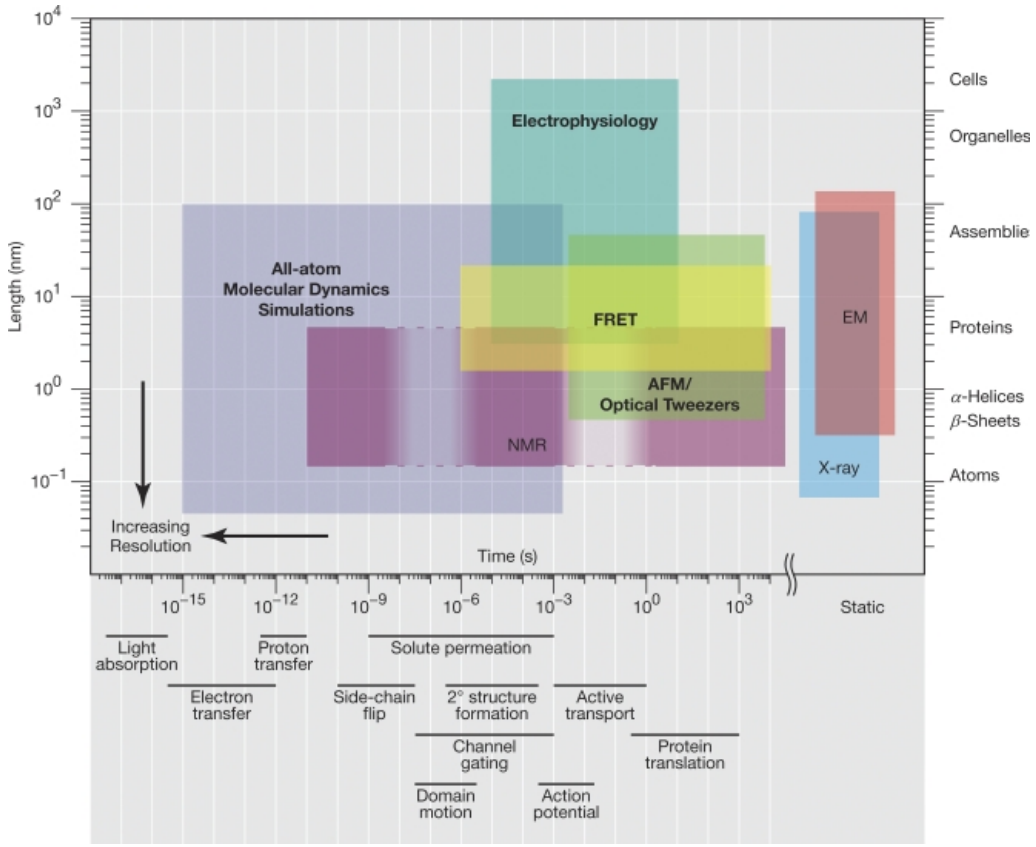


Figure 2.2.: Comparison of spatial and temporal resolution Spatial and temporal resolution of different techniques. Bold text indicates techniques being able to resolve single molecules. For improved interpretation objects corresponding to spatial resolutions are shown on the right and processes corresponding to temporal resolution are shown at the bottom[27].

An overview of the possibilities of current experimental techniques is given in figure 2.2. A possible way to quantify cooperative binding of a ligand to a protein is to use the Hill equation[28]. It is given by:

$$\theta = \frac{[L]^n}{K_d + [L]^n} \tag{2.1}$$

where θ is the fraction of ligand bound two the protein over the maximal number of ligands than can bind. $[L]$ is the concentration of unbound ligands and K_d is the dissociation constant. Experimental dissociation curves can be fitted to this equation giving the Hill coefficient n. The Hill equation describes the direct effect which the concentration of ligands has on the binding to a protein. Therefore it makes the Hill coefficient, a value for cooperativity, computable. This is nowadays the standard measure for cooperative binding, one appearance of allostery.
While the Hill coefficient is a method to determine the presence of cooperativity it does not provide any information about the underlying mechanisms. Therefore other techniques such as time resolved crystallography have been developed. This way processes such as the behavior of dimeric hemoglobin after ligand photodissociation can be studied on a 5 ns timescale[29]. Different types of NMR techniques specialized to detect allosteric networks[30] were developed. This is especially interesting for allostery without conformational change. Other simple NMR techniques are available to detect the different states relevant for the allosteric transition.
Some other methods which should be mentioned here are fluorescence microscopy[31], in which the position detection of fluorescent residues is used to study different allosteric states, and fluorescence resonance energy transfer (FRET)[32], which is a sensitive distance measurement of labeled residues to detect conformational shifts inside the protein.

2.1.4. Computational Approach to Study Allostery

To study allostery on a shorter time scale and in greater detail than currently possible in experiment, different computational approaches have been and are being developed. Even after nearly two decades of computational studies of allostery no best practice to describe allostery in all systems could be identified[33]. Therefore, some of the approaches not used in this thesis are presented in the following to provide the reader with alternative computational strategies to deal with allostery.

The methods are split into Molecular Dynamics (MD) based approaches and non MD based approaches[33, 34].

Non MD Based Approaches

Three different main types of non MD based approaches to study allostery are presented in the following. These are the feature or structural based models, structural surveys as well as the normal mode analysis which is widely used.
Structural surveys use structural information extracted from various conformational states of a protein in the Protein Data Bank (PDB). Daily and Gray used structural differences to identify hot spots of change marking potential allosteric pathways[35]. In a later work they combined this information to find networks of coupling among residues to aim at identifying allosteric pathways[35, 36]. Wolynes et al. introduced the concept of frustration as a basis for conformational change in proteins[37]. This concept suggests that the protein is not in its perfect folding state but rather slightly off, enabling the transition in the first place. It was later adapted to find residues being involved in the frustration to also participate in allosteric regulation[38].
The Elastic Network model is probably the most frequently used approach to allostery nowadays[39–41]. Ming et al. suggested to represent residues of a protein by single points being connected via a spring. This network contains information about elastic deformation modes from which one of the low frequency modes is proposed to be related to large scale conformational changes[42]. These large scale motions are thought to be the allosteric transition[43].

MD Based Approaches

In MD based approaches the dynamics of individual residues can be traced. MD is not specific to the system studied. This creates the advantage of having a general method to study the systems atomistic motions. But challenges are the short timescales a free MD simulation is able to explore and the further analysis. This analysis is needed to identify allosterically relevant motions within the thermal fluctuations. The timescales are often too short for allosteric transitions to occur[34]. Therefore, methods for enhanced sampling along the allosteric coordinate have been developed. The important ones are briefly reviewed here.
Biased MD[44], steered MD[45, 46] as well as targeted MD[47–49] are methods to study transitions in biomolecules. All of them apply an external perturbation along a predefined reaction coordinate between two states. Therefore the two states have

to be known in advance. This is a limiting factor for the systems which can be studied with these methods. While the general idea of the different methods are quite similar they differ in the way how the perturbation is applied[50].
Milestoning[51, 52] defines a reaction coordinate connecting the two states. Along this reaction coordinate different milestones are defined. Starting from each of the milestones simulations are started. These simulations end at one of the next milestones. By concatenating the trajectories a complete transition can be modeled.

With the different approaches the transition can be enforced. A free MD simulation is surely the best approach to the limited cases where a transition from one state to the other can be spontaneously observed, as such simulations have a minimal bias. Therefore it is suited best for a study focusing on extracting information about the allosteric motion. Because having the data of a MD simulation resembling a transition at hand the challenge remains to extract general information about allostery and even on the allosteric motion. The allosteric motion lies hidden beneath thermal fluctuations of the protein. To extract these information and be able to draw conclusions on allostery in general is the major goal of this thesis.

2.2. Hemoglobin - A Well Studied Model System for Allostery

The protein Hemoglobin (Hb) is one of the most essential proteins in the human body because it transports oxygen from the lungs to the tissue.For Hb to be able to transport oxygen the allosteric regulation is fundamental. The reason is that oxygen needs to be bound in the lungs and released in the tissue. This is mediated by the characteristic binding curve for cooperative binding known as the Hill curve.
The structure of Hb was first solved by Puritz et al.[54] in 1960. Since then Hb was one of the working horses for the study of allostery. Its structure (shown in figure

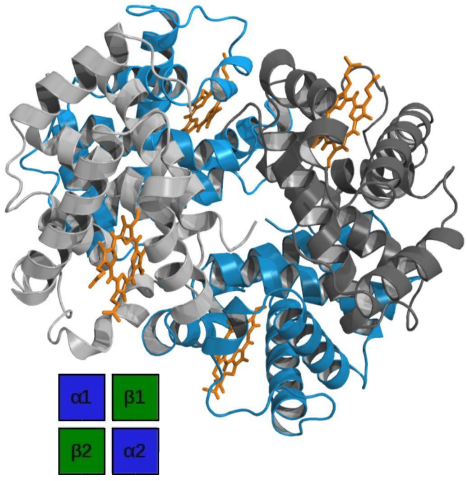


Figure 2.3.: Structure of Hemoglobin
A cartoon representation of the structure of Hemoglobin. The assembly is sketched in the lower left. Figure applied from [53].

2.3) consists of the two heterodimers which assemble into a homodimer. The different subunits consist predominantly of alpha-helical secondary-structure elements. In all subunits a heme group is bound which binds the oxygen.
The identification of conserved allosteric pathways in Hb[55] suggest a structural dominance in its regulation. However, studies show that additional factors such as binding of additional water molecules to the extra surface contribute to the allosteric regulation[56]. Together with the finding of an additional R state named the R2 state[57] it appears that Hb is not as simple a two state protein following the conformational selection model as it was long thought of.

Apart from its model role in the study of allostery Hb was also shown to undergo a T to R transition in free MD simulations on feasible timescales, making it a perfect system for the current study of dissecting the molecular details of the allosteric transition.

2.3. Molecular Dynamics Simulation

To study motions at a high spatial as well as at a high temporal resolution a wide range of techniques has been developed. Some of these techniques are shown in figure 2.2. As can be seen from that figure all techniques are rather limited in spatial as well as in temporal resolution[27]. What is not shown in this figure is an additional limitation due to the need to prepare the samples. For X-ray crystallography for example a crystal is required. This crystal has to be grown, a non trivial process and the real art of crystallographers.
As can be seen from figure 2.2 all-atom Molecular Dynamics (MD) Simulations are especially suitable for small time and length scales. In the following the method of MD simulation is reviewed briefly. Not only from a technical aspect but also looking at its limitations.

2.3.1. Principal Concept

The most detailed and accurate description of atomic motions on short time scales are quantum mechanical (QM) descriptions. In QM simulations the Schrödinger equation is solved by applying a numerical solver. However, these methods are limited to a few hundred atoms only and are computationally very expensive[58]. To

push the time and length scale further, MD simulations have been developed. Here the Schrödinger equation is not solved directly anymore but three major simplifications are applied. These are described in the following.

2.3.2. Born-Oppenheimer Approximation

Using the Schrödinger equation the wave function of electrons as well as for the nuclei have to be solved simultaneously. To reduce the complexity arising from this coupling Max Born and J. Robert Oppenheimer derived the Born-Oppenheimer approximation. Due to the difference in weight between the nuclei and the electrons, and the resulting difference of timescales the motions happen on, the wave functions can be separated and solved independently. For macromolecules the main interest is on the motion of the nuclei. Therefore it suffices to treat the electrons as a reason for a force, i.e. bonded as well as Van der Waals interactions, that act upon the nuclei and to in addition calculate an electrostatic interaction between the nuclei carrying the charge of the nucleus and the electron.
In addition another simplification can be made. The nuclei are treated classically in MD simulations.

2.3.3. Classical Description of Nuclear Dynamics

In nature especially light nuclei are able to tunnel through other nuclei. The simplification made is to excluding this behavior from the possibilities of the system. Another problem arises from the high frequency vibrations of covalent bonds. This quantum oscillator can not be described properly by a classical harmonic oscillator. This can either be corrected by adding an additional energy term or by adding bond and bond angle constraints. Doing this a classical description can be used to calculate the atom dynamics. In Molecular Dynamics simulations all forces are described by classical potentials leading to a classical force which is numerically integrated using for example the velocity-verlet algorithm.
In the next subsection these potentials as well as the treatment of other atom-atom interactions are reviewed.

2.3.4. Classical Potentials

The interaction of atoms are described by classical potentials. These potentials are briefly explained in this section. For further reading we refer to the GROMACS

manual[59] as well as to the relevant papers[60–62].
A reasonable classification of the potentials ($V(x)$) is the classification into (i) bonded and (ii) non-bonded interactions.

$$\begin{aligned} V(x) &= \sum V_i \\ &= V_{\text{bonded}} + V_{\text{non-bonded}} \end{aligned} \tag{2.2}$$

While bonded interactions are the result of the Born-Oppenheimer approximation and imagined as a chemical bond between two atoms, non-bonded interactions are long range interactions such as the Coulomb potential.
In the rest of the subsection bonded as well as non-bonded interactions are described in more detail.

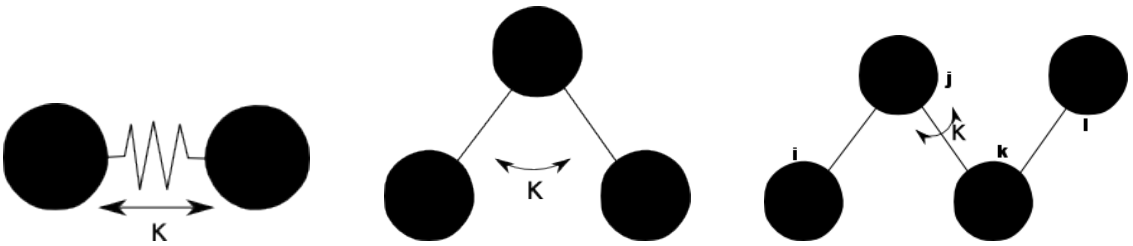
Bonded interactions

Bonded potentials represent chemical bonds. A list of atoms being connected via a chemical bond is fixed. All bonded interactions are described by a harmonic potential.

$$V_{bonded} = \frac{1}{2}\kappa(x - x_0)^2 \tag{2.3}$$

The x_0 in equation (2.3) describes an equilibrium conformation. Due to thermal fluctuations and interactions in and between molecules a motion around this state occurs. The current conformation is given by x. κ is the force constant driving the molecule back to its undisturbed conformation.
The constants κ and x_0 are defined and vary together with x in their meaning for the different types of bonded interactions.
Three types of bonded potentials are described. These are the bond, angle and dihedral potential.

Bond potential The bond potential describes the interaction of two atoms being connected via a chemical bond (see figure 2.4a). The bond is described via a spring. In this simple description of a two body potential the parameter x_0 describes the equilibrium distance between the two atoms.
To improve time performance the bond potential is often not used and the distance between two atoms is constrained. This is done by special algorithms which are



(a) bond **(b)** angle **(c)** dihedral

Figure 2.4.: Bonded interactions Illustration of the different bonded interactions. Taken from [63].

applied for to constrain the distances (i.e. LINCS[64] or SHAKE[65]).

Angle potential Taking three atoms into account an angle between the outer two can be defined (figure 2.4b). The angle potential leads to a vibration around the equilibrium angle x_0.

Dihedral potential Taking one more atom into account leads to the dihedral potential. To define the dihedral angle the orientation of two plains to each other have to be considered. One of these plains is formed by the atoms i, j, k in figure 2.4c. The other by j, k, l. The angle of intersection gives the dihedral angle.
For the angle to be zero the *cis* configuration is chosen as a reference.
Furthermore, a special type of dihedral potential, called improper dihedrals, is applied to keep planar groups planar.

Non-bonded interactions

Non-bonded potentials act on a distance. Therefore their partners are not fixed and can change during the simulation. To improve performance the neighbor list is not updated in every step but only after a predefined number of steps.
In MD simulations the non-bonded interactions described are the Lennard-Jones and the Coulomb potential.

Lennard-Jones The Lennard-Jones potential has two different contributions. An attractive term $1/r^6$ is given due to the polarisability of the electron cloud around the atoms. This term is known as the van der Waals force. The second term taken into account is a repulsive $1/r^{12}$ term. While the attractive term is physical the repulsive term is an approximation of the Pauli exclusion principle preventing the

atoms from getting too close. In total this leads to:

$$V_{LJ} = \frac{c_1}{r^{12}} - \frac{c_2}{r^6}. \tag{2.4}$$

The parameters c_1 and c_2 depend on the atom pairs considered and can be determined experimentally.

Coulomb The Coulomb potential describes electrostatic interactions. Having a $1/r$ law for the coulomb potential far away atoms have to be taken into account. Therefore the calculation of the coulomb force is computationally very expensive ($\mathcal{O}(N^2)$). In modern MD simulation software such as GROMACS the effort is reduced by using the periodicity of the system to go into Fourier space. There a method called Particle Mesh Ewald[66] can be applied. It reduces the computational complexity to $\mathcal{O}(N \times \log N)$.

2.3.5. Force Fields

The parameters described before are collected in so called force fields (FFs). For biomolecules such as proteins specific FFs are designed describing the properties of e.g. the different amino acids. For proteins two of the most widely used FFs are the Amber FF[67] and the Charmm FF[68]. The validity of the FFs arise from the comparison to experimental data. As mentioned before FFs are derived from QM simulations but improved by comparison with different type of experimental data sets. The ways how this is done and the included types of experimental data vary between the FFs. In the past most of the FFs were shown to reproduce a variety of experimental data[69]. What has to be taken into account is that not all FFs are equally suitable to study a given observable but it is important to understand which FF is able to reproduce what kind of experimental data.

2.3.6. GROMACS Algorithms

To perform MD simulations a variety of algorithms have to be applied to decrease artifacts introduced by the simulation. In this subsection a short introduction to these algorithms is provided without going to much into detail. These algorithms together with the FF parameters are used by GROMACS to perform MD simulations.

In MD a limited number of atoms can be simulated only. To prevent artifacts arising from boundaries in the simulation box, periodic boundary conditions (pbc) are applied. If using pbc a copy of the box is considered to sit at any boundary.
The box forms a thermodynamical microcanonical ensemble (NVE number of atoms, volume and energy conserved). But the physiological ensemble is the isothermal-isobaric ensemble (NPT number of atoms, pressure and temperature). To get from the NVE to the NPT a pressure and temperature coupling of the box has to be applied. For this purpose different algorithms exist (e.g. velocity rescaling[70] for temperature and Parrinello-Rahman[71] for pressure).
In addition GROMACS provides a variety of analysis tools to extract observables from the trajectories (i.e. simulated data) as well as an effective framework for programming individual tools.

2.3.7. Limitations

As mentioned above the simplifications made by the FF approach and the numerical integration lead to limitations of MD simulations. Due to the classical description a chemical bond can neither be formed nor broken. The numerical integration leads to an accumulation of the numerical error. If the timestep is chosen too large this results in a numerical error. In addition the accuracy of the FFs is not perfect. It has been demonstrated that the FFs are able to reproduce specific experimental results well. Nevertheless the FFs are approximations of the nature and will therefore always be limited in accuracy.
Furthermore, the polarizability of the molecule is not treated in classical simulations. Nevertheless various models for treating polarization do exist[72–74] but are not generally established yet.
As described previously the timescales which can be explored by the simulations are now reaching the μs length. This is too short for a lot of processes occurring in proteins. Looking into the future one can expect the reachable timescales to coevolve with the computing power.

3. Methods

In this chapter the techniques used in this thesis are introduced. These are Principal Component Analysis, Functional Mode Analysis based on Partial Least Squares, and Essential Dynamics. Furthermore the decomposition into orthogonal motions is reviewed, followed by an overview of the simulation setup used to produce the trajectories used to study Hemoglobin. At the end a short review on relevant findings for this thesis of the work previously done by Hub et al.[75] and Vesper and de Groot [53] is provided.

3.1. Principal Component Analysis

Molecular Dynamics simulations sample a high dimensional space. This space has $3N$ dimensions with N being the number of atoms. Having a system of 10.000+ atoms thus leads to a very high dimension. The high dimensionality imposes a difficulty to distinguish between the important motions and noise.
To face this difficulty Principal Component Analysis (PCA) can be used. First developed by Pearson [76] and further developed by Hotelling [77] PCA identifies large scale linear motions. Therefore the direction of the motion with the largest variance is defined as the direction of the first basis vector of a transformed new basis system. The second basis vector is chosen to be in the direc-

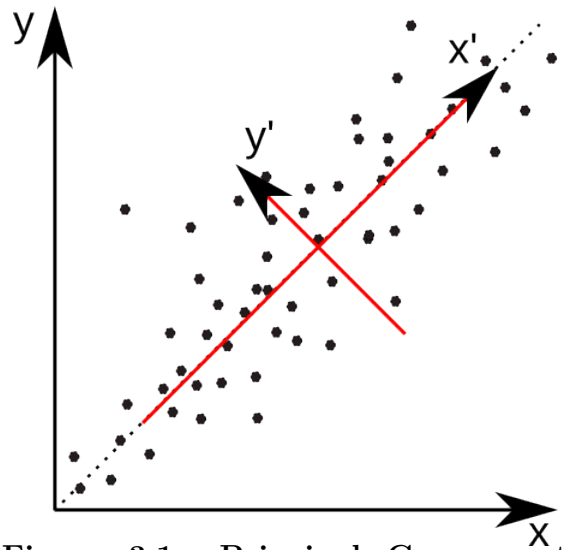


Figure 3.1.: Principal Component Analysis Schematic representation of the transformation of the basis performed by PCA. The coordinate system is shifted such that the first coordinate is along the direction of the greatest variance.

tion of the largest variance orthogonal to the first basis vector. This procedure is followed untill the new coordinate system has the same number of basis vectors as the old one.
In a mathematical sense, a matrix of covariances $\mathbf{C}$, composed of the coordinate vectors of the atoms x, has to be defined:

$$\mathbf{C} = \langle (x(t) - \langle x(t) \rangle_t) \cdot (x(t) - \langle x(t) \rangle_t)^T \rangle_t \tag{3.1}$$

This covariance matrix $\mathbf{C}$ is diagonalized. In the representation of the covariance matrix in the diagonalized basis the elements of the matrix are the eigenvalues. Therefore to find the basis of the covariance matrix a simple eigenvalue, eigenvector decomposition can be applied. The eigenvalues related to the eigenvectors show the variance along the respective eigenvector. Frequently, correlated motions with a large eigenvalue will be limited to the first few eigenvectors. By defining a cutoff for the eigenvalues only the relevant collective coordinates can be selected to be displayed. This can be used to study the important motions hidden in the data i.e. the trajectory.
One of the advantages of PCA is the possibility to find the motions contributing dominantly to the overall motion of the protein. In this study we use the one dimensional subspace with the highest eigenvalue to describe the quaternary transition of Hb. One of the main disadvantages of PCA is the limitation to a linear subspace. If the data lives for example on the popular swissroll subspace (see figure 3.2), a PCA would not be able to find a suitable basis to represent the motions on the 2 dimensional subspace embedded in the full 3 dimensional space.

3.2. Functional Mode Analysis based on Partial Least Squares

PCA was developed to find the largest variance along a linear vector in the data space. A priory it is often not clear if the motion with the largest variance is indeed the functionally most relevant motion. Therefore more advanced methods which are able to identify motions being correlated to a functional property are useful. To fulfill the need for this a method called Functional Mode Analysis (FMA) has been developed[78].
FMA uses a one dimensional functional property f and tries to predict this property

by a linear combination of the data. The high dimension of the data in general leads to the problem of overfitting. To overcome this problem a PCA can be conducted first. Only the first n numbers of components of the new basis, given by the eigenvectors, are used as a dimensionality reduced data set. FMA builds a model that describes the functional property by a linear combination of the data. The goodness of fit can be tested by crossvalidation. To do so, some of the data points are left out of the fitting procedure and subsequently the functional property for them is recalculated by applying the model.
In FMA quite a large number of PCA components is needed. Therefore, the Partial Least Squares extension of FMA (PLS) was introduced[79]. In PLS the dimensionality is reduced by a partial least squares regression instead of using PCA. In a mathematical sense a model is defined such that

$$\mathbf{f} = \mathbf{b}^T \cdot \mathbf{V} + \boldsymbol{\epsilon} \tag{3.2}$$

maximizes the covariance of the new coordinates $\mathbf{V}$ in reference to the functional property $\mathbf{f}$. $\mathbf{b}$ is the model that describes the mapping from the coordinates to the functional property while $\boldsymbol{\epsilon}$ describes the residual to be minimized in PLS. To calculate the new coordinates an iterative procedure is applied which defines the new basis such that the basisvectors are orthogonal to each other.

3.3. Essential Dynamics

Essential Dynamics (ED) uses predefined collective coordinates to alter the motions sampled by MD. This collective coordinates can come from other analysis tools e.g. PCA or PLS. In the first implementation sampling was limited to the principal components only[80]. By the reduction of conformational space transitions in the simulation might actually be prevented artificially. Therefore later implementations also included the other degrees of freedom and only enhance or restrict the sampling in the direction of the collective coordinate.
One way of enhancing the sampling into the direction of one collective coordinate is to grow a potential at the point of dense sampling. This additional repulsive potential increases the possibility to move away from the potential minimum. This method is called conformational flooding[81, 82].
The method used in this theses uses an additional potential along a set of collective

coordinates. This potential is chosen to be attractive and therefore the dynamics of the protein is reduced in one dimension. A harmonic potential is the simplest choice. For the spring constant of the potential different values can be tested. While the collective coordinate chosen is restricted to stay close to its target value, all other coordinates are free to move. Due to the orthogonality of the dimensions, the motion in the other dimensions should not be affected if they are independent.
Therefore, if in contrast, a large effect is induced along one of the other coordinates, this hints at an interplay between the motion being altered directly by ED and the motion being influenced by this. While a lot of methods are able to detect a correlation it is often hard to identify the causal relationship between one motion and another. This is made accessible by this method.

3.4. Disassembly into Orthogonal Motions

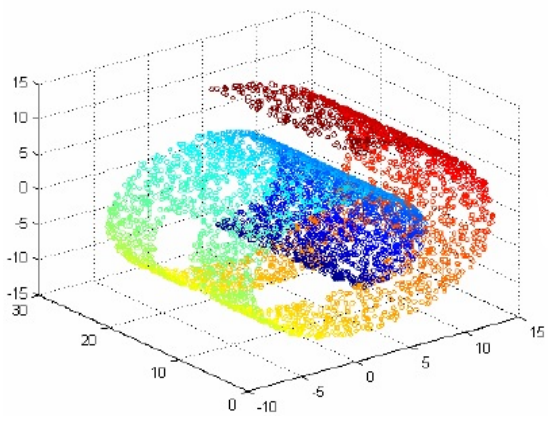
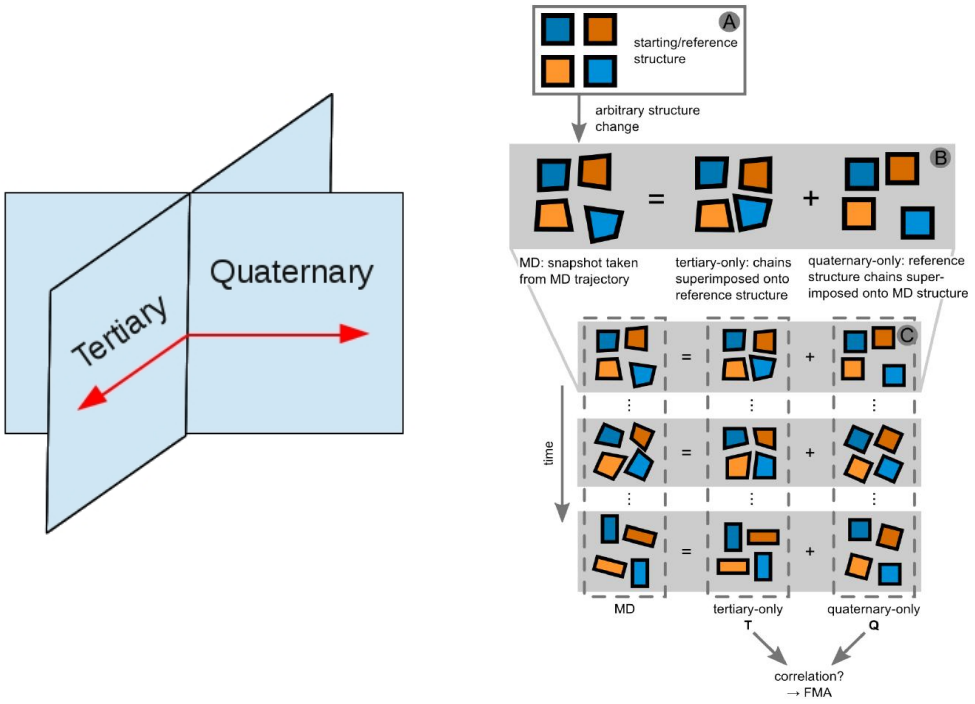


Figure 3.2.: Swissroll The popular swissrole embedding of 2D subspace into a 3D space. *Source: macs.citadel.edu*

The relative motion of the subunits, called the quaternary motion, and the internal motions of the subunits, called the tertiary motion, are motions in orthogonal spaces. To separate the quaternary from the tertiary motions an algorithm was developed by Vesper and de Groot[83]. Motions in the tertiary space can easily be detected by solely locking at the subunits independently. This can be done by splitting the trajectory into trajectories consisting of one subunit only. By fitting the trajectory of the single subunits onto the starting structure of the simulation, the overall rotational and translational degrees of freedom are removed. The removed degrees of freedom then resemble the basis of the quaternary motion. To get the quaternary motion the starting structures of the subunits are fitted independently onto the trajectories of the single subunits and subsequently the motion of the rigid subunits are merged.



(a) The motion of the subunits relative to each other can be separated from the internal motions of a subunit. The reason for that is the orthogonality of the vector spaces spanned by the subspace of relative (quaternary) and internal (tertiary) motions.

(b) The overall motion of the protein is disassembled into a quaternary and a tertiary part for each frame individually. Combining the quaternary and tertiary motion the overall motion can be recovered. Adapted from [83].

Figure 3.3.: Tertiary - Quaternary disassemble of a trajectory

3.5. Simulation Setup

The simulations used for the majority of the analyses are the same as used before by Vesper and de Groot[53]. The starting structure was the T state structure taken from the Protein Data Bank (PDB) with the ID 2HHB[84]. While the reference structure for the R state often used in the analysis has the PDB ID 1IRD[85]. All simulations were run using the GROMACS 4.5.3 software. Simulations which are simulated later (noted in the respective results section) are simulated using GROMACS 4.6. For all simulations the force field used was GROMOS 43a2[86]. The simulations were performed in explicit SPC water[87] in a box with periodic boundary conditions. As a salt concentration 150 mM of sodium chloride was chosen and sodium ions are added to neutralize the simulation box. The box was chosen to be dodecahedral. For the electrostatic interactions PME[66, 88] was applied. Short range non polar interactions were described by the Lennard-Jones potential and were cut off at 1.4 nm. LINCS[64] and SETTLE[89] were used to constrain protein and water bond lengths allowing a time step of 2 fs. The neighbor lists were updated every 10 steps. The simulation temperature was hold constant at T=300 K using the velocity rescaling[70] algorithm with a time constant of $t = 2.5$ ps. For the pressure coupling the Parrinello-Rahman barostate[71] was applied to keep the box at a pressure of 1 bar. Here a time constant of $\tau = 5$ ps was applied.
For the simulations run by Vesper and de Groot a simulation of 200 ps with position restrains on the backbone was used after the energy minimization for relaxation. These restrains are not applied for the new set of simulations.

3.6. Basis

The basis of the work presented in this thesis is the work by Hub et el.[75] and Vesper and de Groot[53]. Hub et al. found a spontaneous transition of Hb from the T to the R state. When starting the simulation from the R or the R2 state the simulations sample the R state only. Hub et al. concluded that the simulation of Hb shows a preference for the R state. This contradicts the experimental prediction to find the unliganded Hb predominantly in its deoxy T state. Nevertheless it is not an isolated result as various other simulation based studies on Hb showed the same tendency towards the R state[90, 91]. Hub et al. detected the correlation between the quaternary or intersubunit motion and the tertiary or intrasubunit

motion. In this stage of the work the tertiary motion was defined as the motion within the subunits. This has not changed in the following, but Hub et al. defined the quaternary motion as the overall motion of the protein. Vesper and de Groot[53] developed a method to define the quaternary motion as the intersubunit motion only (see section 3.4). While Hub et al. searched for correlations between the first PCA vector of the tertiary and quaternary space Vesper and de Groot used the PLS analysis (see section 3.2) to identify the motion in tertiary space being most correlated to the quaternary motion. They found this motion to be quite different to the crystallographic transition of the subunits as described by the difference vector of the R and T state. On the other hand the quaternary motion could be identified to be similar to the crystallographic T→R transition.

4. Results and Discussion

In the preceding chapters an overview of the state of the research in general as well as of the state of hemoglobin research was reported. The following sections report on the new findings as obtained in the course of the master project. All methods used in these sections are described in chapter 3.
The main branches of the project are (i) the investigation of a functional relation between the quaternary and the tertiary motion which goes beyond a correlation, as well as (ii) an investigation on the backbone - sidechain coupling in Hemoglobin to find communication-pathways in the protein and (iii) a PLS based contact analysis to address the intersubunit communication via formation and breaking of contacts.

4.1. Tertiary - Quaternary coupling

This section is split into two logical subsection. In the first subsection a preliminary result from Hub et al.[75], who reported a stronger coupling of the tertiary motion of the β subunits than of the α subunits to the quaternary motion, is investigated with the method of PLS. This is followed by a further investigation of the coupling between tertiary and quaternary motion in general. As Vesper and de Groot[53] found a strong correlation between the first eigenvector of the quaternary motion and a collective vector in tertiary space the question of functional coupling beyond correlation remained unanswered. This will be addressed in the second subsection.

4.1.1. Correlational asymmetry of α and β subunits

In 2010 Hub et al. investigated the coupling between the tertiary and quaternary motion in Hb[75]. In their study they used the difference vector between the R and the T state of Hb as represented by the X-ray structures 2HHB (T state [84]) and 1IRD (R state [85]). For the tertiary motions the difference vector of the single subunits was used. Onto these two vectors the trajectories were projected.

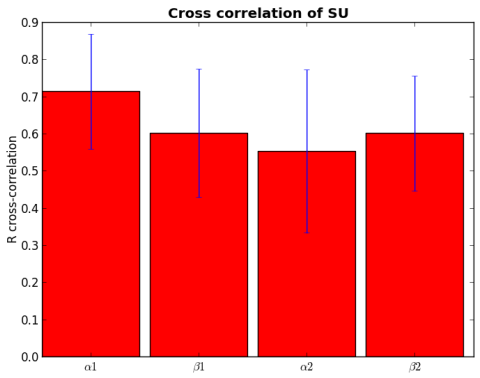


Figure 4.1.: No d ifference in correlation between subunits The histogram shows the Pearson correlation of the original data for Qv and the predicted PLS model data constructed from half of the trajectories. Different bars correspond to different subunits. Values displayed are from the half of the trajectories not used for model building (crossvalidation only). Errorbars are standard deviation.

For the shared information between the projections of the trajectories onto the two vectors, Mutual Information[92] was applied. Mutual Information is a measure of the content of information that is shared between two properties. In this case the two properties under investigation had been the tertiary and quaternary space. When investigating the mutual information of the quaternary - tertiary coupling, Hub et al. found that the quaternary motion shares more information with the tertiary motion of the β subunits than with the tertiary motion of the α subunits. This resulted in a T→R transition model where the tertiary transition (t→r) of the β subunits precede the tertiary transition of the α subunits.

Vesper and de Groot introduced the usage of PLS (see section 3.2) as well as a novel method to separate the trajectory into orthogonal vectorspaces of quaternary and tertiary only motions (see section 3.4)[53]. By having the quaternary-only and tertiary-only motion a more appropriate disassembly of the vector space was found to address the same question. The analysis is further improved by PLS to find the highest possible correlation between the functional mode of the quaternary transition as described by the first eigenvector of the PCA analysis. This enables us to study a higher dimensional subset.

The first eigenvector of the PCA analysis of the quaternary only motion was taken and the trajectory of the quaternary motion was projected on that motion. This formes the functional property needed as a one dimensional input data for the PLS analysis. Subsequently the PLS analysis was run on the tertiary only motions of the different subunits. To calculate the error as well as the mean, a subset, consisting of the half of the trajectories, was used to build a model. This was followed by the prediction of a single trajectory by the model calculated. For the prediction a value for the correlation between the prediction and the original data was calculated us-

ing Pearson correlation. This was done for all trajectories independently. From the correlation values calculated for the crossvalidation, a mean was calculated as well as the standard deviation. The result is shown in figure 4.1. Here the bars are the mean values of the Pearson correlation for the different subunits and the errorbars are the standard deviations.

In figure 4.1 the mean of the different subunits varies notably. The trend reported by Hub et al.[75] is not be confirmed by this analysis. Neither a trend towards a stronger coupling between the quaternary motion and the tertiary motion of the β subunits can be identified nor a difference between the **1** (i. e. α_1 and β_1) and **2** subunits is found. All the differences between the subunits are within the error.
While Hub et al. predicted a concerted motion in which the β subunits of Hb undergo the transition before the α subunits do so this could not be confirmed looking at the quaternary - tertiary coupling in the orthogonal vector space and with the motions of highest functional correlation.
A possible explanation for the results is the difference of tertiary motion. Vesper and de Groot found that the quaternary transition vector, as given by a PCA analysis on the quaternary only motion, is highly similar to the X-ray difference vector. But a similar analysis on the vector found by the PLS analysis to identify the most correlated motion in tertiary space revealed the difference for the vector of the tertiary motion to the tertiary crystallographic difference vector. As the PLS vector is more strongly coupled to the quaternary motion, it is more likely to be the correct tertiary transition taking place in the allosteric transition of Hb than the difference vector given by the X-ray difference between the R and T state. This adds weight to the analysis of the transition by the method presented here and therefore leaves it unclear which role the difference vector between the X-ray structures plays. These results motivated us to revise the model of Hub et al. including the new findings resulting in the model shown in figure 4.3 a). Here the transition of the subunits occurs simultaneously and is either preceded or followed by the quaternary change.

In addition Hub et al. investigated the sequence of events in which the tertiary and quaternary transition of the different subunits occurs. As the interpretation of the tertiary motion changed drastically, a new investigation on the sequence of events was therefore performed. To detect a delayed coupling between the quaternary and tertiary motion the single trajectories were taken and frames of the tertiary motion

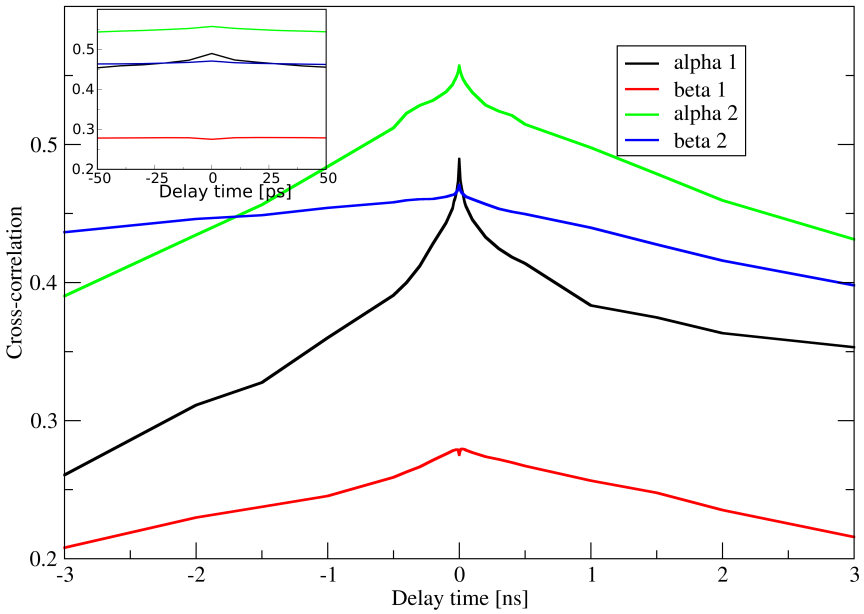


Figure 4.2.: No evidence for delay To detect a delay of the quaternary or tertiary motion a timeshifted PLS analysis was run to find the best fitting time shift. The inset is a zoom-in at the time around zero. Here the results are demonstrated independently for the different subunits. No better match for the PLS analysis can be found than the data without any timeshift. However referring to figure 4.1 it has be noted that the results have to be interpreted with care as the errors seem to be non-negligible.

were removed at the end or the beginning. To remain with the same number of datapoints as frames, the same number of points were removed from the functional property only at other end (i. e. if a frame is removed at the beginning a data point is removed at the end and vice versa). This resulted in a shift of the assignment between the functional property and the frame in the trajectory. Subsequently the trajectories were concatenated and half of them were used to build a model while the other was used to crossvalidate the model. The results for the crossvalidation values by the different time used for shifting are plotted in figure 4.2. For the analysis 10 ps were the shortest delay possible to analyze. This was due to the trajectory output frequency.
As can be seen for all subunits but the β_1 subunit the best value for the crossvalidation is at a 0 ps time delay. For the β_1 subunit the 0 ps time delay is minimally worse than the values close to it. In general it seems as if the minimal time delay gives the best results to predict the quaternary motion by the tertiary motion of the different subunits.

The results suggest that a delay of the tertiary and quaternary coupling is not present in the given trajectories. However, the results have to be interpreted with care. Especially looking at the error of the individual subunits as demonstrated by figure 4.1, a delay in the coupling can not be excluded at this point.
Therefore, we modified the transition model once more to also include this result. This leads us to a transition from the T to the R state as depicted in figure 4.3 b). It is likely that the subunits transit simultaneously from the t to the r state. Furthermore a simultaneous transition of the quaternary and the tertiary structure from the T to the R state is suggested by the absence of a delay in the simulations. Therefore the model depicted in figure 4.3 b) is the most likely way the Hb transits from the T/t to the R/r state. However the model of 4.3 a) where a delay between quaternary and tertiary transition is present can not be ruled out.

4.1.2. Essential Dynamics Simulations

So far the coupling of quaternary and tertiary motion was inspected by the correlation of these motions in a free MD simulation. However the causal effect one motion has on the other can not be identified by this analysis. To move beyond correlation and to be able to address the interplay between the subspaces, formed by the disassembly into quaternary and tertiary motions, Essential Dynamics (ED)

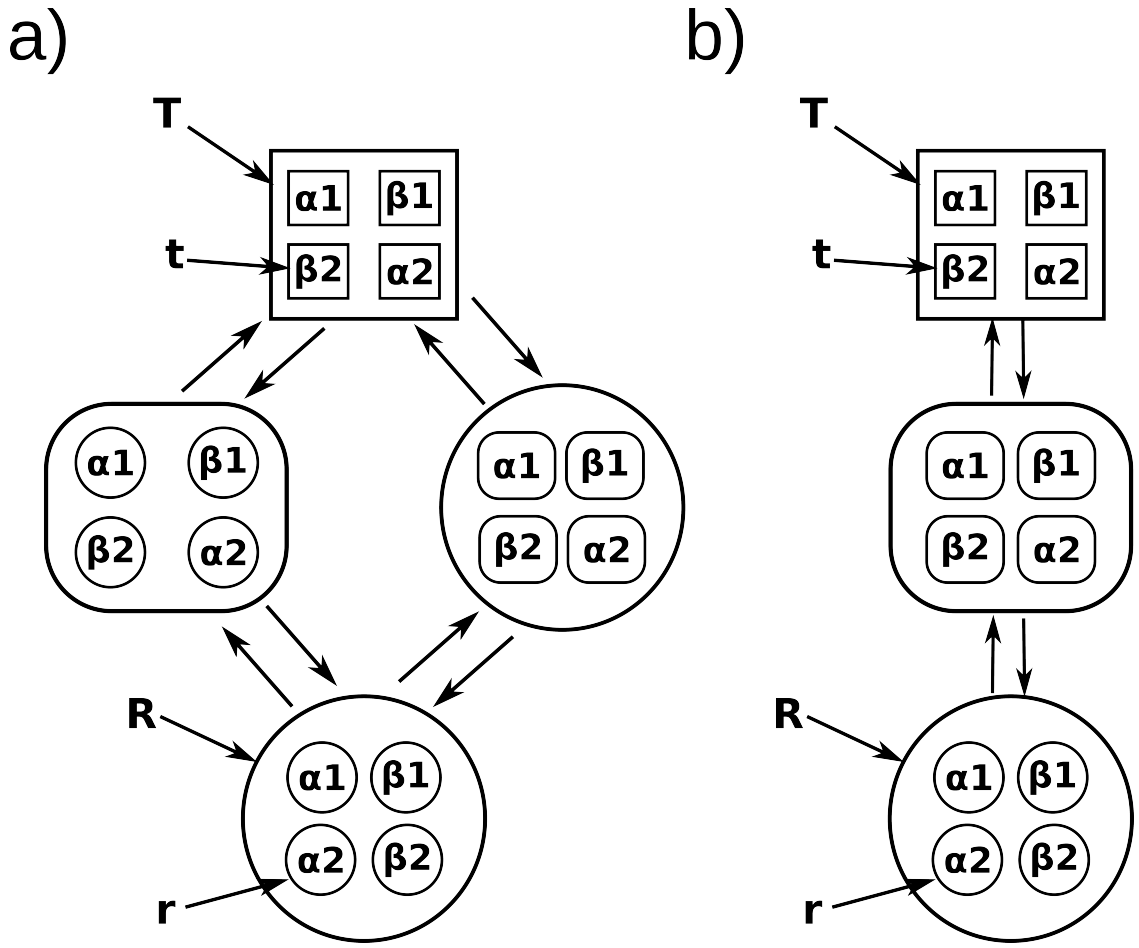


Figure 4.3.: Schematic representation of transition pathway between R and T state The figure is applied from [75]. In the original form a dominance of the β tertiary - quaternary coupling led to the conclusion of a more likely tertiary β r state than t state. As the result could not be confirmed in the current PLS model the symmetry between the subunit is not broken. In a) the variant with a timeshift involved in the transition is depicted. Taking the left route the transition of the tertiary state of the independent subunits is followed by the transition of the quaternary state. Conversely the right route shows the dominance of the quaternary state transiting first to the R state and followed by the tertiary transition.
In b) the simultaneous route as suggested by figure 4.2 is depicted.

was used to modify the simulations. By manipulating one collective coordinate the effect on the other, orthogonal coordinates can be studied by ED. Therefore we are able to move beyond correlation and get to an interaction between the orthogonal motions.

To study the transition a set of MD simulations was performed. The simulations were set up according to the protocol previously reported by Hub et al.[75] and described in section 3.5. No equilibration with restrains on heavy atoms was performed. Therefore the free simulations previously reported by Vesper and de Groot [53] can not be used directly for comparison. The vectors used for the projection (Tv and Qv) are the vectors previously obtained by Vesper and de Groot.

We performed 16 free MD simulations (free, black) as a reference set. In addition, 15 simulations were set up with 3 of the subunits restricted in their respective tertiary t states (4 for α_1, β_1, α_2 and 3 for β_2) while one subunit is simulated freely (3 SU restricted in t, red). These simulations were performed to investigate the effect of the tertiary motion on the quaternary motion as well as on the tertiary motion of the other subunits. The restrictions are inflicted on the protein by applying a harmonic potential on the collective coordinate obtained from the PLS analysis (Tv). After using different force constants we chose it to be 200 kJ/mol/nm^2. This force constant was found to be enough to hold the subunit in the respective state but does not restrict it too much. For the same purpose another set of simulations was performed with one subunit being chosen from a random structure being clearly in the r state. This r state subunit is fitted on the respective subunit in the Hb starting structure and subsequently the subunit of the starting structure is replaced by the r state structure. The result is a starting structure that has three subunits in the t state, one subunit in the r state and has the quaternary structure of the T state. That gives an additional starting structure for a set of simulations where the remaining t state subunits are restricted once more in their starting states. This gives another set of 3 simulations with α_1, 3 with β_1, 4 with α_2 and 4 with β_2 taken from r. The same setup procedure was used without restricting the subunits starting from the t state yielding another set of 16 simulations with 4 simulations with each of the 4 subunits starting from the r state (same, but 1 started from r, green). One more set of simulations was performed where all of the subunits were chosen with r state starting structures. They were all fitted onto the quaternary T state to have a quaternary T state starting structure with a tertiary r state starting structure (free,started from r, blue). The last set of simulations were free simulations with all

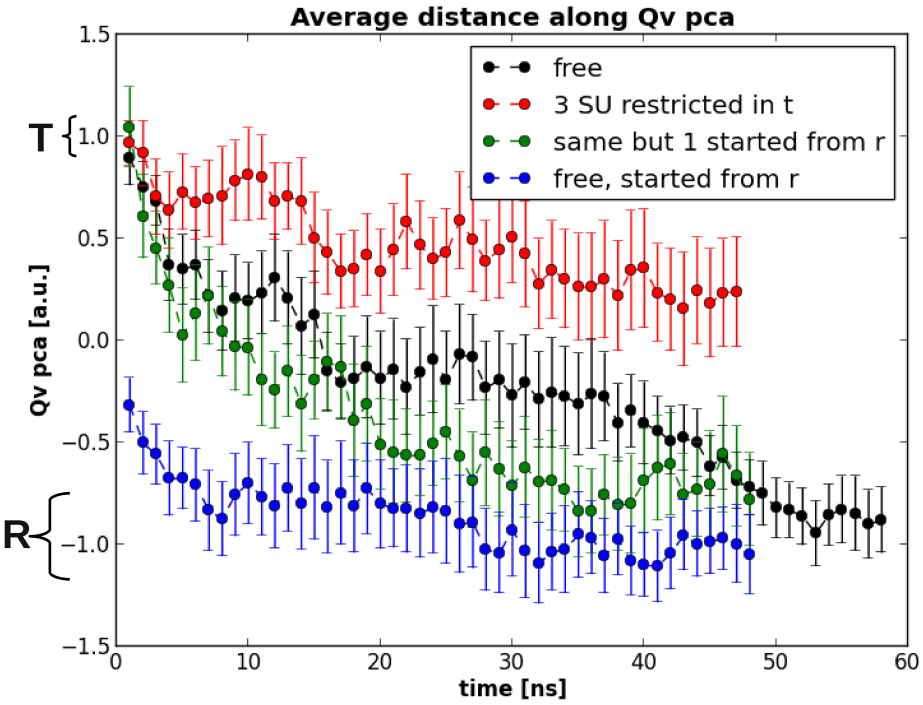


Figure 4.4.: Quaternary projection of ED simulations The black line shows the transition along the first PCA eigenvector of the quaternary motion (Qv) for a free MD simulation. In red the transition for 3 subunits being restricted in tertiary t state are shown. The green curve shows the same setup as the red curve only with the free subunit being started from the tertiary r state. The last, blue curve, is again a free MD simulation. Here the subunits are chosen from a structure being in the R state and are then fitted to a quaternary T state starting configuration. Error-bars are errors of the mean.

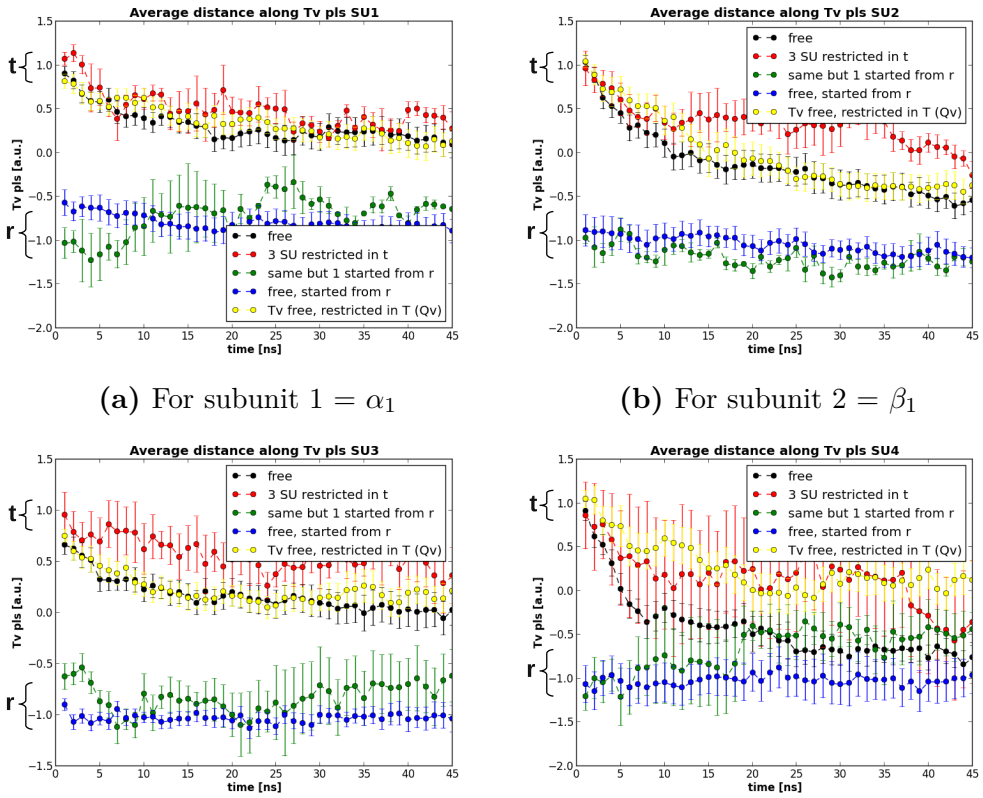
of the subunits freely moving but with the quaternary T state being fixed by the same harmonic potential before applied to the tertiary motion (Tv free, restricted in T(Qv), yellow). This yields another 16 simulations.

To analyze the trajectories the simulations are projected onto the first quaternary PCA eigenvector (Qv) already used by Vesper and de Groot. From the simulations the average projection at a certain time is calculated as well as the standard error of the mean. The evolution of the different simulations along the quaternary vector (Qv) can be read from figure 4.4. For the projection on the tertiary vector the respective PLS vector (Tv) proposed by Vesper and de Groot was used. The results are shown in figures 4.5 a)-d). For the simulations where 3 of the subunits were manipulated by ED only the freely moving subunits were used to calculate the

mean and its error in the projection to the Tv vector. All of the projections were scaled such that +1.0 resembles roughly the T/t state and -1.0 is around the R/r state.
The projection of the free simulations is the reference for the simulations with the manipulated subunits. Among them are the simulations started from the same 2HHB initial T state structure but with the harmonic potential applied to three of the four subunits to stay in the tertiary t state. As can be seen in figure 4.5 these simulations do also show a restriction on the quaternary transition. While the free simulations transit from the T to the R state the restricted simulations have a significantly reduced tendency for the transition. The late states, after 30 ns, seem to be equilibrated and it is unlikely that a further transition will occur. For the same setup where the freely simulated subunit was taken from the r state, the projection onto the Qv vector does not show a significant difference to the free simulations. Comparing the restricted simulations with the free subunit started from t and from r a significant difference can be noted. A much faster transition from the T to the R state along the Qv vector can be found in the simulation when all four subunits started from the r state. This set of simulations moves within the first 10 ns from the T to the R state and remains there.

It seems that a tertiary restriction of three of the subunits leads to a significant restriction of the quaternary motion. This restriction can be overcome by one single subunit starting from the tertiary r state. Apparently in the case the entire T to R transition is enabled by the subunit starting from r. When looking at the simulations where all the subunits are started from the r state, the very fast transition from the T to the R state is to be noticed.
All these findings together hint at a strong coupling of the quaternary motion to the tertiary motion. The question from the beginning if the coupling goes beyond a correlation can at least be confirmed in the sense that the tertiary motion can alter the quaternary motion. Apparently one subunit moved to the tertiary t state suffices to overcome the restriction by three of the subunits being kept in the t state.

The projections on the Tv vector are displayed in figure 4.5. When comparing the free simulations with the simulations having three subunits in the tertiary t state no significant difference can be found in any of the different subunits. For the simulations where the restriction was applied on the T state of the quaternary Qv



(a) For subunit 1 = α_1

(b) For subunit 2 = β_1

(c) For subunit 3 = α_2

(d) For subunit 4 = β_2

Figure 4.5.: Tertiary projection of the ED simulations The progression of the different subunits along the difference vector predicted by PLS (Tv) are shown. In black the free simulations are shown. The red curve demonstrates the behavior of a free subunit if the other three subunits are restricted to the tertiary t state. Simulations displayed in green have the same setup but the free subunit initially starts from the tertiary r state. Blue curves show free simulations where all subunits are started from a tertiary r state. In the yellow colored trajectories the protein was retained in the quaternary T state.

vector no significant effect could be identified either. The two sets of simulations where the subunits started in the r states do not show a significant difference to each other nor do they show any transition. Both of them stay within the region defined as the r state.

The finding that the tertiary motion can alter the quaternary motion significantly, might be in accord with the finding of Olsen et al.[93](information only from secondary sources) who found that the tertiary motion is needed to stabilize the quaternary state. In contrast it is rather surprising that we could not identify a significant influence of the quaternary motion on the tertiary motion nor could we find a modification of the tertiary motion by the tertiary motion of the other subunits. This is in contrast to the tertiary two-state (TTS) model[94] that suggests the quaternary motion as a key player for the tertiary transition.
However, it can not be excluded that the a different tertiary motion is effected either by one of the other tertiary motions or by the quaternary motion. This problem arises from the limitation to only one collective coordinate given by the tertiary PLS vector (Tv) for the subunits.

4.2. Extension to Backbone - Sidechain coupling

In previous studies chain-like networks in proteins connecting the active side to distant sites were identified[95]. If these networks exist, the conformation in the residues forming that network has to be changed due to the transition of the protein. Therefore it appears to be a reasonable approach to try to directly identify the conformational change on the level of involved residues. This conformational change would be correlated to a functional property that is in our case given by the quaternary transition. To identify the most correlated change due to the transition of the states in quaternary space the previously described PLS analysis appears to be a suitable approach. The change in the backbone of the protein is mainly the transition from one state to the other. Apart from direct interactions the information of the conformation of the backbone can indirectly be passed on by the sidechain conformation. If the information would only travel along the backbone it would have to travel a long way from the reactive center (in Hb being the Heme group) to the intersubunit surface. A much shorter pathway can be found taking into account sidechain interactions. Therefore it is likely that the sidechain - sidechain commu-

nication plays a crucial role in the transmission of information from the reactive center to the intersubunit interface.
To identify if such pathways exist the trajectories were disentangled into a backbone-only and sidechain-only motion. These can be seen once more as a disassembly into orthogonal motions as previously used for the tertiary - quaternary separation. As the quaternary motion forms a subspace of the backbone motion the PCA projection, previously used to describe the T to R transition in hemoglobin, it can be used once more as the transition vector describing the transition in a one dimensional way and representing the backbone space. The information discarded in this analysis is the backbone rearrangement within a subunit. This is done on purpose as it was taken into account before when studying the quaternary - tertiary coupling.

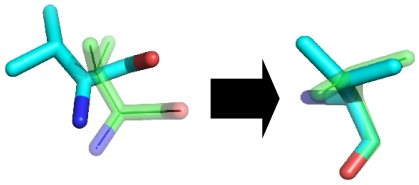
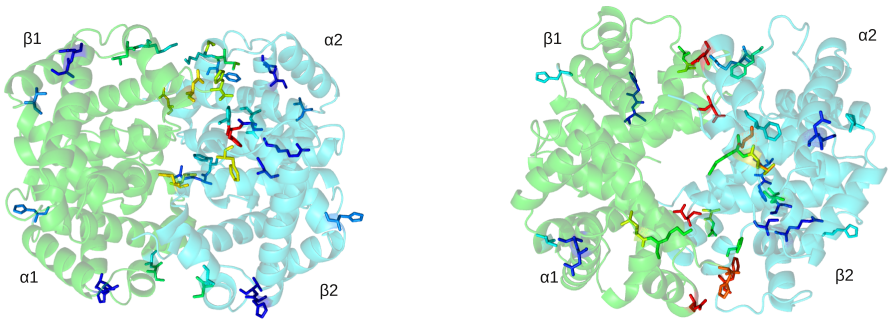


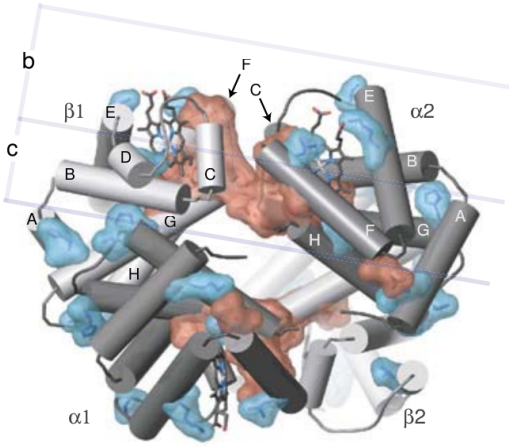
Figure 4.6.: Backbone fitting The backbone is fitted to the backbone of the reference structure. Therefore the sidechain motion only can be analyzed by the PLS tool.

To be able to get the sidechain-only motion the residues are written to individual files. On these individual trajectories a fitting of the backbone to the original starting conformation is performed. This is used to remove the backbone motion from the overall motion of the residue. What remains is the sidechain rearrangement. To address the question if the sidechain motion is induced by the conformational shift of Hb the PLS analysis was run on the trajectories containing the sidechain motion only. Half of the trajectories were used for building the PLS model while a second half was used to calculate crossvalidation values. To extract a value giving the correlation to the conformational transition, the PLS analysis was run with different numbers of PLS vectors. The value for the number of PLS components that gave the best result was taken to represent the correlation of the residue to the conformational transition. The number of PLS components describing the transition best were typically between 2 and 5 components.
To visualize the values found the crossvalidation values were projected as b-factors onto the structure of hemoglobin (see figure 4.7). To be able to read the projection, residues with a crossvalidation value larger then 0.1 were represented in a stick representation. In fig 4.7a one of the sidechains (see table 4.1) is dominating the



(a) All sidechains with crossvalidation > 0.1 shown.

(b) The sidechain with the highest crossvalidation is set to 0.



(c) Reference data from Suel et al.[55].

Figure 4.7.: PLS crossvalidation for single sidechains The crossvalidation values for the single sidechains are projected onto the structure of Hb. The red color corresponds to high values while the blue values correspond to low values. All sidechains not shown are below the cut off value of 0.1 for the cross validation.
After removing the sidechain with the highest crossvalidation is removed all sidechains with a high value are on the interface of the 1 and 2 subunits. The data can be compared to the findings from Suel et al. represented by c.

Subunit	Residuetype	Residuenumber	Cross validation
α_2	PHE	36	0.51
α_1	ASP	94	0.43
α_2	PHE	43	0.42
β_2	GLY	83	0.41
β_2	LYS	82	0.41
β_1	HIS	146	0.41
β_2	LEU	81	0.40

Table 4.1.: Residues with highest cross validation The table of residues correspond to figure 4.7. The 7 residues highlighted red are identified here and given together with their corresponding values for the cross validation.

color scheme. Therefore a second representation is given where the b-factor of this residue was set to 0 (see figure 4.7b).
As can be seen the majority of the contributing residues can be found on the interface between the **1** and **2** subunit. Only a single residue with a high contribution is rather far away in the α_2 subunit. Other mildly contributing sidechains can be found at other places far from the **1/2** interface as well.

These findings are fitting the general view on key residues for allosteric pathways found in the literature. Different studies showed that the important residues which stabilize the T state can be found at the $\alpha_1\beta_2$ and $\alpha_2\beta_1$ tetramerization interface[96–98]. These residues rearrange, leading to the elimination of key energetic interactions which subsequently allow the relaxation of Hb to the R state[96, 99].
The finding of a single residue, not being on the interface but being majorly correlated with the transition, in the α_2 subunit appears strange on first sight. But a similar finding was reported before by Süel et al.[55]. Even if the residue is not the same as found in their study, it appears to be in the same region of the α_2 subunit. Another residue which was found previously to play a role in allosteric pathways is the ASP94 of the α subunit. Here we only found ASP94 to contribute in the α_1 subunit but out of symmetry reasons a contribution in both α subunits is more likely. The other subunits contributing majorly could not be identified in a literature research.

This method seems to give a result that is in the case of Hb in good agreement with the literature. Nevertheless the results should be considered with care as no

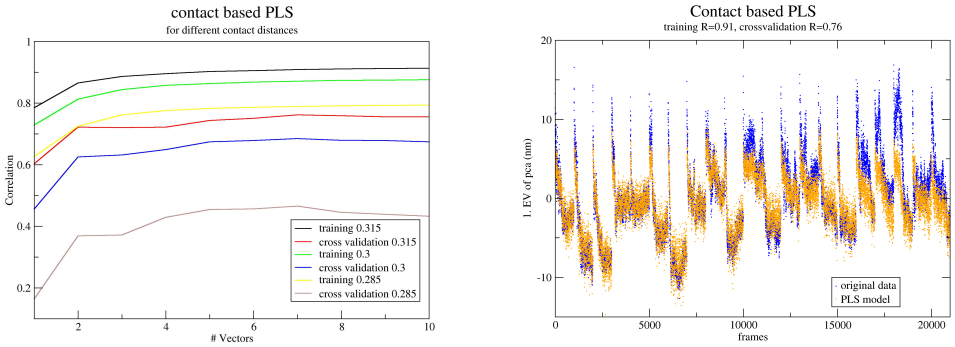
errorbars were calculated so far. The only justification for the method comes from the agreement with previous results. Even if it is unlikely that out of roughly 580 residues seven are identified to be majorly important which are in agreement with previously found results in other studies or at least at hot spots, it could still be coincidence.
To quantify this further we estimate that at each interface we find roughly 20 residues. Therefore by having two important interfaces we end up with 40 interface residues which are possibly important. Not taking into account the residue which probably corresponds to the residue found by Süel et al., we found six majorly important residues at the two interfaces. To make a guess on the likelihood of our observation being a chance result we calculate the probability of finding 6 residues to come from these 40 residues at the interface. The overall number of residues from which we could choose is 574. Thus we have:

$$p = \prod_{i=0}^{5} \frac{40 - i}{574 - i} = 7.93 \cdot 10^{-8} \tag{4.1}$$

Therefore, we can almost exclude that the residues we found are coincidentally at the interface.

4.3. Contact based PLS

In previous work the breaking of contacts, salt bridges or hydrogen bonds where found to enable the allosteric transition[100–102], while others found the formation of contacts on interfaces of a multimeric protein to be correlated to the quaternary rearrangement[103]. To identify the formation or breaking of contacts in correlation to the functional property of the quaternary transition (Qv) PLS was applied. Here we try to identify the contacts which represent a single state (either T or R) to be able to distinguish between R and T by the contact information only.
A modified version of the g_contacts[104] tool was used to extract contacts from the trajectories. For this analysis only contacts on the interfaces between the different subunits were taken into account. The list of contacts was converted into a matrix consisting only of 0 and 1 for a contact being either absent or present in a frame. To find a fitting interatomic distance to define a contact a set of PLS analysis were performed with 3 different distances. We defined a distance cutoff to be 0.285, 0.300 and 0.315 nm. The results obtained for the correlation in the original data set with

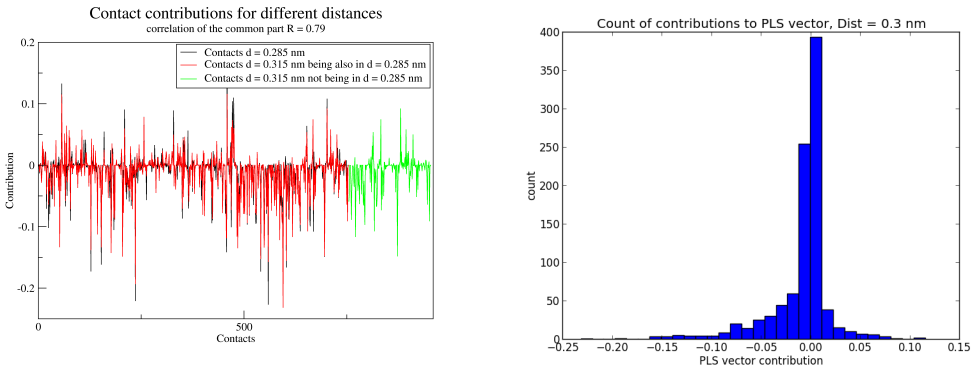


(a) Evaluation of the number of PLS components **(b) Projection of the PLS model**

Figure 4.8.: PLS contact analysis The PLS contact analysis tool is used to identify a suitable cutoff distance as well as the best number of components (subfigure a). To see how well the best combination of number of components and cutoff distance (#components=7, cutoff=0.315 nm) works, a projection of the model (orange points) together with the original data (blue points) is given in subfigure b.

the model built by PLS is shown in figure 4.8a separately for the training set, consisting of the first half of the frames and the crossvalidation set, the second half of the frames. The model is plotted in orange on top of the original data in blue. The PLS analysis gives the vector that is used to recalculate the functional property. This PLS vector is given in figure 4.9a. The black curve gives the contribution of the contacts for the 0.285 nm analysis. In red the same contacts as already found for 0.285 nm are plotted for 0.315 nm. For 0.315 nm more contacts are found to break or form than with the smaller distances. These additional contact changes, not being present for 0.285 nm, are displayed in green. In figure 4.9b the contributions distribution is shown for the 0.3 nm analysis.

A small contribution (i.e. around 0) demonstrates the lack of importance for these residues. They do not help to distinguish between the different states. Contacts having a negative contribution are predominantly present in the R state. In contrast contacts with a positive value break at the T→R transition. As can be seen from figure 4.9b the majority of the contacts do not contribute to the PLS model. Only some of the contacts show a major contribution. When defining a larger distance not only residues with a minor contribution are added. This leads to the difference of the PLS crossvalidation values between the 0.285 nm cutoff distance and the 0.315 nm cutoff distance. For the shared contacts the correlation between



(a) Independent contributions **(b) Histogram of contributions**

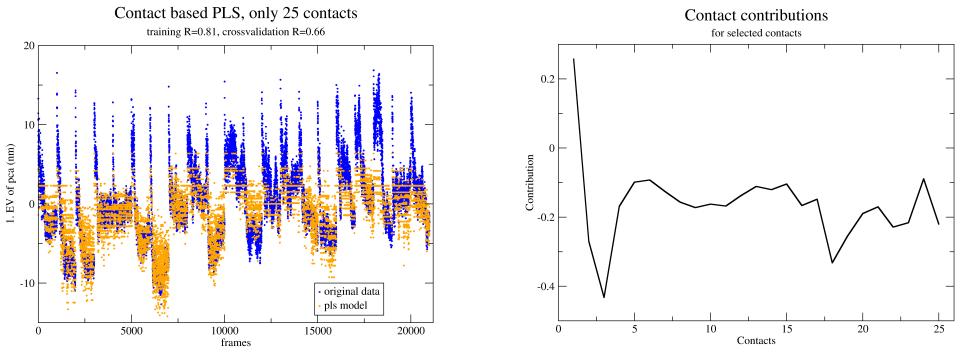
Figure 4.9.: Contributions of contacts (a) When performing the PLS contact analysis the contribution of the different contacts to the states is calculated. The contribution to the contacts for the distance of 0.285 nm are given in black while the same contacts for 0.315 nm are given in red and the contacts not being present in the 0.285 nm analysis are shown in green. (b) The contributions to the 0.3 nm analysis is plotted as a histogram. The majority of the contacts have show a contribution around 0.

the two distances is 0.79 which shows that a contact identified as important at the smaller distance is likely to be also identified at the larger cutoff distance.
To verify the importance of the contacts with a high contribution we defined a cut-off at an absolute contribution value of 0.1. In the initial matrix we eliminated all contacts below this threshold. With only 25 contacts that remained we ran the same PLS analysis again.
The results of the PLS contact analysis (see figure 4.10a) show, that the 25 contacts with an absolute contribution above 0.1 still result in a model with a crossvalidation value of 0.66. To map the result to the identified contacts table 4.2 is provided. It gives the residues involved in the contact as well as the contribution to the PLS model originally derived from the whole set of contacts.

As can be seen from table 4.2 the majority of the contacts identified as important can be found at the intersection of the multimer (i. e. α_1/α_2 or β_1/β_2). How to interpret this finding is unclear.
To analyze the effect on of the contacts further, the contact with the highest positive and negative contribution were chosen and the presence of the contact was calculated. The result (see figure 4.11) shows that the negative values are apparently only present in clearly R state like structures while the transitions not progressing

Nr.	residue 1	residue 2	contribution
1	α_1 HIS103	β_1 GLN131	0.111
2	α_1 SER131	α_2 ARG141	-0.118
3	α_1 ARG141	α_2 LEU2	-0.159
4	α_1 THR38	β_2 HIS97	-0.158
5	α_1 TYR42	β_2 TRP37	-0.119
6	β_1 TYR145	α_2 THR39	-0.103
7	β_1 TYR145	α_2 LYS40	-0.103
8	β_1 VAL1	β_2 LYS82	-0.120
9	β_1 VAL1	β_2 ASN139	-0.133
10	β_1 VAL1	β_2 HIS143	-0.156
11	β_1 LEU81	β_2 ASN139	-0.138
12	β_1 LYS82	β_2 ASN80	-0.139
13	β_1 LYS82	β_2 LEU81	-0.150
14	β_1 THR84	β_2 ASP79	-0.108
15	β_1 LYS132	β_2 HIS146	-0.126
16	β_1 ASN139	β_2 LEU81	-0.118
17	β_1 ASN139	β_2 ASN139	-0.113
18	β_1 ASN139	β_2 TYR145	-0.184
19	β_1 ALA142	β_2 ASN139	-0.110
20	β_1 HIS143	β_2 VAL1	-0.117
21	β_1 HIS143	β_2 ASP79	-0.108
22	β_1 HIS143	β_2 HIS143	-0.120
23	α_2 ARG31	β_2 GLN131	-0.129
24	α_2 HIS103	β_2 TYR35	-0.104
25	α_2 ASP126	β_2 TYR35	-0.144

Table 4.2.: Contacts identified by PLS contact analysis The 25 contacts identified by the PLS contact analysis and later used for further analysis are listed here. Given are the two residues involved in the contact as well as the contribution to the PLS vector. A positive value represents the breaking of a bond on the T→R transition while a negative value represents the formation of a bond. For this analysis the cutoff for a bond was set to 0.315 nm.



(a) PLS model with 25 contacts **(b) Contribution of top 25 contacts only**

Figure 4.10.: PLS contact analysis after 0.1 cutoff After defining an absolute cutoff value of 0.1 the PLS analysis was run again on 25 contacts only. The results show that these 25 contacts suffice to get a model that is close to the original data.

far enough on the quaternary transition vector do not correspond to a contact. The positive contribution vector can be seen to describe a contact in the quaternary T state. This contact appears to be present at the beginning of all trajectories. On trajectories which do not undergo the full T→R transition the contact is not found to break.

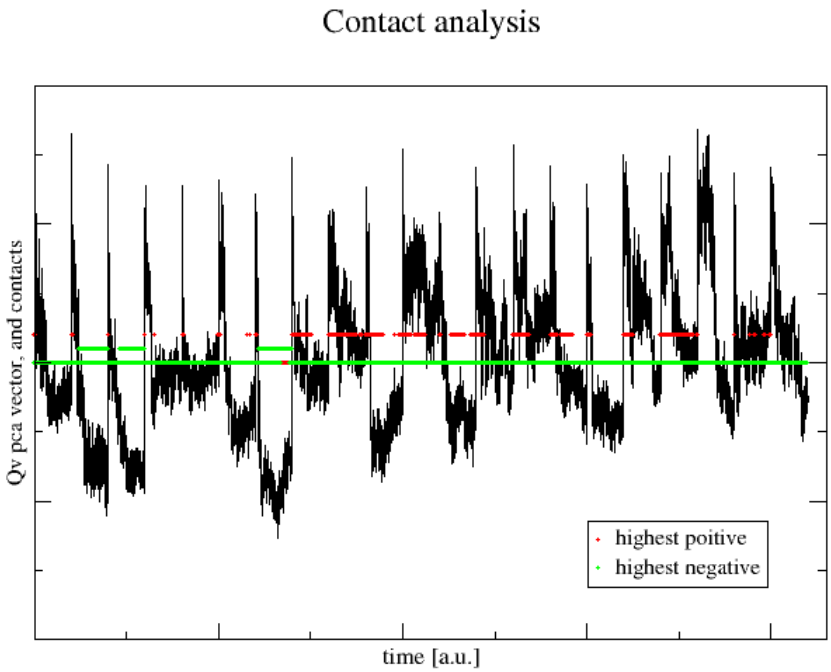


Figure 4.11.: Mapping of contacts to transition For the analysis of the role of the contacts the only contact with a positive contribution (α_1 HIS103 - β_1 GLN131) and the one with the second highest negative contribution (β_1 ASN139 - β_2 TYR145) are plotted on top of the T→R transition. As before the T state is at the top of the graph while the R state is closer to the bottom. The line in the middle shows the absence of a contact. If points are given in the sparsely populated lines above they stand for a formed contact. As can be seen, if the positively contributing contact is present the trajectory is mainly in the T state. If the negative contributing contact is present the contact is always in a R state. But not always if the trajectory is in an R state the contact is present as well.

5. Conclusions and Outlook

In this chapter we are going to review the main conclusions that can be drawn from the presented studies and the effect it has on the understanding of allostery in Hemoglobin. Furthermore we look into possible ways how to proceed with the different branches being followed.

5.1. Conclusions

Our analysis of the coupling between the tertiary and quaternary motion suggests that the most likely order of events in Hb is given by a simultaneous transition of all subunits between the states. This transition is likely to occur together with the quaternary transition as presented in figure 4.3. It is important to consider figure 4.3 as the average transition. This does not mean that in a single transition it is excluded that a sequence of events take place rather than a simultaneous transition. In contrast, we found the tertiary motion which is most correlated to the quaternary transition not to be strongly coupled between the subunits as it would be expected by a model that suggests simultaneous transitions. However that does not exclude a tertiary coupling between motions being not captured by the tertiary PLS vector (Tv). Only the quaternary motion could be identified to be indeed influenced by the respective tertiary states of the subunits. The reverse, i. e. the influence of the quaternary motion on the respective tertiary motions of the subunits, was not found. This does not answer the central question how information is transmitted from one binding site to another in Hb.
The method of backbone - sidechain coupling identified 7 residues to be majorly important in the transmission of information along the sidechains. The residues identified are in good agreement with literature results. Nevertheless, the quality of the method remains unstudied as no errors could be provided.
The PLS analysis used on the contacts is a systematic approach to study contacts in any two-state system. We managed to find a small subset of contacts which cor-

relates majorly with the quaternary transition. These contacts are mainly between the same type of subunit (α_1 - α_2 and β_1 -β_2). However, the role of these contacts in allostery remains unknown.

5.2. Outlook

The method of backbone - sidechain coupling was able to identify residues previously found in Hb. Thus it appears that the method was suitable for the studied system. The first goal should be to calculate errors for this method. Therefore an approach being close to the one used before to calculate the errors of the tertiary - quaternary coupling for the different subunits could be considered. For this purpose the half of the trajectories could be chosen to be used for model building while a single further trajectory could be used for crossvalidation. The errors resulting from that would give a hint at the reliability of the method.
If this analysis would give satisfying results (i. e. a significant difference between sidechains being considered as important and sidechains being considered as unimportant) the method is in principal not limited to the study of Hb only or even to allostery and it can also be applied to identify pathways in other proteins. The possibility to get results as good as in this case remains to be tested. It should be worth a try.
In addition the residue found to contribute but does not sit on the interface should be studied further. Therefore a restriction on that sidechain should be tested with a comparable setup as used before for the ED simulations.

With the PLS contact analysis an analysis of all contacts in the protein should be performed as well. So far only contacts on the interface were considered. Here it is of special interest if there are also contacts within the different subunits that are formed or broken at the transition. Furthermore a Monte-Carlo scheme could be applied to identify if the subset of contacts which are defined by a symmetric cutoff are really the optimal subset. We would expect an asymmetrically chosen cutoff to perform better than the symmetric one. The reason is that the asymmetry might arise from the asymmetry of the projection onto the quaternary vector. More data points do represent the R state as the T state. Most likely a scaling needs to be applied to correct for it. Another explanation for the asymmetry would be that the ratio of contacts being broken is higher than the ratio of contacts being

formed because of the rearrangement motion during the transition. However these are speculations that remain to be verified.
To study the influence of the contacts being broken or formed, an ED simulation in which a subset of the contacts are kept either formed or unformed would be a possible setup. With such a study contacts could be found to use for example cross-linking to stabilize one of the states in a protein. A possible application for such analysis could be the improvement of crystallization.

In all parts of the presented study PLS was used as a key method. Therefore a new multidimensional PLS, currently under development, could improve all of the results by including more quaternary dimensions and thus use the entire quaternary transition as the functional property.

A. Acknowledgments

First and above all I thank my supervisor Bert de Groot. Thanks for giving me the opportunity to work in this amazing and interesting field of science. I always appreciated his help, ideas and guidance.
Furthermore, I would like to thank Hadas Leonov for fruitful discussions and also for never getting tired to answer my questions. Further thanks go to the entire group for the inspiring atmosphere that contributed to my joy of working here.

Finally I thank my girlfriend as well as my parents for their support throughout the whole time of my studies. Without their backup I would not have been able to handle my studies as easily.

Bibliography

[1] R. Nussinov, C.-J. Tsai, and B. Ma. The underappreciated role of allostery in the cellular network. *Ann Rev Biophys*, 42(1):169–189, 2013. PMID: 23451894.

[2] R. Nussinov and C. J. Tsai. Allostery in disease and in drug discovery. *Cell*, 153(2):293 – 305, 2013.

[3] S. S. Taylor and A. P. Kornev. Protein kinases: Evolution of dynamic regulatory proteins. *Trends Biochem Sci*, 36(2):65–77, 2011. cited By (since 1996)110.

[4] A. W. Fenton. Allostery: an illustrated definition for the 'second secret of life'. *Trends Biochem Sci*, 33(9):420–425, September 2008.

[5] V. J. Hilser, J. O. Wrabl, and H. N. Motlagh. Structural and energetic basis of allostery. In Rees, DC, editor, *Ann Rev Biophys, Vol 41*, pages 585–609. Hilser, 2012.

[6] J. Monod, J. Wyman, and J.-P. Changeux. On the nature of allosteric transitions: A plausible model. *J Mol Biol*, 12(1):88 – 118, 1965.

[7] D. E. Koshland. Enzyme flexibility and enzyme action. *J Cell Compar Physl*, 54:245–258, 1959.

[8] D. E. Koshland, G. Némethy, and D. Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry-US*, 5(1):365–385, 1966.

[9] Q. Cui and M. Karplus. Allostery and cooperativity revisited. *Protein Sci*, 17(8):1295–1307, August 2008.

[10] J. Kister, C. Poyart, and S. J. Edelstein. Oxygen-organophosphate linkage in hemoglobin a. the double hump effect. *Biophys J*, 52(4):527–535, October 1987.

[11] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *P Natl A Sci USA*, 44(2):98–104, 1958.

[12] R. Nussinov, B. Ma, and C.-J. Tsai. Multiple conformational selection and induced fit events take place in allosteric propagation. *Biophys Chem*, 186(0):22 – 30, 2014. Special issue : conformational selection.

[13] J. E. Haber and D. E. Koshland. Relation of protein subunit interactions to molecular species observed during cooperative binding of ligands. *P Natl Acad Sci USA*, 58(5):2087–&, 1967.

[14] A. Cornish-Bowden. *Fundamentals of enzyme kinetics*. Portland Press Ltd., 1995.

[15] Y. Ikeda, N. Taniguchi, and T. Noguchi. Dominant negative role of the glutamic acid residue conserved in the pyruvate kinase m-1 isozyme in the heterotropic allosteric effect involving fructose-1,6-bisphosphate. *J Biol Chem*, 275(13):9150–9156, MAR 31 2000.

[16] B. Santamaria, A. M. Estevez, O. H. Martinez-Costa, and J. J. Aragon. Creation of an allosteric phosphofructokinase starting with a nonallosteric enzyme - the case of dictyostelium discoideum phosphofructokinase. *J Biol Chem*, 277(2):1210–1216, January 2002.

[17] X. J. Wang and R. G. Kemp. Reaction path of phosphofructo-1-kinase is altered by mutagenesis and alternative substrates. *Biochemistry-US*, 40(13):3938–3942, April 2001.

[18] W. A. Lim. The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr Opin Struc Biol*, 12(1):61–68, February 2002.

[19] C. M. Falcon and K. S. Matthews. Engineered disulfide linking the hinge regions within lactose repressor dimer increases operator affinity, decreases sequence selectivity, and alters allostery. *Biochemistry-US*, 40(51):15650–15659, December 2001.

[20] H. Frauenfelder, B. H. McMahon, R. H. Austin, K. Chu, and J. T. Groves. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *P Natl Acad Sci*, 98(5):2370–2374, 2001.

[21] A. Christopoulos. Allosteric binding sites on cell-surface receptors: novel targets for drug discovery. *Nat Rev Drug Discov*, 1(3):198–210, March 2002.

[22] S. Lazerno and N. J. M. Birdsall. Detection, quantitation, and verification of allosteric interactions of agents with labeled and unlabeled ligands at g-protein-coupled receptors - interactions of strychine and actylcholine at muscarinic receptors. *Mol Pharmacol*, 48(2):362–378, AUG 1995.

[23] J. L. Galzi and J. P. Changeux. Neurotransmitter-gated ion channels as unconventional allosteric proteins. *Curr Opin Struc Biol*, 4(4):554–565, AUG 1994.

[24] J. Ellis. Allosteric binding sites on muscarinic receptors. *Drug Develop Res*, 40(2):193–204, FEB 1997.

[25] K. Gunasekaran, B. Ma, and R. Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins*, 57(3):433–443, 2004.

[26] A. Cooper and D. T. F. Dryden. Allostery without conformational change - a plausible model. *Eur Biophys J Biophy*, 11(2):103–109, 1984.

[27] R. O. Dror, Morten Ø. Jensen, D. W. Borhani, and D. E. Shaw. Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations. *J Gen Physiol*, 135(6):555–562, 2010.

[28] A. V. Hill. Proceedings of the physiological society: January 22, 1910. *The Journal of Physiology*, 40(Suppl):i–vii, 1910.

[29] J. E. Knapp, R. Pahl, V. Šrajer, and W. E. Royer. Allosteric action in real time: Time-resolved crystallographic studies of a cooperative dimeric hemoglobin. *P Natl A Sci*, 103(20):7649–7654, 2006.

[30] G. Manley and J. P. Loria. Nmr insights into protein allostery. *Arch Biochem Biophys*, 519(2):223 – 231, 2012. Allosteric Regulation.

[31] J.-M. Jault, S. Fieulaine, S. Nessler, P. Gonzalo, A. Di Pietro, J. Deutscher, and A. Galinier. The hpr kinase from bacillus subtilis is a homo-oligomeric enzyme which exhibits strong positive cooperativity for nucleotide and fructose 1,6-bisphosphate binding. *J Biol Chem*, 275(3):1773–1780, 2000.

[32] C. K. Johnson. Calmodulin, conformational states, and calcium signaling. a single-molecule perspectiveâ . *Biochemistry-US*, 45(48):14233–14246, 2006. PMID: 17128963.

[33] G. Collier and V. Ortiz. Emerging computational approaches for the study of protein allostery. *Arch Biochem Biophys*, 538(1):6 – 15, 2013.

[34] R. Elber. Simulations of allosteric transitions. *Curr Opin Struc Biol*, 21(2):167–172, April 2011.

[35] M. D. Daily and J. J. Gray. Local motions in a benchmark of allosteric proteins. *Proteins*, 67(2):385–399, May 2007.

[36] M. D. Daily and J. J. Gray. Allosteric communication occurs via networks of tertiary and quaternary motions in proteins. *Plos Comput Biol*, 5(2):e1000293, February 2009.

[37] J. D. Bryngelson and P. G. Wolynes. Spin-glasses and the statistical-mechanics of protein folding. *P Natl Acad Sci USA*, 84(21):7524–7528, November 1987.

[38] D. U. Ferreiro, J. A. Hegler, E. A. Komives, and P. G. Wolynes. On the role of frustration in the energy landscapes of allosteric proteins. *P Natl Acad Sci USA*, 108(9):3499–3503, March 2011.

[39] W. J. Zheng, B. R. Brooks, and D. Thirumalai. Allosteric transitions in biological nanomachines are described by robust normal modes of elastic networks. *Curr Protein Pept Sc*, 10(2):128–132, April 2009.

[40] C. Y. Xu, D. Tobi, and I. Bahar. Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin t <-> r2 transition. *J Mol Biol*, 333(1):153–168, October 2003.

[41] Z. Yang, P. Majek, and I. Bahar. Allosteric transitions of supramolecular systems explored by network models: Application to chaperonin groel. *Plos Comput Biol*, 5(4):e1000360, April 2009.

[42] D. Ming, Y. F. Kong, M. A. Lambert, Z. Huang, and J. P. Ma. How to describe protein motion without amino acid sequence and atomic coordinates. *P Natl A Sci USA*, 99(13):8620–8625, June 2002.

[43] A. Thomas, M. J. Field, L. Mouawad, and D. Perahia. Analysis of the low frequency normal modes of the t-state of aspartate transcarbamylase. *J Mol Biol*, 257(5):1070 – 1087, 1996.

[44] M. Marchi and P. Ballone. Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *J Chem Phys*, 110(8):3697–3702, February 1999.

[45] H. Grubmuller, B. Heymann, and P. Tavan. Ligand binding: Molecular mechanics calculation of the streptavidin biotin rupture force. *Science*, 271(5251):997–999, February 1996.

[46] J. Leech, J. F. Prins, and J. Hermans. Smd: Visual steering of molecular dynamics for protein design. *Ieee Comput Sci Eng*, 3(4):38–45, 1996.

[47] J. Schlitter, M. Engels, P. Kruger, E. Jacoby, and A. Wollmer. Targeted molecular-dynamics simulation of conformational change - application to the t $\leftrightarrow$ r transition in insulin. *Mol Simul*, 10(2-6):291–&, 1993.

[48] J. Schlitter, M. Engels, and P. Kruger. Targeted molecular-dynamics - a new approach for searching pathways of conformational transitions. *J Mol Graphics*, 12(2):84–89, June 1994.

[49] A. van der Vaart and M. Karplus. Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. *J Chem Phys*, 122(11):114903, March 2005.

[50] H. Huang, E. Ozkirimli, and C. B. Post. Comparison of three perturbation molecular dynamics methods for modeling conformational transitions. *J Chem Theory Comput*, 5(5):1304–1314, May 2009.

[51] P. Majek and R. Elber. Milestoning without a reaction coordinate. *J Chem Theory Comput*, 6(6):1805–1817, June 2010.

[52] A. M. A. West, R. Elber, and D. Shalloway. Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide. *J Chem Phys*, 126(14):145104, April 2007.

[53] M. D. Vesper. *Collective Dynamics in Allosteric Transitions: A Molecular Dynamics Study*. PhD thesis, Georg-August University School of Science (GAUSS), 2012.

[54] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North. Structure of haemoglobin -3 -dimensional fourier synthesis at 5.5 å resolution, obtained by x-ray analysis. *Nature*, 185(4711):416–422, 1960.

[55] G. M. Suel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1):59–69, January 2003.

[56] M. F. Colombo, D. C. Rau, and V. A. Parsegian. Protein solvation in allosteric regulation - a water effect on hemoglobin. *Science*, 256(5057):655–659, May 1992.

[57] M. M. Silva, P. H. Rogers, and A. Arnone. A third quaternary structure of human hemoglobin-a at 1.7-å resolution. *J Biol Chem*, 267(24):17248–17256, August 1992.

[58] R. A. Friesner and V. Guallar. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (qm/mm) methods for studying enzymatic catalysis. *Annu Rev Phys Chem*, 56(1):389–427, 2005. PMID: 15796706.

[59] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C. Berendsen. *Gromacs User Manual version 4.5.4*. www.gromacs.org, 2010.

[60] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. Gromacs: Fast, flexible and free. *J Comp Chem*, 26:1701–1718, 2005.

[61] C. Kutzner, D. van der Spoel, M. Fechner, E. Lindahl, U. W. Schmitt, B. L. de Groot, and H. Grubmuller. Speeding up parallel gromacs on high-latency networks. *J Comp Chem*, 28:2075–2084, 2007.

[62] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. Gromacs 4.0: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput.*, 4:435–447, 2008.

[63] J. T. Brennecke. Inhibition of peptide aggregation. Bachelor Thesis.

[64] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *J Comput Chem*, 18(12):1463–1472, 1997.

[65] J.P. Ryckaert, G. Ciccotti, and H.J.C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comp Phys*, 23(3):327–341, 1977.

[66] T. Darden, D. York, and L. Pedersen. Particle mesh ewald: An n·log(n) method for ewald sums in large systems. *J Chem Phys*, 98:10089, 1993.

[67] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins*, 78(8):1950–1958, 2010.

[68] P. H. Nguyen, M. S. Li, and P. Derreumaux. Effects of all-atom force fields on amyloid oligomerization: replica exchange molecular dynamics simulations of the a*beta*16-22 dimer and trimer. *Phys Chem Chem Phys*, 13:9778–9788, 2011.

[69] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw. Systematic validation of protein force fields against experimental data. *PLoS ONE*, 7(2):e32131, 02 2012.

[70] G. Bussi, D. Donadio, and M. Parrinello. Canonical sampling through velocity rescaling. *J Chem Phys*, 126(1):014101, January 2007.

[71] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys*, 52:7182, 1981.

[72] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv Protein Chem*, 66:27–+, 2003.

[73] A. Warshel and S. T. Russell. Calculations of electrostatic interactions in biological - systems and in solutions. *Q Rev Biophys*, 17(3):283–422, 1984.

[74] F. S. Lee, Z. T. Chu, and A. Warshel. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the polaris and enzymix programs. *J Comput Chem*, 14(2):161–185, February 1993.

[75] J. S. Hub, M. B. Kubitzki, and B. L. de Groot. Spontaneous quaternary and tertiary t-r transitions of human hemoglobin in molecular dynamics simulation. *Plos Comput Biol*, 6(5):e1000774, May 2010.

[76] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philos Mag*, 2(7-12):559–572, July 1901.

[77] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*, 24:417–441, 1933.

[78] J. S. Hub and B. L. de Groot. Detection of functional modes in protein dynamics. *Plos Comput Biol*, 5(8):e1000480, August 2009.

[79] T. Krivobokova, R. Briones, J. S. Hub, A. Munk, and B. L. de Groot. Partial least-squares functional mode analysis: Application to the membrane proteins aqp1, aqy1, and clc-ec1. *Biophys J*, 103(4):786–796, August 2012.

[80] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins*, 17(4):412–425, December 1993.

[81] H. Grubmuller. Predicting slow structural transitions in macromolecular systems - conformational flooding. *Phys Rev E*, 52(3):2893–2906, September 1995.

[82] O. E. Lange, L. V. Schafer, and H. Grubmuller. Flooding in gromacs: Accelerated barrier crossings in molecular dynamics. *J Comp Chem*, 27(14):1693–1702, November 2006.

[83] M. D. Vesper and B. L. de Groot. Collective dynamics underlying allosteric transitions in hemoglobin. *PLoS Comput Biol*, 9(9):e1003232, 09 2013.

[84] G. Fermi, M. F. Perutz, B. Shaanan, and R. Fourme. The crystal-structure of human deoxyhemoglobin at 1.74å resolution. *J Mol Biol*, 175(2):159–174, 1984.

[85] S. Y. Park, T. Yokoyama, N. Shibayama, Y. Shiro, and J. R. H. Tame. 1.25 angstrom resolution crystal structures of human haemoglobin in the oxy, deoxy and carbonmonoxy forms. *J Mol Biol*, 360(3):690–701, July 2006.

[86] W. van Gunsteren, S. Billeter, A. A. Eising, P. Hünenberger, P. Krüger, and et al. *Biomolecular Simulation: The GROMOS96 manual and user guide*, 1996.

[87] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. *Interaction Model for Water in Relation to Protein Hydration. Intermolecular Forces.* D. Reidel Publishing Company, 1981.

[88] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J Chem Phys*, 103(19):8577–8593, November 1995.

[89] S. Miyamoto and P. A. Kollman. Settle - an analytical version of the shake and rattle algorithm for rigid water models. *J Comp Chem*, 13(8):952–962, October 1992.

[90] L. Mouawad and D. Perahia. Motions in hemoglobin studied by normal mode analysis and energy minimization: Evidence for the existence of tertiary t-like, quaternary r-like intermediate structures. *J Mol Biol*, 258(2):393–410, May 1996.

[91] M. Laberge and T. Yonetani. Molecular dynamics simulations of hemoglobin a in different states and bound to dpg: Effector-linked perturbation of tertiary conformations and hba concerted dynamics. *Biophys J*, 94(7):2737–2751, April 2008.

[92] T. M. Cover and Thomas J. A. *Elements of Information Theory.* Wiley & Sons, 2006.

[93] K. W. Olsen, S. Fischer, and M. Karplus. A continuous path for the t -> r allosteric transition of hemoglobin. *Biophys J*, 78(1):394A–394A, January 2000.

[94] E. R. Henry, S. Bettati, J. Hofrichter, and W. A. Eaton. A tertiary two-state allosteric model for hemoglobin. *Biophys Chem*, 98(1-2):149–164, July 2002.

[95] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, October 1999.

[96] M. F. Perutz, A. J. Wilkinson, M. Paoli, and G. G. Dodson. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annu Rev Bioph Biom*, 27:1–34, 1998.

[97] M. F. Perutz, G. Fermi, B. Luisi, B. Shaanan, and R. C. Liddington. Stereochemistry of cooperative mechanisms in hemoglobin. *Accounts Chem Res*, 20(9):309–321, September 1987.

[98] R. Liddington, Z. Derewenda, E. Dodson, R. Hubbard, and G. Dodson. High-resolution crystal-structures and comparisons of t-state deoxyhemoglobin and 2 liganded t-state hemoglobins - t(alpha-oxy)haemoglobin and t(met)haemoglobin. *J Mol Biol*, 228(2):551–579, November 1992.

[99] M. Paoli, R. Liddington, J. Tame, A. Wilkinson, and G. Dodson. Crystal structure of t state haemoglobin with oxygen bound at all four haems. *J Mol Biol*, 256(4):775–792, March 1996.

[100] X. Fei, D. Yang, N. LaRonde-LeBlanc, and G. H. Lorimer. Crystal structure of a groel-adp complex in the relaxed allosteric state at 2.7 angstrom resolution. *P Natl A Sci USA AMERICA*, 110(32):E2958–E2966, August 2013.

[101] G. Balakrishnan, C. H. Tsai, Q. Wu, M. A. Case, A. Pevsner, G. L. McLendon, C. Ho, and T. G. Spiro. Hemoglobin site-mutants reveal dynamical role of interhelical h-bonds in the allosteric pathway: Time-resolved uv resonance raman evidence for intra-dimer coupling. *J Mol Biol*, 340(4):857–868, July 2004.

[102] T. J. Wendorff, B. H. Schmidt, P. Heslop, C. A. Austin, and J. M. Berger. The structure of dna-bound human topoisomerase ii alpha: Conformational mechanisms for coordinating inter-subunit interactions with dna cleavage. *J Mol Biol*, 424(3-4):109–124, December 2012.

[103] J. Baldwin and C. Chothia. Hemoglobin - structural-changes related to ligand-binding and its allosteric mechanism. *J Mol Biol*, 129(2):175–+, 1979.

[104] C. Blau and H. Grubmuller. g_contacts: Fast contact search in bio-molecular ensemble data. *Comput Phys Commun*, 184(12):2856–2859, December 2013.

Erklärung nach §18(8) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:

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

Göttingen, den September 26, 2014

(Julian Tim Brennecke)