Protonation patterns in reduced and oxidized forms of electron transfer proteins

Dissertation

For the award of the degree
"doctor rerum naturalium"
Division of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

submitted by

Plamen Dobrev

from

Harmanli

Göttingen 2012

Prof. Dr. Helmut Grubmüller (Reviewer)
Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry Göttingen

Prof. Dr. Marcus Müller (Reviewer)
Institute for Theoretical Physics, Georg August University Göttingen

Prof. Dr. Claudia Steinem
Institute for Organic and Biomolecular Chemistry, Georg August University Göttingen

Date of the oral examination: 08.05.2012

It is declared that the presented thesis has been written independently and with no other sources and aids than quoted.

Göttingen, 10.04.2012 ____________________
Plamen Dobrev

Contents

1 Introduction 6

2 Theory and Methods 13

2.1 Molecular dynamics 13

2.2 pK$_a$ calculations of an acid in solution 15

2.3 Free Energy calculation of deprotonation of an amino acid in MD simulation 17

2.4 pK_a calculation of ionizable groups in proteins using constant pH MD 19

2.5 Construction of the model compounds and the titratable amino acids in the protein 23

2.5.1 Residues with Carboxyl titratable group 24

2.5.2 Histidine 26

2.5.3 Tyrosine 27

2.5.4 Model compounds 27

2.6 Keeping the simulation box neutral upon protonation or deprotonation 29

2.6.1 The change of the ionic strength 29

2.6.2 Restraing the coupled water molecules and the entropy change due to the restraining potential 30

2.6.3 Distance between the titrating site and the coupled water molecule 31
2.7 Barrier potential 32
2.8 pK_a calculations 33
2.9 Principal component analysis 36
2.10 Simulation details 38
2.10.1 Integrating algorithm and time step 38
2.10.2 Force field 38
2.10.3 Periodic boundary conditions 39
2.10.4 Thermostat 40
2.11 Software 41

3 Method tests results and test systems for method accuracy estimation 42
3.1 Method tests results 42
3.1.1 The effect of the ionic strength 42
3.1.2 Entropy effect of the restraining potential 44
3.1.3 Distance between the titrating site and the coupled water molecule 46
3.2 Method accuracy estimation 49
3.2.1 Epidermal Growth Factor 49
3.2.2 Cardiotoxin V 55
3.2.3 The Gly-His-Ala-His-Gly pentapeptide 61

4 Cytochrome C 66
4.1 Structure and Function 66
4.2 Force Field of the HEME 69

4.3 Protonation state of the titratable groups of Cytochrom C in reduced and oxidized state 71
4.4 Simulation Details and Setup 72
4.5 Data Analysis and Results 74
4.6 Discussion 75

5 Conclusions **78**

Acknowledgements **84**

Bibliography **91**

Chapter 1

Introduction

Electron transfer is a key element of the bioenergetics of the respiratory and photosynthetic organisms. It allows external sources of energy like solar radiation or highly reduced inorganic compounds to be utilized by the living organism. At the same time it makes possible the full oxidation of the organic compounds, produced by the photosynthetic organisms, by the subsequent members of the food chain, thus playing a crucial role in the energy flow within ecosystems.

However, despite its great importance, due to the complexity of the protein systems responsible for the electron transfer, as well as the fast rates and the non-equilibrium effects in this process, it is very difficult to obtain experimentally all of its structural, dynamical and thermodynamic aspects. Although the electrode potentials among the redox pairs are known from experiment, the different interactions that give rise to these potentials are far from being completely characterized. In order to fully understand the nature of the strong electrostatic interactions and the large shifts of the electrostatic potentials, created by the subsequent oxidation and reduction of the electron carriers, it is important that we go in atomistic details.

Changing of the redox state of the carriers, takes place in the low dielectric environment of the protein, which in many cases is embedded in hydrophobic membrane. This leads to significant shift of the electrostatic potential, which in turn, will result in strong interaction with the ionizable groups of the protein. Free energy calculation methods developed in all atom MD simulations, offer a powerful tool for quantitative estimation of these interactions, while at the same time provide correct dynamical description of the system as well. However, for correct treatment of the electrostatic interactions in protein, one needs to know the protonation state of the titratable groups of the protein. The protonation state at a given pH will be determined by the individual pK_a of each group in its specific environment. Unfortunately, in the established force fields, used in MD simulations, the protonation state of the titratable residues is assigned at the beginning of the simulation and it is not change during its course. Thus, changes in the protonation pattern of the protein, that would occur upon electron transfer, will not be accounted in standard MD simulation. Moreover, the electron transfer events are related to generating pH shifts in the environment, and often such systems function in pH regions below or above seven. In that case the protonation state of the acidic and basic residues is ether guessed or taken from experiment. However, experimental data for microscopic pK_a determination, is very sparse. Most of the experimentally available pK_a values are obtained via NMR [39], but for large proteins, assigning the hydrogen atoms in the structure, in order to determine the protonation state, becomes very involved in NMR experiments.

For that reason, in this work we aim at using a new method, previously developed in our group [7], of a dynamic protonation explicit solvent MD, which allows for constant pH MD of titratable sites in proteins, for evaluation of the protonation effect on the electron transfer. Thus by calculating the difference of the proton occupation for each titratable site for a given redox state, the response of the protein environment on the electron tranfer event, can be evaluated.

Similar studies have been carried with other theoretical approaches and the relation between the electron transfer and pK_a shifts has been partly elucidated. The most commonly used methods for pK_a calculations rely on continuum electrostatics models such as Poisson-Boltzmann(PB) or Generalized Born(GB) [1], [2], [3] or implicit solvent molecular dynamics [4], [5], [6].

The PB models treat the protein as a low dielectric environment in which fixed charges are distributed. Then the protein is solvated in the high dielectric environment of the solvent, in which distribution of movable ions is implicitly modelled. The resulting potential of the two distributions of charges is then solved on a grid. This potential is used to calculate the free energy of a protonation or deprotonation event, which is treated as a fixed charge appearing or disappearing in the low dielectric of the protein. The first drawback of this method is that the hydrogen bonding effect is not included in the model, which can be crucial for the correct treatment of the proton and its interaction with the surrounding environment. The second drawback is that the entropy effects, that come from the dynamics of the protein and the solvent are not included in the free energy estimation of the protonation event and this might give a wrong pK_a estimation.

The GB methods rely on calculation of the Born energy, or the interaction of the protonatable group with the field it creates, which depends on the dielectric of the environment. Thus the shift of the Born energy, due to transferring the protonation event, in the low dielectric environment of the protein can be calculated. Additionally the interactions with all the other groups and partial charges are included as a separate term. This approach suffers from the same disadvantages as the one mentioned above for PB models.

In implicit solvent MD, the movement of the particles is propagated in a similar way to the one in explicit solvent MD, but the solvent molecules are not included in the simulation. Their effect is modelled, by correcting the forces acting on the atoms, by a term in the force field, corresponding to the Born energy that will come from interaction with the solvent. In this method, protonation is also treated dynamically and the dynamics of the protein is also included. The lack of solvent ensures faster sampling of the protonation space, but it neglects the changes of the entropy of the solution during protonation or deprotonation as well as the hydrogen bonding between the titration sites and the solvent.

The constant pH MD method, used in this work, is based on an approach similar to the one described above but the solvent molecules are treated explicitly. The great advantage of this implementation is that interactions with the water molecules and the ions in the solution are treated with atomistic details. Thus, significant contributions such as hydrogen bonding, salt bridge formation between ions and titratable residues as well as entropy of the solvent molecules are also included in our model.

The drawback in the explicit solvent case is that not only the protein configurational space has to be sufficiently sampled, but also the solvent one, and therefore the sampling has to be much longer. In order to see a transition from protonated to deprotonated state or vice versa, not only the protein, but also the solution has to be in a thermodynamically favourable configuration to accommodate the change of the charge. However, due to the fact that structurally, the protonation is guided mainly by the side chain orientation, if there is no large scale conformational change in the protein, the equilibration of the protonation state is relatively fast.

For all of the reasons described above, our method of constant pH MD, includes important contributions of the proton-coupled electron transfer in the model of electron transfer proteins. The dynamic protonation states and the explicit treatment of the solvent are an important prerequisite for accurate pK_a calculation, hence, correct electrostatic treatment of the electron transfer.

The system, that we selected for initial study of the response of the protein environment to the change of the redox state of the electron carriers, was Cytochrome C from *Rhodopseudomonas viridis.* Its electron carrier is a heme group, whose iron is ligated by a histidine and metionine residues from the protein. The iron can be in two redox states, and thus the heme group switches between charge 0 in reduced and +1 in oxidized state. The heme is buried in the low dielectric environment of the protein core. In this case, the change of the charge upon redox reaction, will have a significant effect on the electrostatic potential on the titratable groups. The resulting shift in their pK_a's will provide quantitative measure of the response of the protein environment on the electron transfer.

Moreover, the small size, of 107 amino acids and the fact that the titratable residues have relatively high solvent exposure, ensures fast sampling of the dynamic and protonation configuration space of the protein. This makes Cytochrome C, a good starting case, to study the protonation effect of the ionizable residues on the electron transfer, using explicit solvent constant pH MD.

In order to validate our method, we selected two prototypic proteins, Cardiotoxin V from *Naja naja atra* [40] [42] [39] and Murine Epidermal Growth Factor [9] [38] [39]. Each of the two proteins was chosen to test a particular issue: first, how accurately the force field models the dynamics and second how accurately it models the electrostatics of the protein.

The Cardiotoxin is a relatively rigid protein and its pKa's will depend mainly on the charge interactions among the titratable sites and the partial charges of the environment, without too many conformational changes that would otherwise be additional test parameter for the correct titration. The protein has only three acidic residues whose pKa's vary and deviate considerably from their solution values and thus it is a very suitable test case for evaluation of the force field ability to correctly represent the elctrostatic interactions in proteins.

The Epidermal Growth Factor is a very serious challenge for pK_a calculation, due to its large flexibility and unstructured C-terminus. The latter requires a lot of sampling with MD methods and it is difficult to evaluate with continuum electrostatics. The difficulty, when using continuum models, is due to the fact that parts of the protein are unstructured and there is no proper averaged X-ray structure, on which these models rely.

The NMR ensemble for this protein consists of 16 structures and in order to additionally enhance the sampling, all of them were used for titration.

As an additional test, we performed titration simulations of an artificial pentapeptide, which was synthesized and then experimentally titrated by the NMR Department at the Max Planck Institute for Biophysical Chemistry Göttingen, lead by Prof Christian Griesinger. The peptide consists of two terminal glycines followed by two histidines and an alanine residue in the middle. The C and N termini of the peptide are not capped and their charge can interact with the two histidine residues. This and the fact that both titratable residues are identical is a good test case for correct evaluation of a pK_a shift, due to closeness to two different interacting partners - C or N terminus.

The structure of the thesis is as follows: In Chapter 2, after a short description of MD and some background for pKa and free energy calculation techniques, the constant pH MD method is described. In the same chapter, are also provided current solutions for some methodological problems, like neutralization of the charge in the simulation box due to protonation or deprotonation, using four state model to describe residues, with two protonatable sites, the barrier potential used to keep the system in on average completely protonated or deprotonated state etc. In Chapter 3, a description of the results of the titration of the peptide and the prototypic proteins is provided, including the simulation setup and the results from methods, used to estimate the structural convergence of the titrated systems. The results and the discussion of the simulation of Cytochrome C at pH 5 in the two redox states are given in Chapter 4. Short summary and conclusions are provided in Chapter 5.

Chapter 2

Theory and Methods

2.1 Molecular dynamics

The atomistic details of many biological processes provide deeper insight in the mechanism of their work. Techniques such as X-ray crystallography and NMR are a powerful tools for atomic resolution structure determination of biomolecules. However, they produce limited number of molecular structures and the information of the dynamics of the systems is little. These structures however, can be used in molecular dynamics (MD) simulations, which allows studying the structural rearrangement of biological molecules as a function of time in atomistic details. The approach is based on calculating the force acting on each atom with respect to the potential energy of interaction with all other atoms:

$$m_i \frac{d^2 \boldsymbol{r}_i(t)}{dt^2} = \nabla_i V(\boldsymbol{r}_1, ..., \boldsymbol{r}_N) \tag{2.1}$$

where m_i is the mass of the i-th particle and V is the potential energy which is function of the position of all of the particles of the system $\boldsymbol{r}$. The potential V can be calculated using quantum mechanics, however due to the size of the system that will be computationally too expensive. For that reason, the potential used in MD simulations is a sum of analytical functions treating the interactions among the particles classically - Fig 2.1.

These potentials are the force field which models the interaction properties of the real atomic system. As shown on Fig 2.1, the force field includes terms that describe the properties of bonds and angles of chemically linked

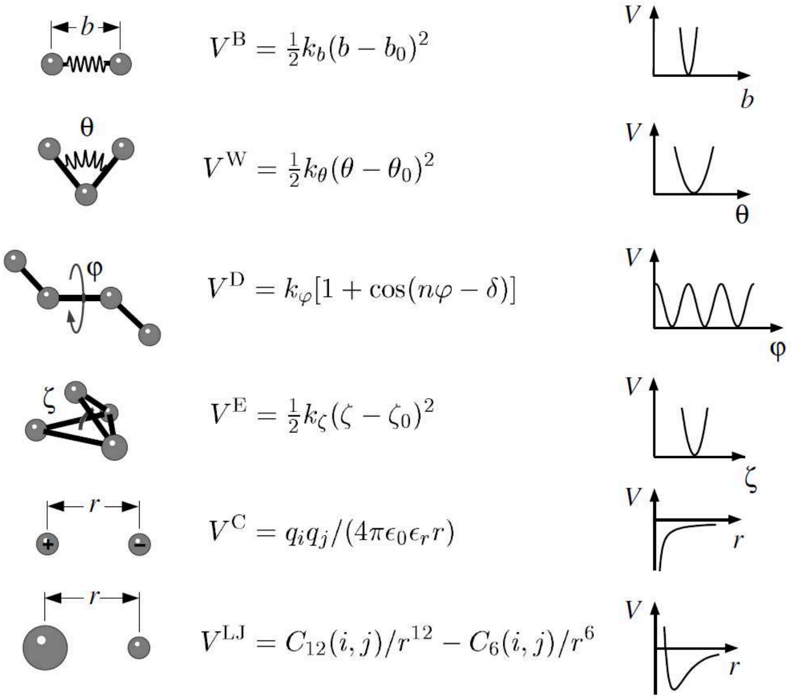

b
$V^{\mathrm{B}} = \frac{1}{2}k_b(b - b_0)^2$
V
b
θ
$V^{\mathrm{W}} = \frac{1}{2}k_\theta(\theta - \theta_0)^2$
V
θ
φ
$V^{\mathrm{D}} = k_\varphi[1 + \cos(n\varphi - \delta)]$
V
φ
ζ
$V^{\mathrm{E}} = \frac{1}{2}k_\zeta(\zeta - \zeta_0)^2$
V
ζ
r
+
−
$V^{\mathrm{C}} = q_i q_j/(4\pi\epsilon_0\epsilon_r r)$
V
r
r
$V^{\mathrm{LJ}} = C_{12}(i,j)/r^{12} - C_6(i,j)/r^6$
V
r

Figure 2.1: The four bonded and two non-bonded interactions in a force field (picture taken from ref. [13]). V^B is the potential energy of bond vibration between the atom i and the atom to which it is chemically linked, V^W is the potential energy of the vibration of the angle formed with two other atoms, V^D is the potential energy in a dihedral twisting, V^E is the potential energy due to the displacement of an aromatic carbon atom from the plane of the aromatic ring, V^C and V^{LJ} are the Coulomb and Lennard-Jones potentials between atom i and atom j.

atoms as well as non-bonded interactions like Coulomb and Lennard-Jones potentials.

All of the parameters of a force field have been optimized to represent consistently the physical properties of the biomolecular systems and there are number of force fields which are used in standard MD simulations. More detailed information on the force fields is provided later in this chapter.

Once the force acting on each atom is calculated, the Newtonian equations of motion are used to propagate their movement. However since in the case of many-body system the equations of motion can not be solved analytically, they are solved numerically with discrete time step using specific integrating algorithms (see Section simulation details).

2.2 pK$_a$ calculations of an acid in solution

The process of acid dissociation is given by:

$$AH \rightarrow A^- + H^+ \tag{2.2}$$

where AH is an acid, A$^-$ is the conjugated base and H$^+$ is the proton. In this case the equilibrium constant usually designated as K$_a$ is measure for the strength of the acid and it is defined as:

$$K_a = \frac{[A^-][H^+]}{[AH]} \tag{2.3}$$

However since the equilibrium constant can very many orders of magnitude for different acids, the negative decimal logarithm (pK_a) is often used:

$$pK_a = -\lg \frac{[A^-]}{[AH]} + pH \tag{2.4}$$

where pH is the negative decimal logarithm of the concentration of the protons and it is measure for the acidity of the environment. This formulation is known as the Henderson-Hasselbalch equation and is used to calculate the ratio of protonated and deprotonated species for a given pH.

Eq 2.4 is correct only if there is one type of acid in solution. However in protein, ionizable amino acid residues interact with each other and often their behaviour is described more accurately by Hill equation:

$$\lg \frac{[A^-]}{[AH]} = n(pK_a - pH) \tag{2.5}$$

where n is the Hill coefficient and it is measure of the cooperativity of the system: $n < 1$ indicating negative cooperativity and $n > 1$ positive cooperativity.

The equilibrium constant, of a chemical reaction, depends on its free energy. The relationship between the standard free energy of deprotonation and the pK_a of the acid, is given by:

$$\Delta G^o = -2.3RTpK_a \tag{2.6}$$

the 2.3 prefactor is due to the convention of the natural to decimal logarithms. Then from 2.4 and 2.6, the free energy difference of the system being protonated or deprotonated at a given pH is given by:

$$\Delta G = -2.3RTpK_a + 2.3RTpH \tag{2.7}$$

and for the pK_a, one obtains:

$$pK_a = -\frac{\Delta G}{2.3RT} + pH \tag{2.8}$$

As it can be seen the pK_a, or the measure for the affinity of the conjugated base for the proton, is given by the free energy of deprotonation for the given pH and the pH at which the process takes place. In order to obtain the pK_a, one must calculate the free energy of deprotonation. This free energy can be calculated using methods developed to work with MD simulations. A short description of the most frequently used, and particularly the ones used in this work, is provided in the next subsection

2.3 Free Energy calculation of deprotonation of an amino acid in MD simulation

As mentioned in the previous subsection, the free energy of deprotonation is at the base of calculating the ratio of protonated and deprotonated form of an acid at given pH. MD simulations are a tool to generate ensemble of configurations of the system of deprotonating acid in water which are used with statistical methods to calculate the free energy of the process.

For that purpose, statistical thermodynamics offers powerful tools for theoretical calculation of absolute free energy of a system with given parameters (pressure, volume, temperature and given number of particles), which are based on the properties of the ensemble of configurations of particles when the system is in the given conditions. However, in practise, due to the large size of the configuration space, the absolute free energy can not be calculated in MD generated ensembles and only a relative free energy is calculated. Biologically most relevant ensemble is the one generated with constant pressure P, temperature T and number of particles N. The free energy of such an ensemble is called the Gibbs free energy and in its absolute value is given by:

$$G = -kT \ln Z \tag{2.9}$$

where Z is the partition function of the ensemble and is defined as:

$$Z = \int \exp\left(\frac{-H(\boldsymbol{r},\boldsymbol{p}) - pV}{kT}\right) \mathrm{d}(\boldsymbol{r},\boldsymbol{p},V) \tag{2.10}$$

where H is the Hamiltonian of the system and represent its total energy, which is determined by the postions of all the particles $\boldsymbol{r}$ (the potential energy) and their momentum $\boldsymbol{p}$ (the kinetic energy). Since the volume is not constant $p\,\mathrm{d}V$ is the volume work upon expansion of the system with volume $\mathrm{d}V$ but as long as the temperature and the density of the particles are kept fixed, the volume work would average out in time. From Eq 2.9, the free energy difference between the two different Hamiltonians of the system i.e. protonated or deprotonated state is given as:

$$\begin{aligned} \Delta G &= kT \ln Z_B - kT \ln Z_A && (2.11) \\ &= kT \ln \frac{Z_B}{Z_A} && (2.12) \end{aligned}$$

where Z_A is the partition function in protonated state A and Z_B is the partition function in the deprotonated state B. However as already mentioned, since the integrals in the partition functions run over the full configuration space of the system they can not be calculated in practice and therefore the Zwanzig formula [20] is usually used where only the Boltzmann weighted difference between the two Hamiltonians is taken and the relative free energy between the two states is calculated:

$$\Delta G = -kT \ln \left\langle \exp\left(\frac{H_B(\boldsymbol{r},\boldsymbol{p}) - H_A(\boldsymbol{r},\boldsymbol{p})}{kT}\right)\right\rangle_A \tag{2.13}$$

This method is known as Free Energy Perturbation and its main disadvantage is its slow convergence, due to the fact that the difference between the two Hamiltonians is in exponent and high energy states, which have big contribution to the overall partitioning, will be sampled very badly due to their low probability. In an alternative approach not the averaged difference between the weighted Hamiltonians is calculated but rather the average of the derivative of a blending Hamiltonian H:

$$H = (1-\lambda)H_A + \lambda H_B \tag{2.14}$$

where λ is the reaction coordinate going from 0 to 1, is calculated with respect to λ . The derivative is then integrated over the reaction coordinate to obtain the free energy along it:

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\mathrm{d}H}{\mathrm{d}\lambda} \right\rangle_{\lambda} \mathrm{d}\lambda \tag{2.15}$$

This method is called Thermodynamic Integration (TI) [21] and it can be done slowly with gradual change of λ parameter from 0 to 1 in every MD step also known as Slow Growth Thermodynamic Integration or it can be done with discrete steps in which λ is kept fixed:

$$\Delta G = \sum_{i=0}^{N} \left\langle \frac{\mathrm{d}H}{\mathrm{d}\lambda} \right\rangle_{\lambda_i} \Delta\lambda \tag{2.16}$$

which is known as Discrete Thermodynamic Integration. The main problem of the slow growth TI is that the system is never in equilibrium due to the constant change of λ. That is why for better convergence, the discrete thermodynamic integration is used provide there are enough λ points taken between the two states. [25].

2.4 pK_a calculation of ionizable groups in proteins using constant pH MD

Due to their spacial proximity in protein, charged amino acid residues interact with each other and their pK_a values inside the protein might be very different from those in solution. The protonation state of all titratable amino acids determines the spatial distribution of the electrostatic potential in the protein, which is crucial for ligand binding, protein-protein interaction and protein folding. Moreover for residues located in the active site of enzymes, the correct protonation state is essential in order the catalytic act to take place.

Unfortunately in standard MD simulations, the protonation state is assigned at the beginning of the simulation and change of the protonation that would occur upon conformational change or interaction with other ionizable groups is not accounted.

If one wants to calculate the protonation state of a given group it is necessary that the interactions with all other groups are taken into account. Due to the fact that in protein all groups are in spacial proximity and they interact with each other, their free energy of deprotonation hence their pK_a, will depend on the pK_a of all other groups, they interact with. So for the free energy of deprotonation of a given group i one writes:

$$\Delta G_i = -RT \ln \frac{\langle x_i \rangle}{1 - \langle x_i \rangle} \tag{2.17}$$

where $\langle x_i \rangle$ is the partial occupancy of the proton of that group and is defined as:

$$\langle x_i \rangle = \frac{\sum_{n=1}^{2^N} x_i^n \exp(-G_n/RT)}{\sum_{n=1}^{2^N} \exp(-G_n/RT)} \tag{2.18}$$

were N is the number of groups, $\langle x_i^n \rangle$ is 1 or 0 depending if group i is protonated or deprotonated in configuration n and G_n is the free energy of state n. It is clear that calculation of deprotonation free energy with such an approach would give combinatorial explosion and therefore the method we used in this work [7], treats every reaction coordinate of every titratable group as virtual path between two states - protonated and deprotontaed, along which a virtual particle with mass m and speed v can move. Then the force acting on that particle is given by:

$$F = -\frac{\mathrm{d}H(\lambda)}{\mathrm{d}\lambda} \tag{2.19}$$

where H is the Hamiltonian from Eq 2.14 and λ is the reaction coordinate. If one considers n reaction coordinates, one writes:

$$F_{\lambda_i} = -\frac{\partial H(\lambda)}{\partial \lambda_i} \tag{2.20}$$

Thus, all of the λ particles are treated as normal particle in MD simulation and their movement is propagated simultaneously. This approach circumvents the necessity for combinatorial calculation of the free energy of deprotonation of all titratable groups. Furthermore, the method allows dynamic treatment of the protonation state and a given group can switch from protonated to deprotonated state depending on its environment during simulation.

The restriction to travel between 0 and 1 is achieved by introducing special circular coordinate on which a virtual θ particle is moving. The force acting on that particle is the one that is actually calculated and its movement is then projected onto the λ coordinate Fig.2.2 [7] according to:

$$\lambda = 0.5 + r\cos(\theta) \tag{2.21}$$

To prevent the λ particle from staying too long in the non-physical intermediate states of the $0 \div 1$ interval, an energetic barrier is applied centered at $\lambda = 0.5$ which forces the λ particle to go to fully protonated or fully deprotonated state. The properties and height of the barrier are discussed in greater details later in this chapter.

A particular issue of the dynamic protonation approach described so far is the fact that free energy of a chemical reaction, such as deprotonation can not be calculated in MD simulation. The reason is that in the established force fields used today, there is no term that account for formation or breakage of a chemical bond, which happens during titration of a group. So to calculate the free energy of such a process in protein, one needs an estimation

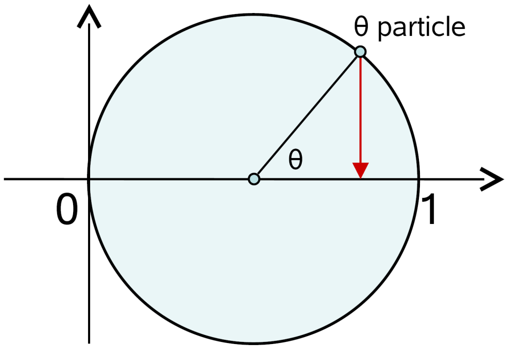


Figure 2.2: Projection of the dynamics of the θ particle on the lambda coordinate. Picture taken from ref. [7].

of the error introduced by the force field. For that reason, the solution experimental value for the free energy of deprotonation (or the pK_a) of a model compound, similar to the amino acid residue that will be titrated, is required. The solution deprotonation free energy for the same model compound is then calculated in MD simulation. The difference between the experimental and simulation value will give the error introduced by the force field - Fig.2.3. Assuming the same error will be made in protein, one can correct for it and get the right value of the free energy of deprotonation. Thus:

$$\Delta\Delta G = \Delta G^{sim}_{model} - \Delta G^{exp}_{model} \tag{2.22}$$

where $\Delta\Delta G$ is the error in the free energy calculation introduced by the force field and it is then added to $\Delta G^{sim}_{protein}$ to correct for the free energy of deprotonation in the protein.

In the present work, the experimental pK_a values for the titratable residues in solution, will be referred as *reference* pK_a's. Their values are known and are discussed in greater details in Section 2.5.

All of the simulations, performed to obtain the free energy of deproto-

$$\begin{tikzcd}
Prot - AH \arrow[r, "\Delta G^{sim}_{protein}"] \arrow[d] & Prot - A^{-} + H^{+} \arrow[d] \\
AH \arrow[r, "\Delta G^{sim}_{model}"] \arrow[d] & A^{-} + H^{+} \arrow[d] \\
AH \arrow[r, "\Delta G^{exp}_{model}"] & A^{-} + H^{+}
\end{tikzcd}$$

Figure 2.3: The thermodynamic cycle describing the error correction for the force field

nation of the model compound in solution, for the force field error correction, will be referred in this work as *parametrization simulations* or *calibrations*.

The contribution to the free energy from the pH, at which one wants to simulates, is also added according to Eq. 2.7. Then the main contribution for the behaviour of the group at the given *pH*, comes from the electrostatics of the protein which is presumably modelled correctly by the force field.

2.5 Construction of the model compounds and the titratable amino acids in the protein

There are three types of residues, modelled in this thesis. First group includes acidic residues with carboxyl group - aspartate, glutamte and the propionic side chains of the heme group. In the second is the histidine residue, having imidazole ring as a functional group and in the third is the tyrosine residue with a phenolic functional group.

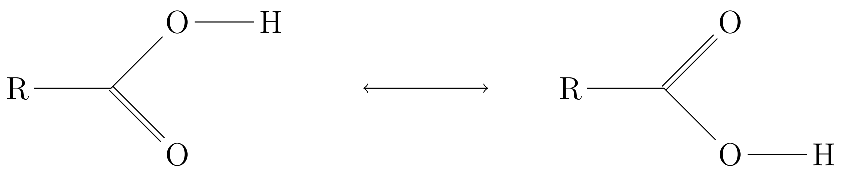


Figure 2.4: Isomerization of the carboxyl group

2.5.1 Residues with Carboxyl titratable group

As shown in Fig 2.4, the carboxyl group have two oxygen atoms, each of which can be protonated. However, in the established force fields the proton is fixed to one of the oxygens and if the interactions between the functional group and the protein environment would be more suitable for the other oxygen to protonate, one may never see this in the MD simulation. This will lead to shift of the free energy of deprotonation and incorrect pK_a estimation. To solve this problem, we introduced a new chimeric "carboxyl" group, which has two protonated oxygens but only one of the hydrogen atoms carries a charge. Moreover, besides the protonation/deprotonation coordinate, we also introduced a second reaction coordinate, which switches the charge of the hydrogen from one oxygen to the other as depicted in Fig 2.5. Thus each titratable site, containing a carboxyl group, is treated with a four state model, which gives a closer representation to the behaviour of the real chemical compound.

The estimation of the force filed error for the switching coordinate is done in the same way as for the protonation one. However, since now there are two oxygen atoms, that can undergo protonation, the reference pK_a for the model compound should also change to give the right correction for each separate oxygen. If K_{a_1} and K_{a_2} are the two microscopic equilibrium constants, which is the ratio of protonated/deprotonated species for each oxygen, then the macroscopic equilibrium constant for the whole group is

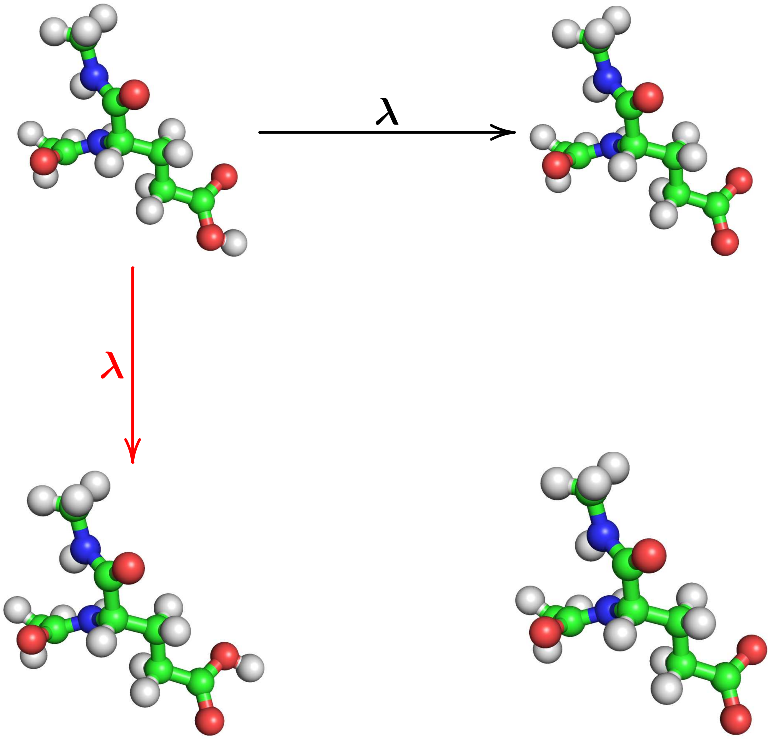


Figure 2.5: The four state model of residue with carboxyl group. The titrating coordinate is depicted in black arrow and the switching one - in red

given as:

$$pK_a = \lg(10^{pK_{a_1}} + 10^{pK_{a_2}}) \tag{2.23}$$

In case where both microscopic pK_a's are the same, as in the case of a carboxyl group, one writes:

$$pK_a = pK_{a_{1,2}} + \lg 2 \tag{2.24}$$

During titration or switching between the two oxygens, only the charge of the atoms of the group change. All of the bonded and Lennard-Jones parameters are kept unchanged.

In this work, we used 3.95 as reference macroscopic pK_a value for aspartate residues and 4.4 for glutamate and the propionic sidechains of the heme [23] [24].

2.5.2 Histidine

Histidine has an imidazole wring as a functional group which has two nitrogen atoms and again both of them can be protonated Fig.2.6.

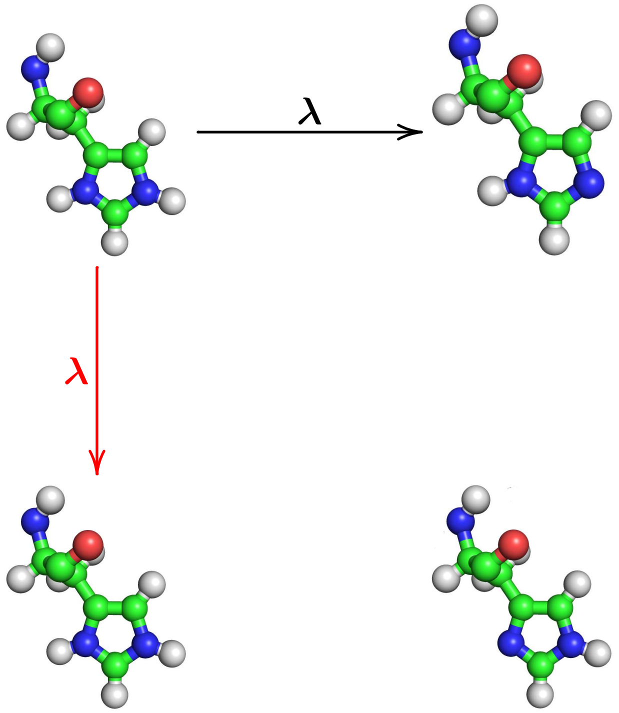


Figure 2.6: The four state model of histidine. The titrating coordinate is depicted in black arrow and the switching one - in red

The scheme of the model is similar to the one used for carboxyl groups, again with two coordinates - switching and titrating. However, since the link to the backbone is at carbon which is closer to one of the nitrogens, their microscopic pK_a's are different. The relation between the microscopic pK_a's of the two nitrogen atoms and the macroscopic pK_a of the whole group is given as:

$$pK_a = -\lg(10^{-pK_{a_1}} + 10^{-pK_{a_2}}) \tag{2.25}$$

where pK_{a_1} and pK_{a_2} are the two microscopic pK_a's of the two nitrogens. Analogous to the carboxyl group, again neither the bond nor the Lennard-Jones parameters are changed during the simulation.

The macroscopic reference pK_a value used for the histidine residue was 6.38 [22].

2.5.3 Tyrosine

The tyrosine residues have a phenolic -OH group which can deprotonate. Since there is only single oxygen atom, on which the proton is located, a two state model was used. The used reference pK_a value was 9.6 [23]

2.5.4 Model compounds

As described in Section 2.4, in order to perform constant pH MD simulations, one needs to calculate the fee energy of deprotonation in solution, of a model compound, similar to the amino acid which will be titrated. This value will be compared with the reference value for the amino acid and the error introduced by the force field will be obtained.

The model compound should be as similar to an amino acid residue in protein as possible. However, if the amino acid that will be titrated is used for the free energy of deprotonation in solution, it will have negative and positive charge for its C and N termini respectively. That configuration is very different from the one in protein, where the amino acid's termini are bound in peptide bonds with the neighbouring amino acids.

In all of the proteins simulated in this work, there were no titratable amino acids at the C on N terminus that would have COO^- or NH_3^+ charged groups. Therefore the model compounds, used to parametrize the ionizable groups of all proteins, are the same amino acids as the one in protein, but with acetyl and methylamine caps for the N and C-terminus respectively. Thus the compound will be neutral, and will model best the residues linked to the polypeptide chain.

For both - the model compounds and the amino acids in the protein, the charge of atoms of the backbone was not changed. This was done, because of the different scaling of the electrostatic interactions in the force field, for atoms that are chemically linked, up to the fourth neighbour. If the backbone atoms are also included in parametrization, then their electrostatic interactions with the adjacent amino acids will be modelled wrong in protein due to the scaling of the Coulomb potentials. It is very difficult to obtain quantitative measurement of this effect and estimate its contribution to the free energy of deprotonation. Therefore the charges of the backbone atoms were kept unchanged.

Unfortunately that approach can not be applied for the model compound of the propionic side chains of the heme group. In that case, butyric acid was used as a model compound, but the charges were fit in such a way, so that there is no change in the end -CH_3 group upon deprotonation. Its carbon atom in protein will be the link to the heme ring, and for that reason its charge was kept unchanged and with the same value as in protein. Even so, for some of the aliphatic atoms of the side chains, the Coulomb interactions will be screened. However, the largest shift of charges upon deprotonation is that of the carboxyl group atoms. The distance between them and the iron and nitrogen atoms, where most significant part of the charge of heme molecule is localized, is larger then four bonds. In that case, the interactions between them are treated without scaling. Thus the Coulomb interactions between the propionic side chain and the heme group will be modelled reasonably well.

2.6 Keeping the simulation box neutral upon protonation or deprotonation

2.6.1 The change of the ionic strength

Due to change of the charge of ionizable groups during constant pH MD simulation the box, in which it is running, becomes charged. In the case of periodic boundary conditions, in which all of the simulations were performed (See Section 2.10.3), that would lead to endless summation of the potential of the charge through the periodic images. As a result endless forces would be generated on the charged atoms. If such a simulation setup is made in standard MD simulations, the contribution from the charge is artificially corrected and the charge of the box is set back to zero. However this approach gives secondary artefacts and its effect on the system is not completely understood. To solve the problem, we coupled each of the titration sites with a water molecule that can be charged in direction opposite to the one of the amino acid it is coupled to. If an ionizable group charges negatively, the water molecule becomes positively charged.

However since the neutral water molecule becomes charged, that changes the ionic strength of the solution. The free energy of an ion will be different in solutions with different concentrations and the concentration in the simulation box will vary upon ionization of the titratable residues.

Quantitatively this effect will be determined by the activity coefficients of the positive and negative ions in the solution:

$$a = \gamma_{\pm} c \tag{2.26}$$

where a is the activity, c is the concentration of the ion species in the solution and γ is the activity coefficient of the positive (γ_+) or negative (γ_-) ions, which is a measure for the deviation of the solution from the ideality. The activity coefficient depends mainly on the ionic strength of the solution and its contribution to the free energy of the ions in the solution is given as:

$$G = G_{ideal} + RT \ln \gamma_+ \gamma_- \tag{2.27}$$

where G_{ideal} is the free energy of the ions in solution with such a high dilution that the γ coefficient equals 1 and there is no difference between activity and concentration.

Thus change of the ionic strength would lead to change of the activity coefficients and the free energy of deprotonation. Unfortunately, this dependence can be estimated from general considerations only for very diluted solutions. For that reason, the effect of different ion concentration on the free energy of deprotonation was systematically tested and the results are provided in Section 3.1.1.

2.6.2 Restraing the coupled water molecules and the entropy change due to the restraining potential

If the coupled water molecules come too close to the protein or to each other, then the free energy of deprotonation of the residue they are coupled to will differ from the one calculated in the calibration simulation. In order to assure that the coupled water molecules and the protein atoms have enough screening solution among them, they were placed as far as possible from each other, and their movement was restrained with harmonic potential.

However, if the volume, accessible for the waters, is different in the calibration simulations and in the constant pH MD simulations, that will gives rise to different entropy contributions in both cases. The available volume for the water molecules in harmonic restraining potential is given by:

$$V = \sqrt{(2\pi)^3 \sigma_1 \sigma_2 \sigma_3} \tag{2.28}$$

where the σ's are the eigenvalues of the three by three covariance matrix of the spacial distribution of the restrained water molecule. Then the contribution to the free energy coming from the volume confinement is given by:

$$\Delta G = \frac{RT}{2} \ln \left(\frac{\sigma_{2;1} \sigma_{2;2} \sigma_{2;3}}{\sigma_{1;1} \sigma_{1;2} \sigma_{1;3}} \right) \tag{2.29}$$

where the first lower indices indicate the different simulations with different spring constants of the restraining harmonic potential and the second indices indicate the three different eigenvalues.

The entropy effect of the restraining potential was tested in a number of simulations, and the results are provided in Section 3.1.2

2.6.3 Distance between the titrating site and the coupled water molecule

Since the simulation box needs to be kept neutral during parametrization as well as during the constant pH MD, the amino acid residue that was parametrized or titrated was in both case coupled to a water molecule. In all of the parametrization simulations the two partners were placed as far as possible to diminish their interaction. However, in constant pH MD simulation with a protein, there will be as many coupled water molecules as titrating sites. Then the distance between the site and the coupled water molecules will be different for every group, hence different from the one in the calibration simulations. To estimate the effect of the distance between the titrating site and the coupled molecule on the free energy of deprotonation, we performed number of free energy simulations. The results are provided in Section 3.1.3.

2.7 Barrier potential

If the lambda particle stays too long in intermediate states, that will result in wrong average charge of the atoms of the titrating amino acids, hence wrong dynamics and wrong behaviour of the protein. In order to prevent that, a repulsive barrier potential is applied with its highest point at $\lambda = 0.5$ and its lowest points toward the two ends of the lambda interval.

The form of the potential should be such, that it doesn't perturb the two Hamiltonians at the end states and thus preserves the free energy difference between them. The form of the potential used for all of the protein simulation in the current thesis is shown on Fig 2.7.

As it can be seen, the potential is applied only for states with $\lambda >0.1$ and $\lambda <0.9$. Thus, there is no energy contribution to the force field at the two ends of the lambda coordinate, which are used for the pK_a calculation (see next Section).

Unfortunately due to its form, in some cases there might be stabilization in states close to $\lambda =0.1$ or $\lambda =0.9$. Additional effort is still made to further adjust the width and the zero potential intervals in order to solve this problem.

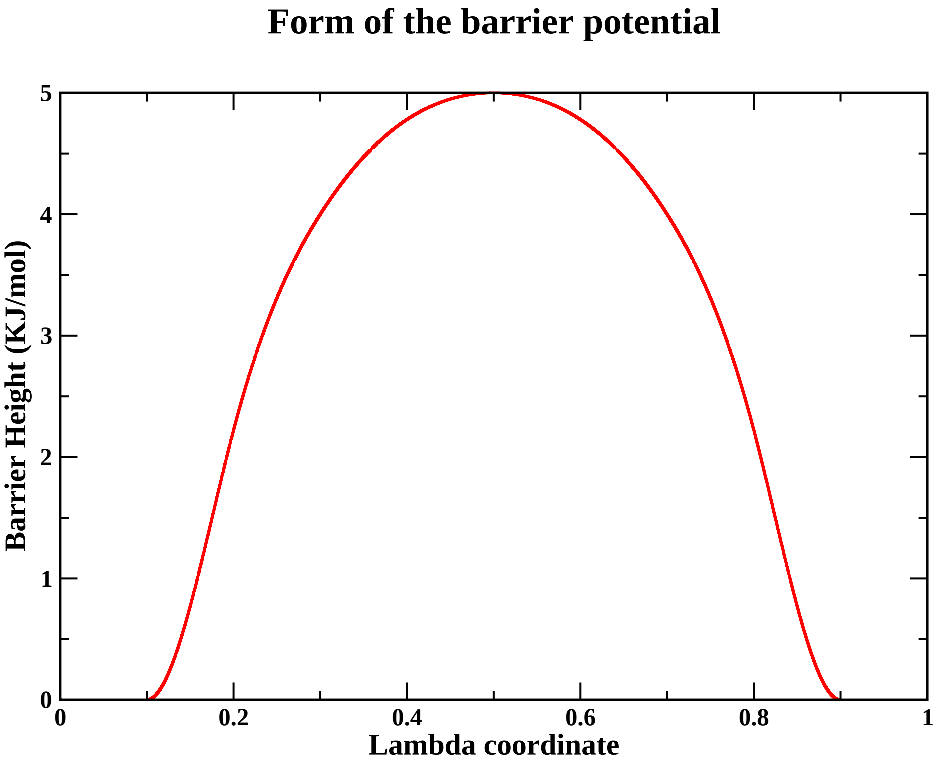


Figure 2.7: The form of the barrier potential

2.8 pK_a calculations

To obtain the pK_a of a given group in protein one needs to perform a number of simulations at different pH points. From the simulations, the trajectory of the lambda particle is obtained (see Fig 2.8). State 0 corresponds to protonated and 1 to deprotonated amino acid and only the time lambda particle spends below 0.1 and above 0.9 were considered for the pK_a calculation. As a criterion for what is the minimal acceptable time that should be spent by a titrating group at the end states, we considered 70 % or above to be the time in which a group should be at states with lambda coordinate larger then 0.9 or smaller then 0.1. Based on the number of transition and the time spent at either protonated or deprotonated state, one obtains the average value of

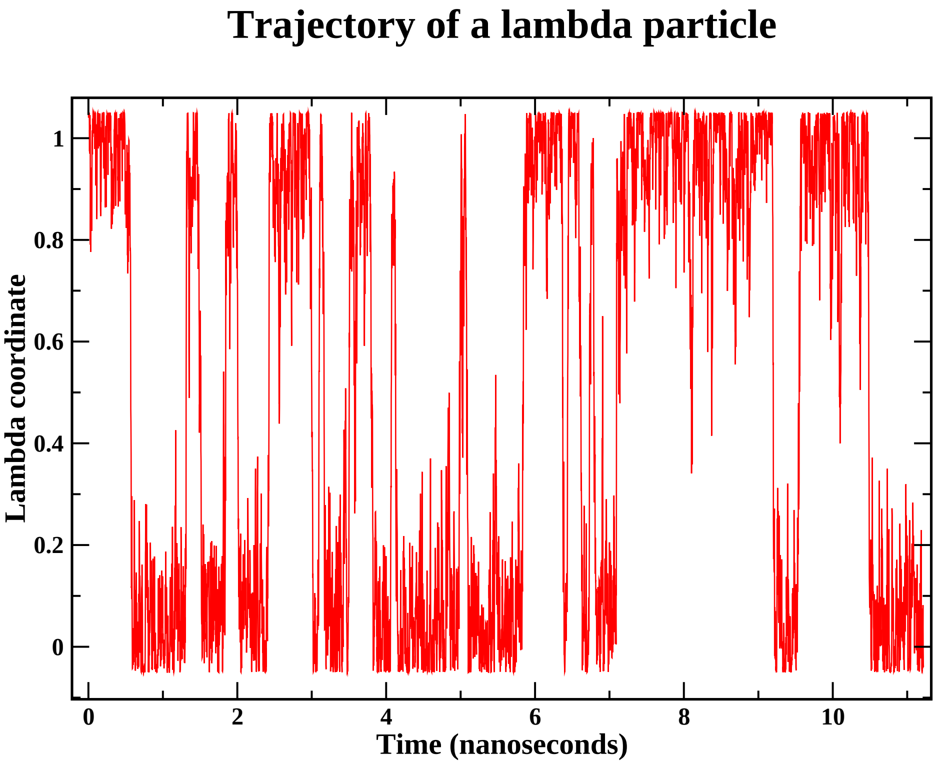


Figure 2.8: Trajectory of a lambda particle during simulation

lambda for that particular pH and the error of the estimation.

The method used for that purpose is described in [7]. It makes use of Poisson fit of the distribution of the rate constants of the two processes - protonation and deprotonation (Fig 2.9), to calculate the probability distribution of lambda. However, since from simulation one obtains only the number of transitions for the given time and given rate constant, the Bays Theorem was used to inverse probability conditions and obtain the probability of the rate constants given certain number of transitions in the time interval. Thus the probability of a certain combination of rate constants is given by:

$$p(k_0, k_1|n_0, n_1) = \frac{T_0^{n_0+1} T_1^{n_1+1}}{n_0! n_1!} k_0^{n_0} k_1^{n_1} \exp[-T_0 k_0 - T_1 k_1] \tag{2.30}$$

where p is the conditional probability, k_0 and k_1 are the rate constants of a $0 \to 1$ and $1 \to 0$ processes respectively, n_0 is the number of deprotonation transitions and n_1 is the number of protonation transitions, T_0 and T_1 are the time spent in protonated and deprotonated state respectively.

In order to calculate $p(k_0, k_1|n_0, n_1)$, an initial guess for the rate constants, calculated as the number of transition over the time, is made. Then, the probability of the rate constants, from 0 up to 20 times larger then the original estimation, are calculated with a 1000 point resolution according to Eq. 2.30. A result of such calculation is shown on Fig. 2.9.

The average value of lambda depends on the two rate constants and its value is given as:

$$\langle \lambda \rangle = \frac{k_0}{k_0 + k_1} \tag{2.31}$$

From 2.30 and 2.31 it follows that:

$$\langle \lambda \rangle = \int_0^\infty \int_0^\infty p(k_0, k_1|n_0, n_1) \frac{k_0}{k_0 + k_1} \, \mathrm{d}k_0 \, \mathrm{d}k_1 \tag{2.32}$$

Once the probability distribution of the rate constants is calculated according to Eq. 2.30 the average lambda and its probabilty distribution is calculated according Eq. 2.31 and 2.32.

In order to obtain the errors of the lambda distribution, we integrated the area below the probability curve, and took the lambda values that lie at 2.5 and 97.5 percent of the area, to be the lower and upper bound of the error respectively. After the probability distribution of lambda and the error estimation for each pH point are known (Fig 2.10), 10000 random points are taken from the lambda distribution for every pH value and are then used to build 20000 Henderson-Hasselbalch fits. These fits are used to calculate the probability distribution of the pKa of the given group - Fig 2.10. The same criterion, used for the error estimation below the probability curve of the lambda distribution, is used to calculate the error of the pKa as well.

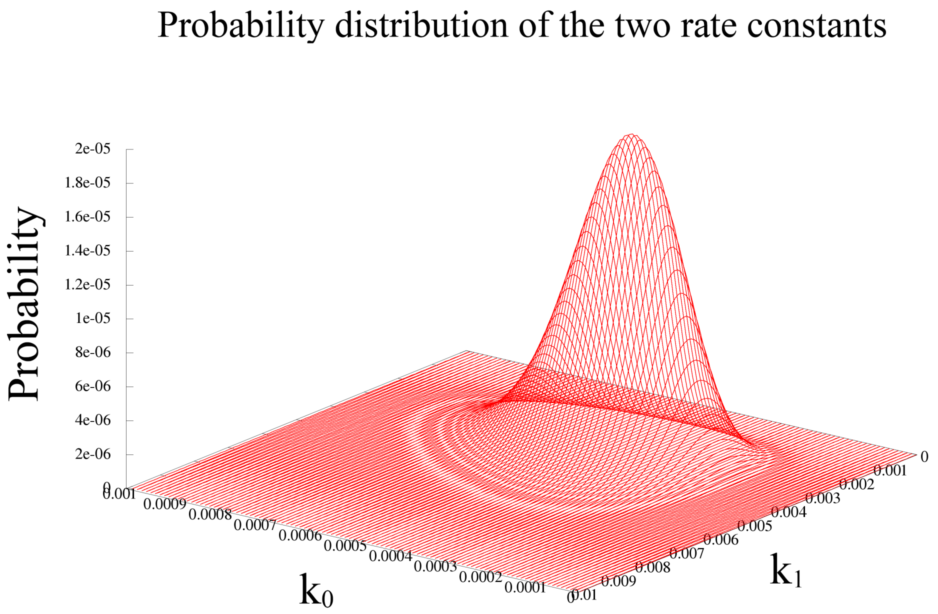


Figure 2.9: The distribution of the two rate constants. Since we only need their ratio, the units on the axes of the constants are relative as every time step in the simulation is considered as 1 and the overall times in Eq. 2.30 are sums of the number of steps.

2.9 Principal component analysis

In order to track the dynamics of the overall structure of the protein, one needs a method to reduce the dimensionality of the observed 3N-dimensional system, where N is the number of the atoms, and focus only on the large conformational changes during simulation. Such a method is the Principal Component Analysis [34][35], which rely on constructing a covariance matrix C, from the coordinates of all atom positions x, averaged over the simulation

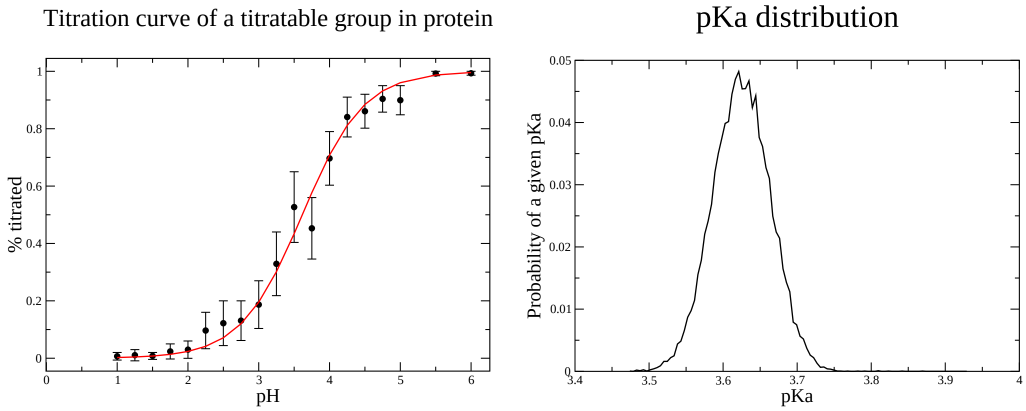


Figure 2.10: A single Henderson-Hasselbalch fit of the average lambda values and the error, based on the lambda distribution for each pH point (left) and the probability distribution of the pKa's based on 20000 fits within the error of the lambda distribution for each pH point (right)

length:

$$C = \langle (x_i - \bar{x})(x_i - \bar{x})^T \rangle \tag{2.33}$$

When the matrix is diagonalized, one obtains the collective deviation form the averaged position of the atoms in the eigenvectors of the matrix. The eigenvectors related to the largest eigenvalues describe the large-scale motion of the protein. When the trajectory of the atom movement is projected onto the eigenvectors, one obtains the different states through which the system migrates during simulation.

Similar to atom positions, protonation states of titratable residues in protein, may also change in collective manner. Therefore similar approach was used to track the overall protonation state of Cytochrome C. However, instead of atom coordinates, we uses lambda coordinates to construct the covariance matrix.

2.10 Simulation details

2.10.1 Integrating algorithm and time step

The equations of motion for a multi-body system can not be solved analytically. For that reason, the movement of the atoms in force field simulations, is calculated numerically, with a discrete time step, using different integrating algorithms. In this work, the *leap frog* algorithm [14] is used, in which the velocity v and the position r of any atom i are calculated for each time step Δt from their previous values according to:

$$v_i\left(t+\frac{\Delta t}{2}\right) = v_i\left(t-\frac{\Delta t}{2}\right) + \frac{F_i(t)}{m_i}\Delta t \tag{2.34}$$

$$r_i(t+\Delta t) = r_i(t) + v_i\left(t+\frac{\Delta t}{2}\right)\Delta t, \tag{2.35}$$

where F_i is the force acting on atom i at time t.

The size of the time step is determined by the fastest degree of freedom, which is usually the vibration of the chemical bond between the atoms. In order to use longer time steps hence access longer simulation times, the bond vibrations were removed and the lengths of the chemical bonds were fixed using the LINCS constraint algorithm [15]. Additionally, the hydrogen atoms were converted to virtual sites, which is their masses were merged to the heavy atoms they are linked to, and their positions during the simulation were geometrically calculated [19].

With that set up, we were able to use a time step of 4 femtoseconds, for all of the simulations in the present work.

2.10.2 Force field

As mentioned in the molecular dynamics subsection, there are a number of established force fields which are used in MD simulations. Some of them are empirical (OPLS, GROMOS) and are parametrized to represent correctly macroscopic observables like free energy of solvation, diffusion constants etc. Others such as the AMBER series are quantum based. Their great advantage is in their flexibility when a new compound needs to be parametrized for force field simulations.

All of the work in the current thesis is done using AMBER99sb [16] force field, with its port to GROMACS [17] [18], and the SPCE water model [28]. Its choice was motivated by the requirement new chemical compounds to be easily parametrized and used in simulation. Thus charges of molecules and model compounds that are not part of the force field, such as the hypothetical positively charged water molecule or the charges of deprotonated tyrosine residue, were calculated *ab initio* with Gaussian 03 [26]. Some of the parameters of the amino acid residues, already existing in the force field, have been altered and some new residues and atom types have been introduced to describe the four-state model of the carboxyl functional group. Furthermore, new bond parameters and atom types for the reduced and oxidized form of the heme group were added[27].

2.10.3 Periodic boundary conditions

In order to prevent introducing of surface tension, molecular dynamics simulations are performed under periodic boundary conditions i.e. the simulation box is surrounded by its mirror images in all directions. The interactions and the particles, crossing the borders are translated through the box as in a crystal lattice. However, if the water molecules coupled to the titrating sites are placed only considering the other coupled waters or the protein in the same simulation box, they will come in immediate contact with their mirror images. That is why during their original placing, not only the distances among the coupled molecules and the protein atoms are calculated, but also their periodic images are considered - Fig.2.11.

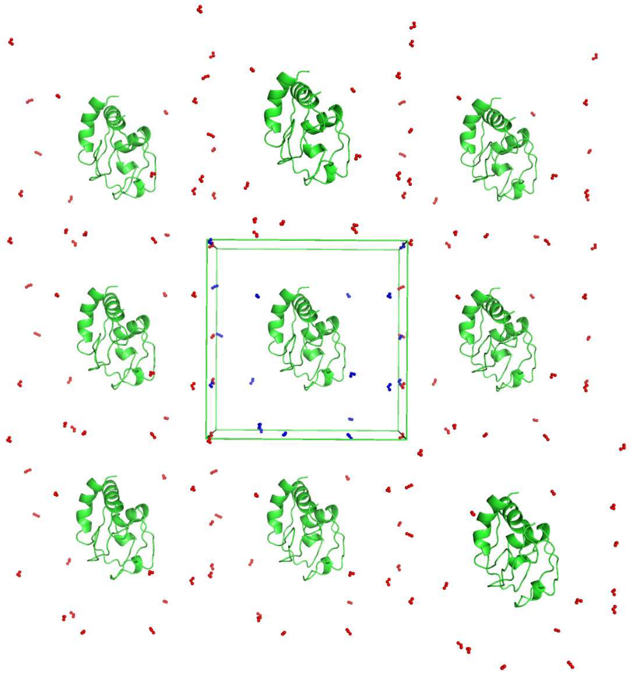


Figure 2.11: The simulation box with its periodic images, the coupled water molecules from the box are colored in blue and their periodic images in red

2.10.4 Thermostat

Only limited number of atoms can be included in a MD simulation. For that reason, the temperature of the system has to be maintained additionally with different thermostat algorithms. The thermostat used to maintain the temperature of the atoms in the simulation box in this work, was the so called Nosè-Hoover[29][30] thermostat. It is based on addition of a friction term that couples the system to a thermal bath and the equation of motion become:

$$\frac{\mathrm{d}^2\boldsymbol{r}_i}{\mathrm{d}t^2} = \frac{F_i}{m_i} - \xi\frac{\mathrm{d}\boldsymbol{r}_i}{\mathrm{d}t} \tag{2.36}$$

where ξ is a dynamic friction parameter with own equation of motion:

$$\frac{\mathrm{d}\xi}{\mathrm{d}t} = \frac{1}{Q}(T - T_0) \tag{2.37}$$

where the difference between the reference temperature T_0 and the current temperatute T is used as a potential gradient in the integrating algorithms. The Q parameter depends on the strength of the coupling to the heat bath and its components are input parameters to the simulation.

The lambda particles are also coupled to heat bath, but in order prevent heat flow from the lambda space to the real space, they are each coupled to individual heat baths. The thermostat used for their coupling is the so called Andersen thermostat[31] and it generates Maxwell-Boltzmann velocity distribution in the time, by correcting the speed of the lambda particle, using Monte-Carlo method [32][7].

2.11 Software

All of the free energy calculations, needed for parametrization or testing in this work were performed with discrete thermodynamic integration in GROMACS-3.3.3 software package [25]. The constant pH MD simulations were performed with module for constant pH MD in GROMACS-3.3.3[7]. Charges of not parametrized protonation states of the compounds were performed with Gaussian 03[26]. The PCA of the atoms was done again with the GROMACS-3.3.3 software package [25] and the one for the lambda particles - with Matlab 7.12.0.635 ©1994-2012 The MathWorks, Inc. The images of protein structures were generated with pymol[33] and edited with GIMP©. All of the graphs were generated with Xmgrace©and Gnuplot©.

Chapter 3

Method tests results and test systems for method accuracy estimation

3.1 Method tests results

3.1.1 The effect of the ionic strength

As discussed in Theory and Methods Chapter, when a water molecule charges in order to compensate for change of the charge of a titratable residue in the protein, that leads to change of the ionic strength of the solution. This shift may influence the activity coefficients of the ions, hence the free energy of deprotonation.

In order to evaluate the dependence of the free energy of deprotonation on the ionic strength, we performed a number of free energy calculations with different salt concentrations. However, since the ions also need time to equilibrate their distribution, the simulations were performed with different lengths in order to estimate the relaxation time of the system as well. The results are shown in Fig.3.1.

As it can be seen, in the longer time simulations the dependence on

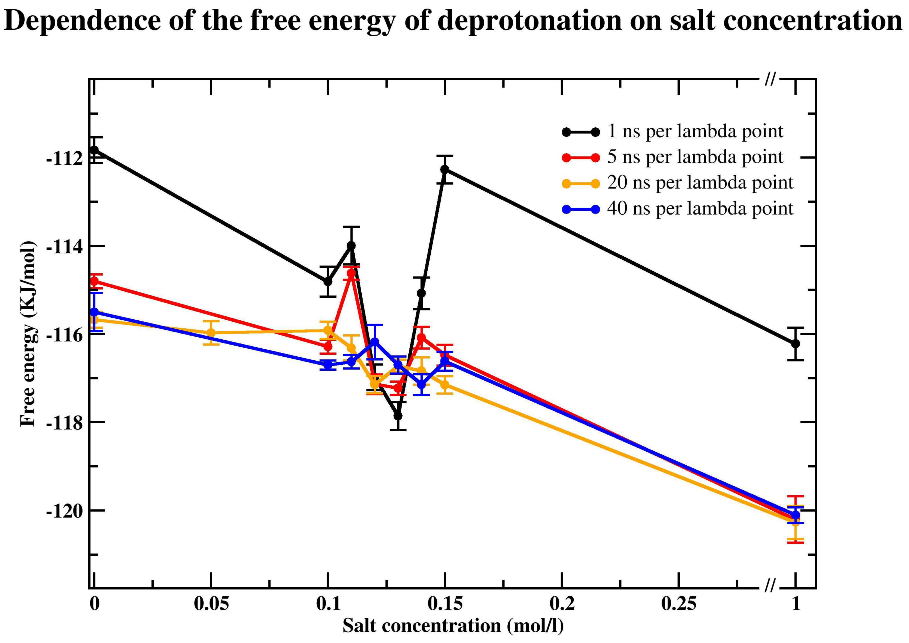

Dependence of the free energy of deprotonation on salt concentration

Figure 3.1: The dependence of free energy of deprotonation on the salt consternation. The free energy method used was discreet thermodynamic integration and the different series illustrate the different simulation lengths for each lambda point

the salt concentration is not very strong. Change of the concentration from 0.1 to 0.15 mol/l, which corresponds to change of 11 positive and 11 negative ions gives no noticeable difference in the free energy of deprotonation. Physiological concentration of 0.15 mol/l was used in all of the constant pH MD simulations and appearance of up to 6 charged molecules in the case of the Epidermal Growth Factor or 3 in the case of Cardiotoxin V will not have a significant effect on the free energy of deprotonation of the titratable groups.

To mimic fully the effect of a deprotonating group in the case of Cytochrome C, an amino acid was deprotonated in simulation where 17 Na^+ ions were fixed and the simulation was performed with 11 Na^+ and 28 Cl^- movable ions. The deprotonating amino acid was coupled to a water molecule that charges during simulation, thus brining the total positive charges in solution to 18 as would be the number of coupled water molecules in cytochrome C. The difference in the free energy between this setup and the one of having only 24 Na^+ and 24 Cl^- (0.15 mol/l for the box size) is approximately 0.5 KJ/mol. This difference is negligible and it will not have big effect on the pK_a of the ionizable groups during constant pH MD.

However, the series of simulations shown on Fig.3.1 also indicate that there is a certain time needed for ions to sample correctly the simulation box. They all started from the same configurations for the given ion concentrations and as it can be seen, most of them required more then 20 nanoseconds to equilibrate. Therefore all of the calibration simulations were done with discrete thermodynamic integration, with a 40 nanosecond sampling window per lambda point.

3.1.2 Entropy effect of the restraining potential

In order to prevent coupled water molecules from coming too close to the titrating sites they were coupled to, their motion was restrained with a three dimensional harmonic potential. However, that way, the allowed volume for the water molecule is limited. To estimate the effect of the restraining, we performed a number of free energy simulations in which we charged a water molecule, which was restrained with harmonic potentials with different spring constants. In order to confirm the proper sampling of the allowed configuration space for the water molecule, we calculated its distribution

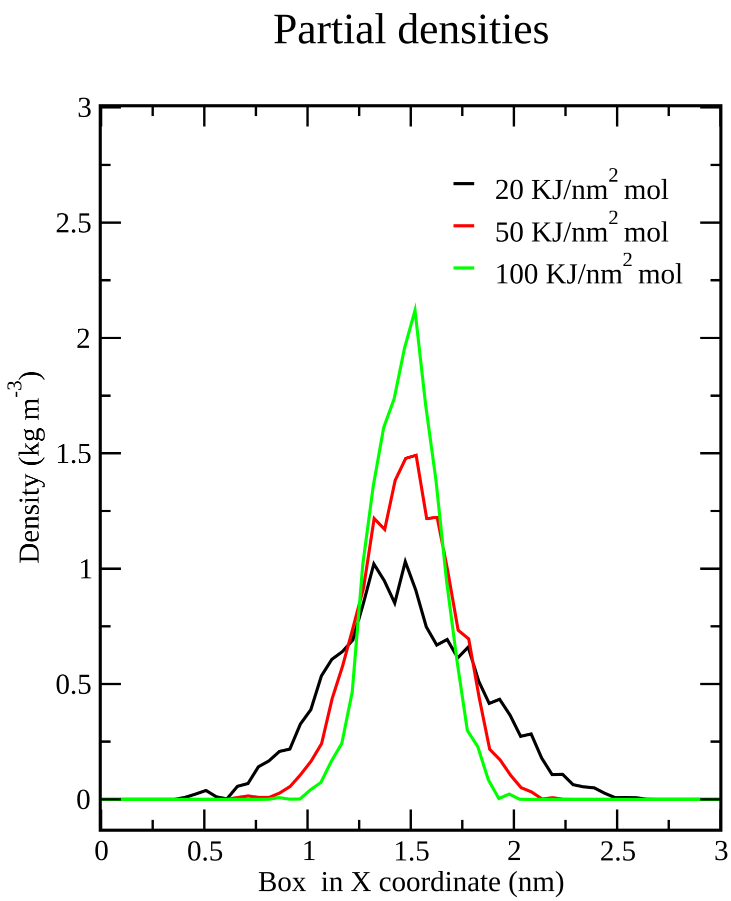


Figure 3.2: The distribution of the restrained water molecule on the X coordinate of the simulation box during the free energy simulation. Here again the discrete thermodynamic integration is used and the distribution plotted is the one of the first lambda point with protonated water molecule. The distribution is similar to the one where the water is completely deprotonated

during the free energy simulation and the results are shown in Fig.3.2 and Table 3.1.

As it can be seen, the Gaussian distribution, which is the result of the underlying harmonic restraining potential, is well converged and the water molecules have sampled their available volume properly. From Table 3.1 it

Table 3.1: The free energy of deprotonation with different spring constants

Spring constant(KJ/nm^{2}mol)	Volume(nm^3)	Free energy (KJ/mol)
20	0.2945	-322.6
50	0.1656	-322.433
100	0.0293	-322.57

can be seen that if the available volume for the water molecules is constant during the simulation, there is no difference in free energy in the simulations with different spring constants. However when in an alternative simulation (data not plotted) the spring constant was changed from 20 KJ/nm^2mol to 50 KJ/nm^2mol, without changing the charge, the free energy of that process was 3.37 ± 0.034 KJ/mol, which agrees with the theoretical value of 3.4 KJ/mol calculated according to Eq. 2.29. This shows that there will be contribution to the free energy of deprotonation or protonation only if the allowed volume for the coupled water molecules is changing during the simulation. This is another reason for restraining their movement to certain volume and position.

3.1.3 Distance between the titrating site and the coupled water molecule

Different distance during the parametrisation simulation and during the constant pH MD simulations may give rise to difference in the free energy of deprotonation in both cases. In order to estimate this effect, we performed number of free energy simulations, in which the two partners were placed at different distances. The results are summarized on Fig.3.3 and Table 3.2.

As it can bee seen the dependence of the free energy on the distance is

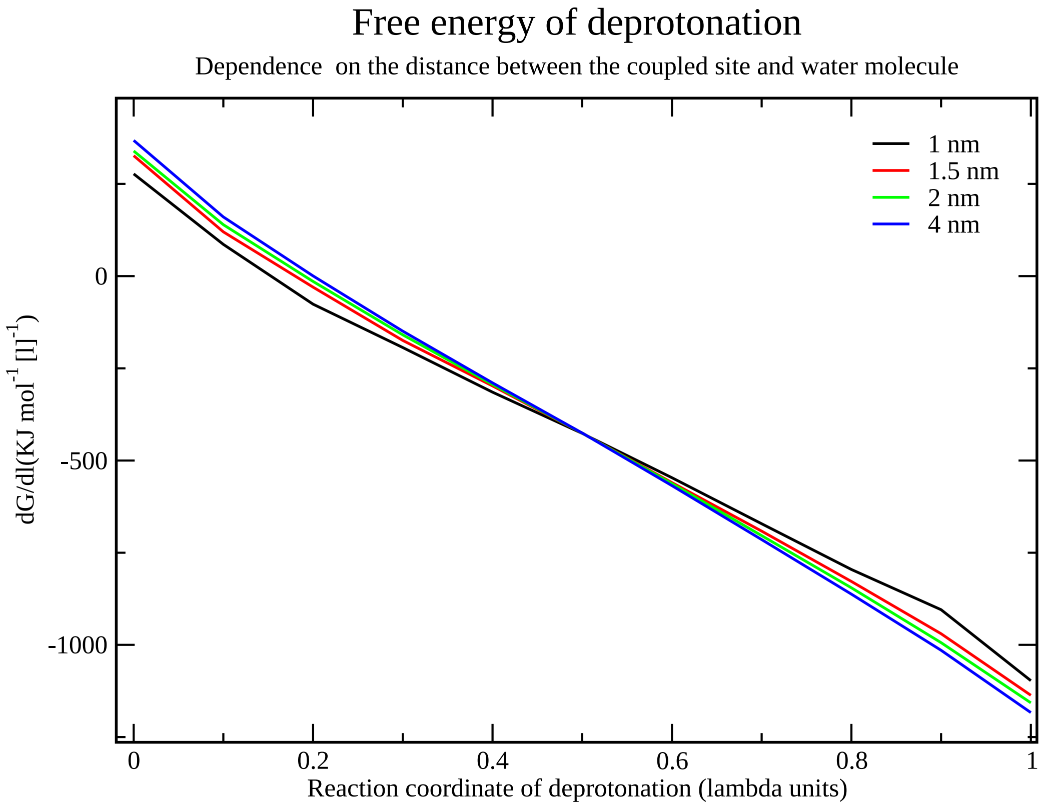


Figure 3.3: The dependence of the gradient of the free energy of deprotonation with respect to the reaction coordinate on the distance between the two coupled sites

Table 3.2: The free energy of deprotonation with different distances between the two coupled sites

Distance (nm)	Free energy (KJ/mol)
1 nm	-425.260 ± 0.148
1.5 nm	-426.178 ± 0.168
2 nm	-426.746 ± 0.149
4 nm	-427.065 ± 0.13

relatively small. Although the slope of the derivative of the free energy with respect to the reaction coordinate changes, after integration, the values of the free energy for the whole process are close. Above 1 nanometer distance they are within 1 KJ/mol.

However, the difference in the slope will lead to wrong free energy correction of the intermediate states. That is to say, if the two coupled molecules are placed closer to each other in constant pH MD than during calibration, that will create an energy minimum in the intermediate states. According to Fig.3.3, the smaller the distance the deeper the minimum. This effect can be seen on Fig.3.4 with the two extreme cases where parametrization was done at 4 nanometer distance and then two simulations were performed placing the two molecules at the same distance of 4 nanometers and then 1 nanometer at pH=pK_a. At this pH point, half of time lambda particle should spend in deprotonated state and half in protonated or in other words the average lambda value should be 0.5. Due to the preservation of the free energy when distance changes, the average value in both cases is close to 0.5 but with significant stabilization of the middle states when the two molecules are closer. Since the free energy is preserved, this problem is easily solved by the hight of the barrier potential, which in cases of stabilization of the intermediate states can be increased to obtaining again the correct distribution. The shortest distance at which coupled water molecule and any protein atom are placed in the present work was 1.6 nanometers.

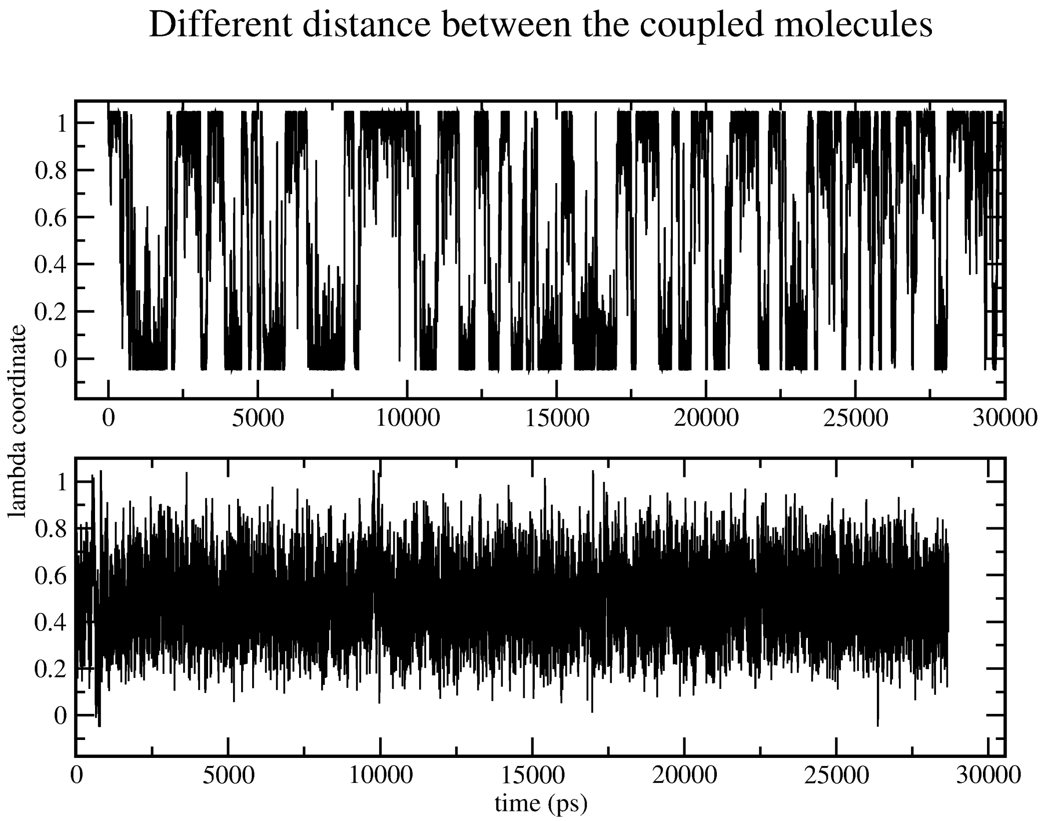


Figure 3.4: The lambda particle behaviour in two simulations with different distance between the titration site and the coupled water molecule. On the upper plot the distance is 4 nanometers, the same as in the calibration simulation and on the lower one is 1 nanometer. The lambda coordinate goes from 0 - protonated to 1 - deprotonated state. The time is in picoseconds.

3.2 Method accuracy estimation

3.2.1 Epidermal Growth Factor

Structure and Function

The Murine Epidermal Growth Factor is a small, 53 amino acid protein which act as an intercellular signaling molecule with mitogenic activity and is believed to play role in the wound healing process[36]. It is composed of three antiparallel β sheets and a hairpin loop close to the C-terminus - Fig 3.5. The overall structure is stabilized by three disulphide bridges and it has a very flexible C-terminus. The structure is solved by NMR[37] and the structural ensemble consists of sixteen members.

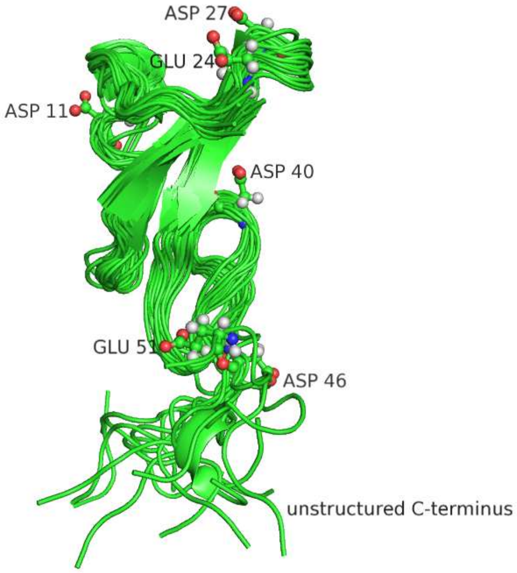


Figure 3.5: The NMR ensemble of the Epidermal Growth Factor with the positions of its acidic residues plotted on the first structure

Constant pH MD Simulation Setup

The protein has 6 acidic ionizable groups that were included in the titration - Fig 3.5. The pH region that was titrated started from pH=1.5 and ended at pH=6, with a step of 0.25 pH units at the regions close to the pK_a of the titrating groups and 0.5 pH units at the end of the interval. Each of the 16 structures was titrated as the length for each pH point simulation was approximately 7 nanoseconds. The charge of the titrating residues was neutralized by coupling them to water molecules, which charge in opposite directions (as described in Theory and Methods Chapter). The rotation and translation of the center of mass of the protein was removed to insure that protein doesn't come close to any of the coupled water molecules responsible for the neutralization of the system.

Estimation of the structural convergence

The two residues GLU 51 and ASP 46, that are located in the unstructured C-terminus, are challenge for a pK_a calculation since their structure space needs a lot of sampling. In order to improve the sampling we took all of the structures of the NMR ensemble and titrate them separately concatenating the lambda trajectories afterwords. However, since the NMR ensemble is not representative for the protein behaviour and the structures inside are not Boltzmann-weighted, we performed a PCA on the atomic positions of the C-terminus in the constant pH MD and included the NMR ensemble in the analysis.

Furthermore, due to some implementation issues, the constant pH MD module[7] is much more expensive then the regular MD. For that reason, we simulated the protein with the two titratable residues in the flexible C-terminus, with all possible combination of them being protonated and deprotonated in free MD simulations, and included them in the PCA.

Data analysis and Results

The constant pH MD simulations from all 16 structures and all pH points, have united length of 1.58 microseconds. Each free MD simulation, representing given fixed protonation state, has a length of approximately 240 nanoseconds. The free MD trajectories were projected on the eigenvectors of the covariance matrix of the united constant pH MD simulations through all pH regions and then overlaid, together with the lambda simulations themselves. The average structure used, for building the covariance matrix, was generated by fitting the more rigid part of the protein and averaging only the structure of the flexible C-terminus. After diagonalization of the matrix, the eigenvectors of the flexible region were used to project its dynamics during simulation - Fig.3.6.

The experimental pK_a values for the protein are known from NMR experiments[38] [39], and the ones we calculated theoretically are based on Henderson-Hasselbalch fitting as described in the Theory and Methods Chapter. A Henderson-Hasselbalch fit to the average occupancy of lambda at each pH point is shown on Fig. 3.7. As a control to justify using of the Henderson-Hasselbalch equation, which assumes no cooperative effect between the sites, we did a single Hill equation fit as well as Henderson-Hasselbalch fit on the averaged lambda for each pH point. The difference between the estimated pK_a's, with the two different fits, was less then 0.05 pK_a units and the cooperativity coefficient was between 0.6 and 0.7. These results suggest that the Henderson-Hasselbalch model can be used without significant loss of accuracy, and since it is computationally much more efficient, it was used for this protein. The comparison between the calculated pK_a values and the experimental ones[38][39] is given in Table 3.3

Discussion

In general, most of the groups through their titration satisfy the condition that their lambda particle be with coordinates smaller then 0.1 or larger then 0.9 more then 70 % of the time. However, due to the specific form of the

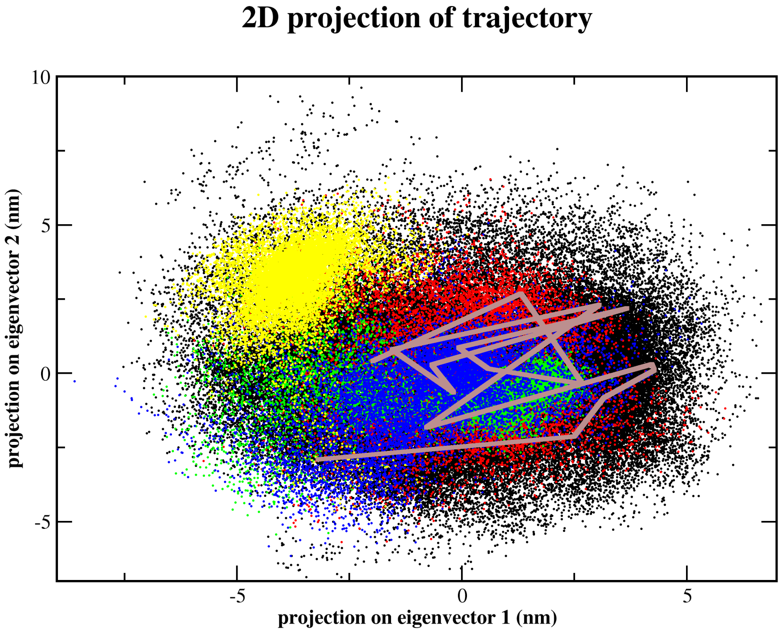


Figure 3.6: The projection of the atomic position trajectory onto the first and the second eigenvector of the covariance matrix, generated from the united trajectories of the titration of the sixteen structures from the NMR ensemble in all pH regions. With black is represented the united constant pH MD trajectory , in red is the simulation with both ASP 46 and GLU 51 deprotonated, in green is the simulation where they were both protonated, in blue - ASP 46 is deprotonated and GLU 51 protonated, in yellow ASP 46 is protonated and GLU 51 deprotonated and in brown is the projection of the NMR ensemble.

barrier some of the groups for given pH points, show stabilization of their lambda particle close to 0.1 and 0.9 as in the worse case, the average drops between 60 % and 70 %. As already mentioned above, we are currently working to solve this problem mainly via adjusting the barrier hight as well as its

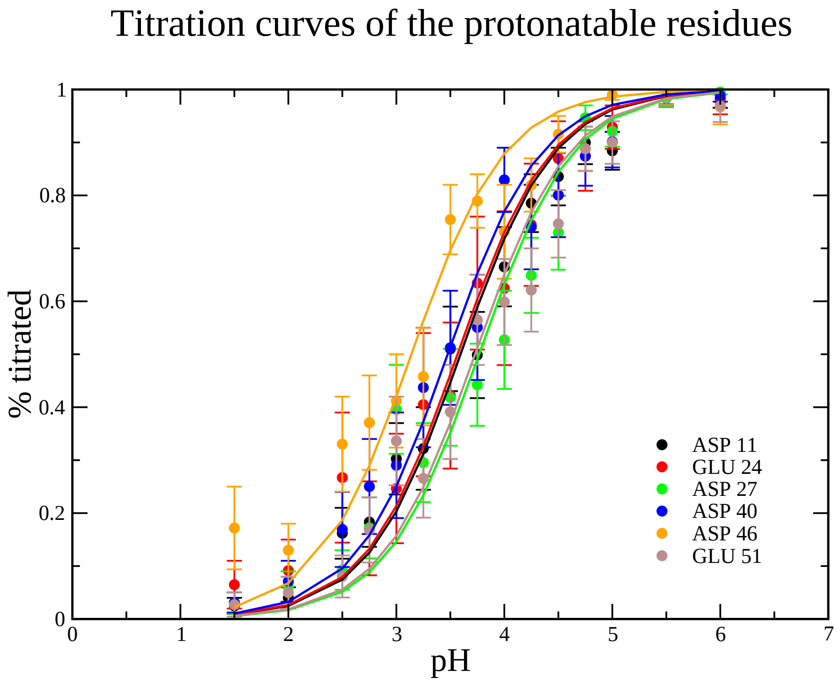


Figure 3.7: The fitted Henderson-Hasselbalch curves of the average lambda occupancy for the six titratable residues of the Epidermal Growth Factor

width and form.

From the PCA plot it can be seen, that the NMR ensemble is located close to the middle of the phase space of the flexible region and it covers a significant part of it. Thus, it is a good starting point to enhance the structural sampling. On the other hand, the constant pH MD simulations cover the entire region of the faster free MD with fixed protonation states. This suggests that the computational issues of a single trajectory in the constant pH MD code is partly compensated by the different starting structures and the small pH interval which also provides better structural statistics.

The calculated pK_a values, agree fairly with the experimental data and the largest deviation of 0.65 pK_a units, found in ASP 46, is sufficiently low. One possible reason for the deviation is the insufficient sampling of the

Table 3.3: Comparison between experimental and theoretical data for Epidermal Growth Factor

	Const pH MD	experiment(NMR)
ASP 11	3.6 +0.07 -0.06	3.9
GLU 24	3.6 +0.12 -0.1	4.1
ASP 27	3.8 +0.08 -0.07	4.0
ASP 40	3.5 +0.09 -0.08	3.6
ASP 46	3.15 +0.07 -0.07	3.8
GLU 51	3.7 +0.08 -0.07	4.0

fluctuating C-terminus. Although GLU 51 shows better agreement with experiment, despite its location in even more unstructured region, its complete exposure to water provides faster sampling. ASP 46 interacts closer with the other regions of the protein and it requires longer simulation times to equilibrate. Another explanation would suggest a too strong interaction with the located nearby positive arginine residues and a possible flaw in the forcefield describing their correct sidechain dynamics, leading to incorrect electrostatic interactions.

3.2.2 Cardiotoxin *V*

Structure and Function

The cardiotoxin family are proteins which are found in snake venom. Although most of them have hemolytic activity, Cardiotoxin *V* from *Naja naja atra* does not show depolarization or lysis of red blood cells[40]. Instead it induces aggregation and fusion of sphingomyelin vesicles[41] and its toxicity is mostly neurological.

The structure of the protein is solved by NMR [40]. It consists of 62 amino acid residues and it is has four disulphide bridges. The bridges stabilize a core with no definitive structure, from which antiparallel β strands form a β-sheet - Fig.3.8.

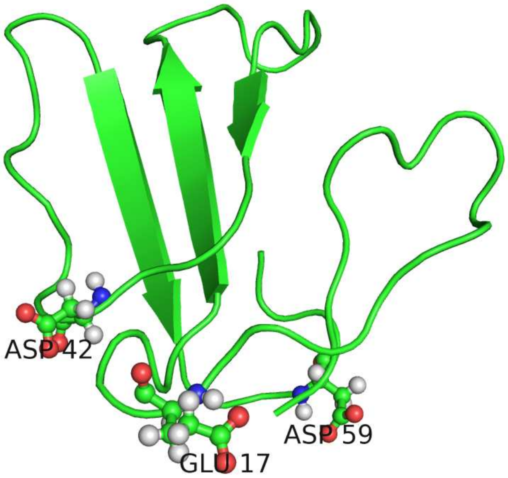


Figure 3.8: The structure of Cardiotoxin V with its three acidic residues

Simulation Setup

The Cardiotoxin V has three acidic residues that were used for titration - Fig.3.8. Furthermore it contains 13 positively charged residues at the acidic pH region and a very high net charge. Thus it is a system with very strong electrostatic interactions and interesting case for constant pH MD.

All of the three titrated residues were coupled to a water molecule to neutralize the charge change during titration. Again the translation and rotation of the center of mass was removed to avoid protein coming too close to the coupled water molecules. The titration simulations were performed similar to the scheme described for the Epidermal Growth Factor. The were five titrations each starting from pH=1 up to pH=6. The length of each simulation was approximately 12 nanoseconds. For the overall titration of each group, the lambda trajectories were concatenated for the given pH point.

Data analysis and Results

A Henderson-Hasselbalch fit of the average occupancy of lambda for each pH point is shown on Fig.3.9. The experimental pK_a values of the titrated residues are obtained from NMR experiments [42][39] and their comparison with the calculated values is shown in Table 3.4.

For all three residues, the Hill coefficient, from the Hill equation fit of the average occupancy of lambda, for each pH point, are in the range $0.7 \div 0.9$. The difference between the pK_a values calculated with them and the one calculated via Henderson-Hasselbalch fit are negligible.

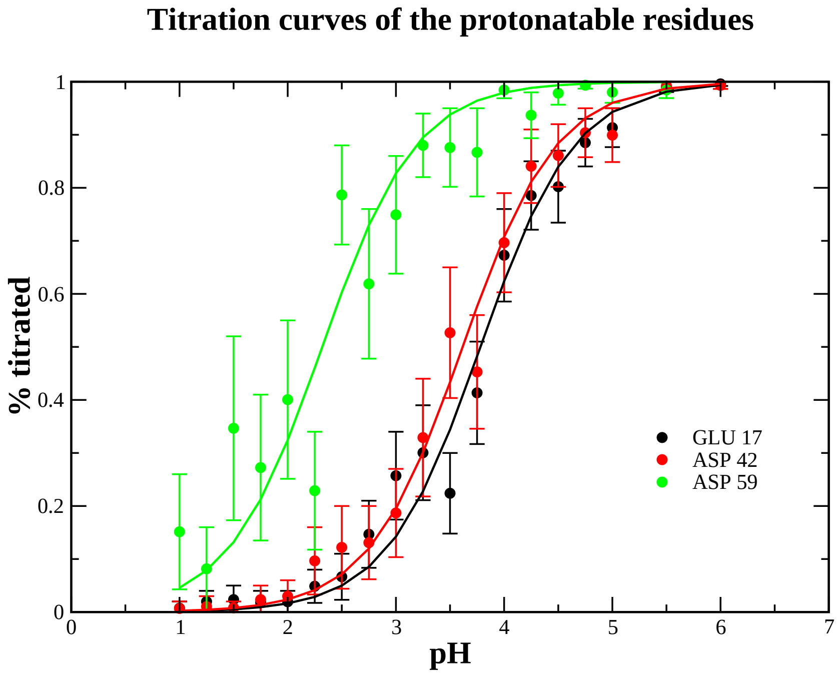


Figure 3.9: The fitted Henderson-Hasselbalch curves of the average lambda occupancy for the three titratable residues of Cardiotoxin V

Table 3.4: Comparison between experimental and theoretical data for Cardiotoxin V

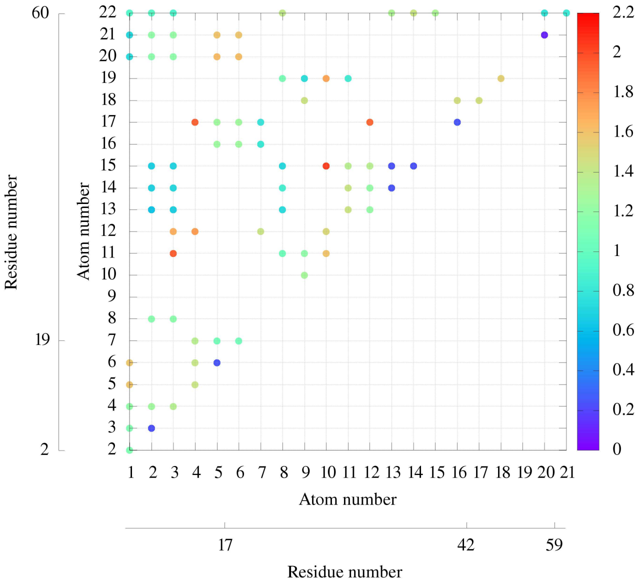
	Const pH MD	experiment(NMR)
GLU 17	3.8 +0.07 -0.06	4
ASP 42	3.6 +0.08 -0.07	4.1
ASP 59	2.3 +0.1 -0.1	2.3

The low pK_a value of ASP 59 is most probably due to the closeness of LYS 2 and LYS 60 as also suggested by [42]. In order to estimate if the pK_a shift is due to closeness of oppositely charged residues, the distance among all acidic and basic residues was calculated. On Fig.3.10 is shown the averaged distance of the oxygen and nitrogen atoms of all charged residues at two extreme pH points: pH 6, where all titrating sites are deprotonated and pH 2 where only ASP 56 is partly deprotonated. The spacial distribution of the titrated amino acids is shown on Fig.3.11.

Discussion

The microscopic treatment of the electrostatic interactions and the atom charges optimized for the specific force field, seems to give a good representation of the actual electrostatics of the protein even in such highly charged system. The electrostatic interactions are modelled correctly, and the result is a good agreement between experiment and theory.

Additionally, the relative structural stability of Cardiotoxin V and the fast sampling of the conformational space of the protein is another reasons for the good convergence and the good agreement between experimental and



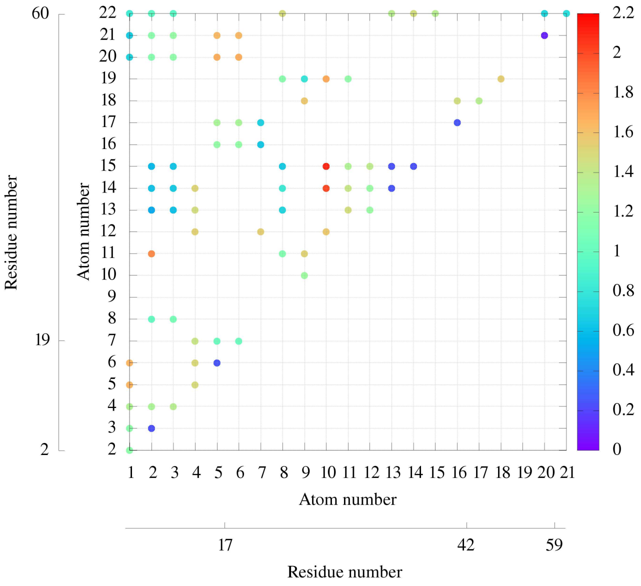


Figure 3.10: The average distance in nanometers (color coded) between the oxygen and nitrogen atoms of the charged groups at pH 2(upper graph) and pH 6 (lower graph) of Cardiotoxin V. The number of the atoms, included for the calculation, is shown on the first x-axis and y-axis, the second x-axis depicts the titrating residues and the second y-axis represents the positively charged residues which comes in close contact to the titrating sites.

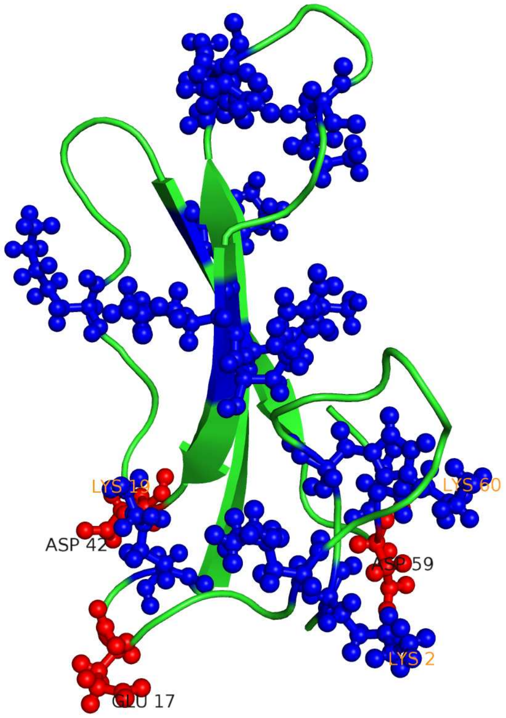


Figure 3.11: The distribution of ionizable residues of Cardiotoxin V. In red are the acidic residues which were titrated and in blue are the positive at acidic pH LYS, ARG and HIS. The residues that were described above are labelled accordingly

theoretical data.

From Fig 3.10 one can see that the closest positive residue to GLU 17 and ASP 42 is LYS 19 and at the same time ASP 59 is in close proximity with two positively charged resides - LYS 2 and LYS 60. As it can be seen the distance doesn't significantly change in the whole titrating range and the interaction with the two positive charges can explain the pKa shift of ASP 59. On the other hand the both negative GLU 17 and ASP 42 have only one positive LYS in proximity and that can explains why their pKa's are close to their solution values.

3.2.3 The Gly-His-Ala-His-Gly pentapeptide

Structure and Properties

The histidine residues are the most challenging ones for simulation since not only do they have two atoms that can be protonated, but these two atoms have different affinities toward proton hence different pK_a's. Therefore, they require longer sampling in order the switching coordinate to equilibrate, since one of its states will be visited much more rarely then the other.

In order to test our model, a synthetic pentapeptide containing two histidine residues was ordered and commercially synthesised for an NMR titration experiment. The later is carried with collaboration with Prof. Christian Griesinger's NMR II Department of Max Planck institute for Biophysical Chemistry. Similarly to the proteins described above, we performed constant pH MD simulation and the results are to be compared with experiment, when the experimental data is available.

The sequence of the peptide consists of two flanking glycine residues, followed by the two titrating histidine residues and an alanine residue in the middle - Fig.3.12. The C and N termini are synthesised charged, without

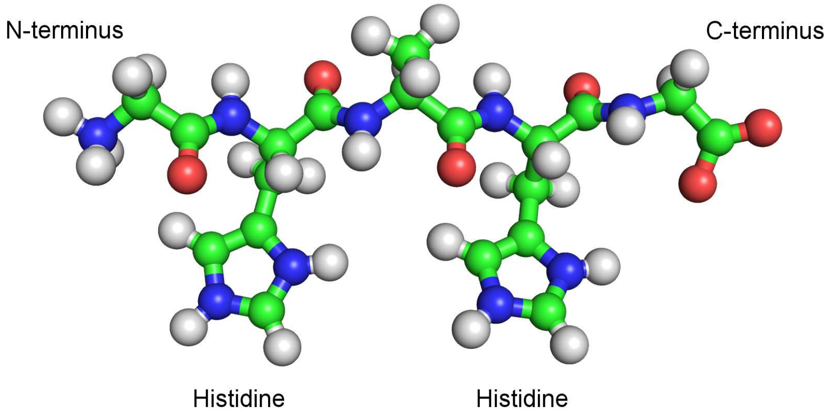


Figure 3.12: The structure of the pentapeptide

methyl caps. This was done due to experimental issues on one hand and the opportunity to estimate the influence of the two opposite charges on the pK_a's of the two histidine residues on the other.

Simulation Setup

Each of the two titrating residues was coupled to a water molecule like in the protein simulation described above. To prevent the peptide of coming too close to the coupled water molecule, a harmonic restraining potential was applied to the C-α carbon atom of the central alanine residue. Since the potential is applied to only one atom this should not effect the sampling of its conformational space.

There were two titration performed, starting from pH 4 to pH 8 with 0.25 pH unit step at the pH regions close to the solution pK_a of histidine and 1 pH unit step at the far ends of the interval. The length in each separate simulation was approximately 17 nanoseconds.

Data analysis and Results

Due to its small size and therefore large flexibility, the peptide may need long simulation time to sample properly its conformational space. Similar approach, as the one described for estimation the coverage of the configuration phase space of the Epidermal Growth Factor, was implemented here as well. Each histidine in the peptide has two nitrogen atoms in its imidazole ring that can protonate. Since they are not symmetric, there are overall four different sites that can undergo protonation. All possible combinations of protonated and deprotonated nitrogen atoms were simulated with the faster free MD. Then their atomic trajectories were concatenated and projected onto the eigenvectors of the atom positions covariance matrix of the united trajectory in all pH points of the constant pH MD. Their projection as well as the projection of the united constant pH MD simulation is shown on Fig.3.13.

The pK_a's of the two histidine residues were calculated again with the

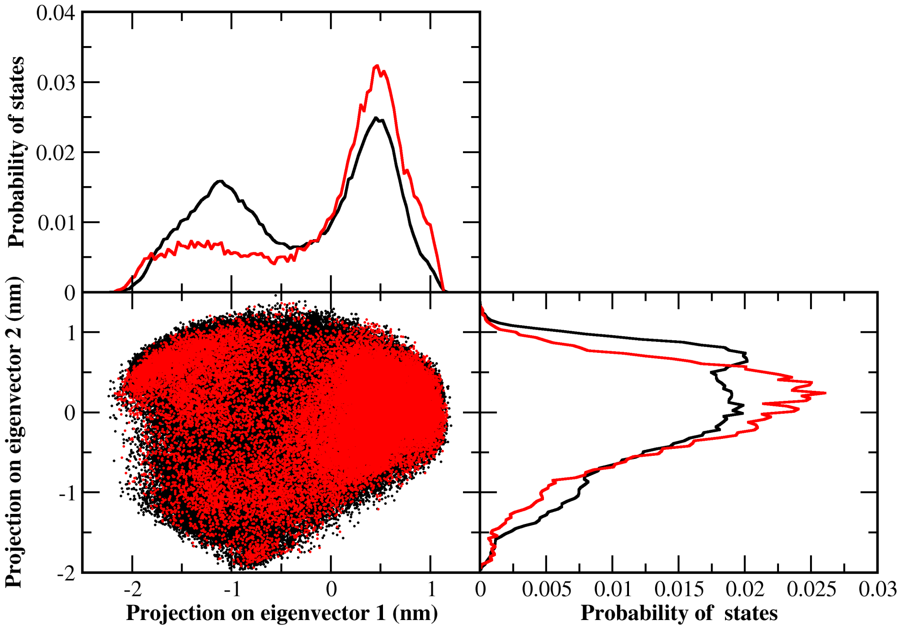


Figure 3.13: The projection of the atomic position trajectory of all combinations of fix protonation state free MD simulations (black) and the united constant pH MD simulations (red) on the first two eigenvectors of the atomic position covariance matrix of the constant pH MD simulations. On the right and top of the projection plot, are shown the probability distributions of a given trajectory projection onto the eigenvector.

approach described in the Theory and Methods Chapter. The Hill coefficients of the Hill equation fit to the average occupancy of lambda at each pH point are close to 1 and the difference of pK_a's calculated with them and calculated via Henderson-Hasselbalch fits are negligible. The results from the Henderson-Hasselbalch fits of the average lambda occupancy for each pH point and the calculated pK_a values are shown on Fig.3.14 and Table 3.5.

Table 3.5: Theoretical pK_a values of the pentapeptide

	Const pH MD	experiment(NMR)
N-term His	5.7 +0.1 -0.09	-
C-term His	6.2 +0.08 -0.07	-

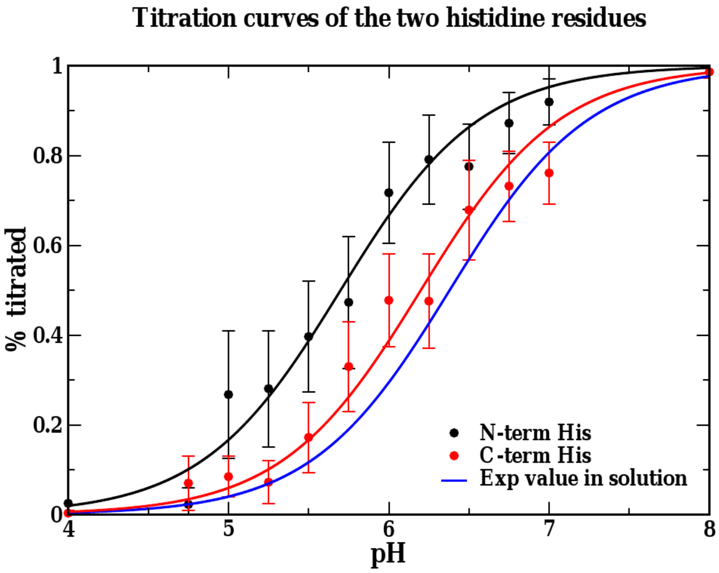


Figure 3.14: The fitted Henderson-Hasselbalch curves of the average lambda occupancy of the two histidine residues and Henderson-Hasselbalch curve of the reference pK_a in solution

Discussion

The first observation one can make is the down-shift of the pK_a's of both residues with respect to their reference values of 6.38 - Fig.3.14. This can be explained by the fact that there are two positive charges that have to appear next to each other in order to protonate both residues. The electrostatic repulsion makes this process unfavourable. The pK_a of the N-terminus histidine is shifted more then the one of the C-terminus histidine and that can be explained with the stabilization effect of the negative charge of the carboxyl group on the residue.

Another interesting fact seen from the simulation is that despite the two residues interact (seen from the pK_a shift), their titration can still be described reasonably well by Henderson-Hasselbalch curves. The Hill equation fits that were done on the average lambda for each pH point are close to 1 and the pK_a values are practically the same with the one obtained from Henderson-Hasselbalch fits. In practise this means that the two groups see each other only through their averaged electrostatics and their interaction is not directly co- or anti- correlated.

Chapter 4

Cytochrome C

4.1 Structure and Function

Cytochrome C from *Rhodopseudomonas viridis* is a small, globular, 107 amino acid protein that play a key role in the bacterial photosynthesis, carried by this specie. The photosynthesis type in this microorganism is a cyclic type and the electrons, which are transported through the reaction center during photosynthesis, are recycled back to it as shown on Fig.4.1. The excitation energy, of the absorbed light, is transferred to the located close to the periplasmic side of the plasma membrane special chlorophyll pair. There it causes shift of the chlorophyll's electrode potential, allowing subsequent oxidation from the next electron carriers located closer the cytosol side of the membrane. After that, the electrons reduce quinone to quinol molecules taking protons from the cytosol. The quinol molequles are transferred as neutral species through the membrane and oxidized by Cytochrome bc_1 complex releasing the protons on the periplasmic side of the plasma membrane. The result of this proton transfer is a proton gradient across the membrane and an elecric field created by this gradient. This electric field drives the

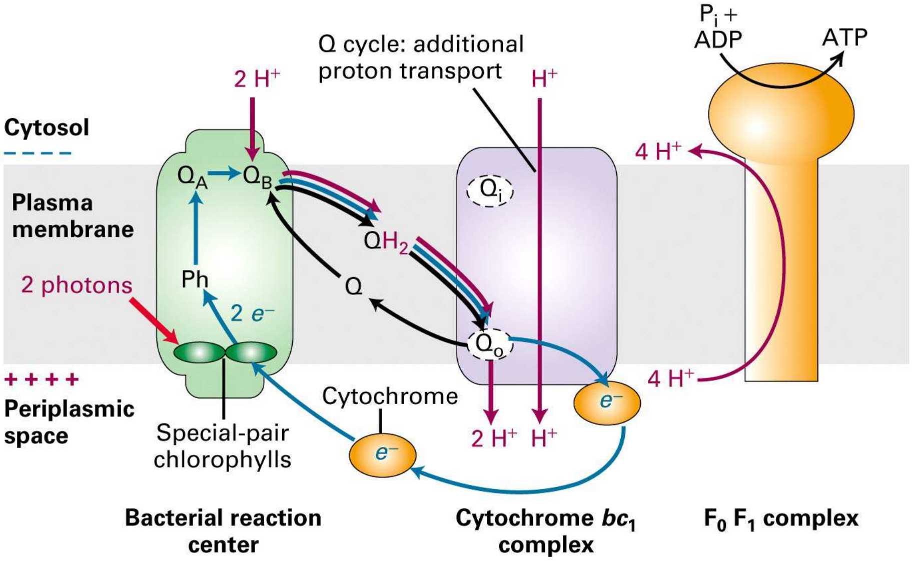


Figure 4.1: The scheme of the bacterial photosynthesis of *Rhodopseudomonas viridis* [43].

protons back to the cytosol side of the membrane and this flow is utilized from the ATP-synthase complex to convert ADP and inorganic phosphate to ATP. The electrons from the oxidation of the quinol molecules are recycled back to the reaction center by Cytochrome C, which is subsequently reduced and oxidized in the process.

Both structures of reduced[45] and oxidized[46] cytochrome C are solved by X-Ray crystallography and are shown on Fig 4.2. The structural differences between the two forms are very small and some significant deviations in the two structures can be seen only in the side chain orientation of GLU 65 and ARG 68. However, the authors of the two models suggest that the differences in the orientation of the side chains of amino acids, located near the surface, are due to their large flexibility and the uncertainty of electron density interpretation in these cases. They do not necessary have physiological meaning.

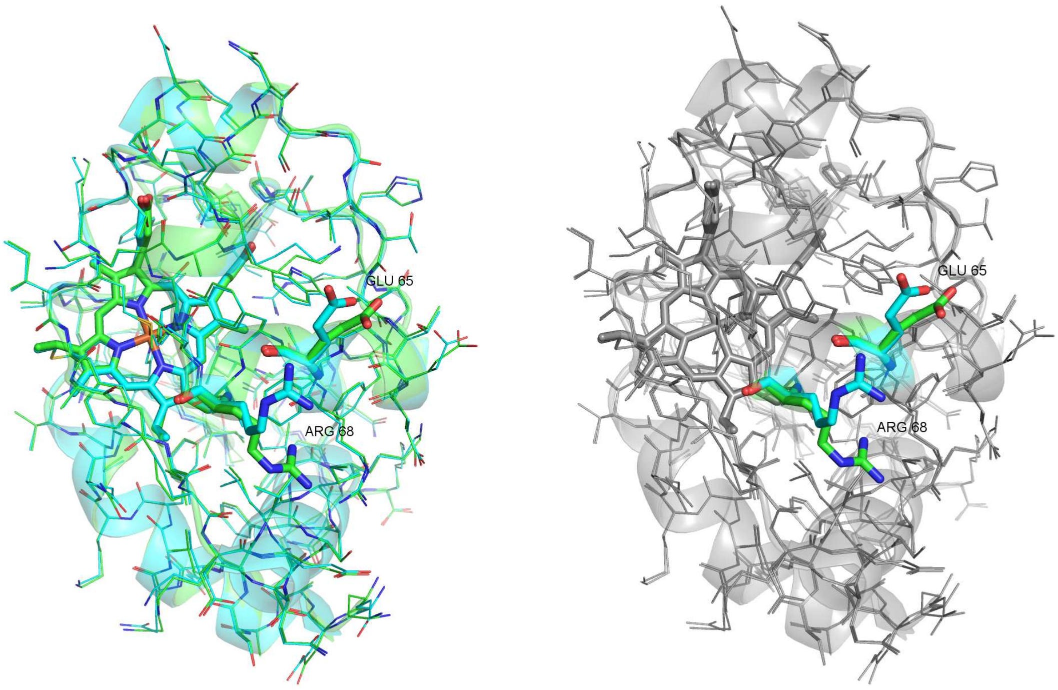


Figure 4.2: The structural difference between the reduced (green) and the oxidized (cyan) forms of cytochrome C (left) and the underlined GLU 65 and ARG 68 (right)

The electron carrying moiety of the cytochrome is a heme group type C, with iron center ligated by the sulphur atom of a methionine residue and a nitrogen atom of a histidine residue. Another two chemical bonds, with two cysteine residues, link the tetrapyrrole ring to the protein as shown on Fig.4.3. The heme has also two propionic side chains that carry two carboxyl functional groups. Upon oxidation and reduction the iron changes from Fe^{2+} to Fe^{3+} form, as the charge is delocalized in the tetrapyrrole ring.

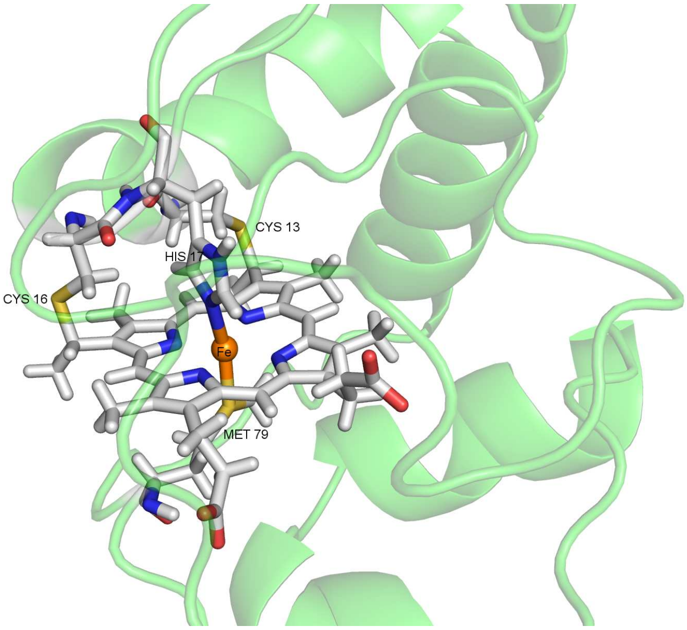


Figure 4.3: The heme group and its chemical bonds with residues from the protein moiety of Cytochrome C

4.2 Force Field of the HEME

The charges as well as the bond parameters of the heme group in both reduced and oxidized form, are not part of the standard Amber99sb force field, used for the current work. The charges for the heme group, the ligated histidine and methionine as well as the chemically linked cysteine were taken from[47] - Fig.4.5. and adapted to Amber99sb so that they give whole charge of the amino acids forming chemical bonds with the heme.

However for titrating the Propionic side chains (Fig.4.4) they had to be considered as separate residues that can undergo protonation. Since the charges of the respective atoms in the reduced and oxidized form are very similar, we took the average of them for the deprotonated form and adapted the charges from a protonated glutamate i.e. glutamic acid, from the Am-

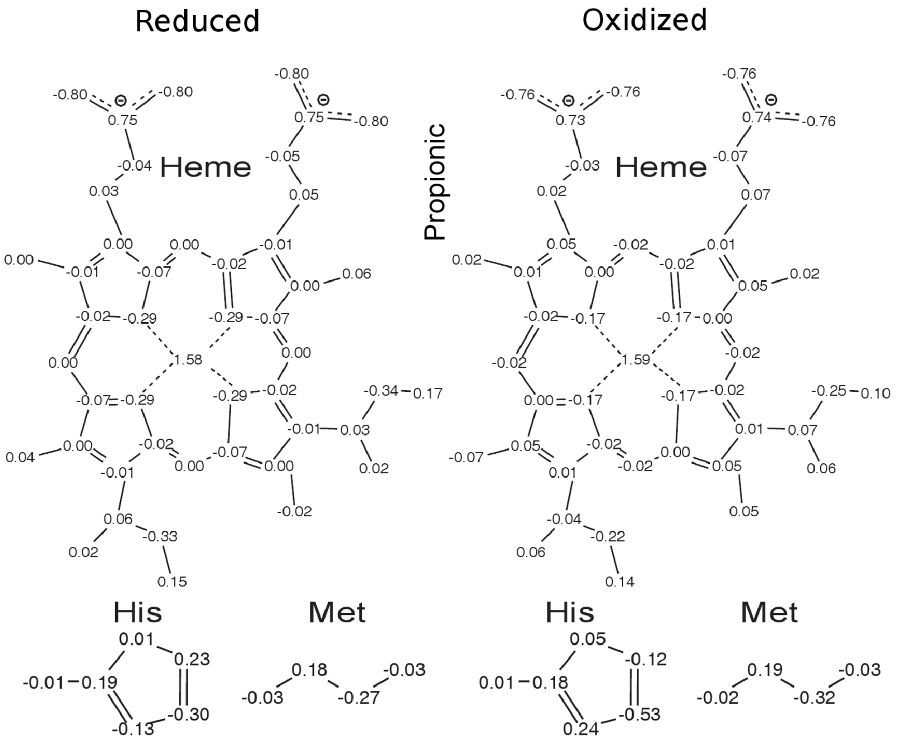


Figure 4.4: Charges of the heme and the chemically linked to it amino acids as originally calculated in [47]

ber99sb force field, for the protonated form. A small amount of charge, that is left due to the fact that difference between the charge of protonated and deprotonated form has to be exactly one, was added to the carbon atoms of the tetrapyrrole ring they were linked to, observing that there is no change of the sign of this atom's charge.

The reference pK_a value for the propionic side chains was also the one used for glutamate, and similar to glutamate, the four state model was used to simulate the carboxyl group.

The used force field parameters of the chemical bonds of the heme were the one calculated in [48] and their porting in Amber99sb was the one made by the authors of [44].

4.3 Protonation state of the titratable groups of Cytochrom C in reduced and oxidized state

The free energy of deprotonation of each titratable group in the protein will depend on the electrostatic field created by all of the charges in the protein. In the case of cytochrome C, there is a change of the the charge upon oxidation of the heme group, which is located in the low dielectric core of the protein. The shift of the electrostatic potential, on each protonatable group due to that process, can be enough to cause a shift of the deprotonation free energy and hence the protonation state of that particular group. Then the overall protonation pattern of the protein would rearrange in such a way so that it minimizes the effect of the oxidation. Thus the shift of the free energy of oxidation, or in other words the shift of the electrode potential of the heme, due to the protein environment and its titratable sites, can be a crucial player in the thermodynamics of the electron transfer. In previous studies, such an effect has been shown in the reaction center of *Rhodopseudomonas viridis*, where a charge dislocation from the special pair of chlorophylls to the next carrier - pheophytine, cases a large shift of the pK_a of Tyr L162, allowing it to deprotonate and thus stabilize the charge of the special pair of chlorophylls after photo-oxidation [44].

In order to investigate the change of the protonation state of the titratable groups upon oxidation, we performed constant pH MD simulations on both reduced and oxidized forms of the protein at pH 5. The acidic environment was chosen as in a time of ongoing photosynthesis, due to the proton pumping, the pH will be lower on the periplasmic side of the membrane, were cytochrome C functions.

4.4 Simulation Details and Setup

Constant pH MD simulations of both forms, reduced and oxidized were performed using the respective X-ray structures. All acidic residues including the propionic side chains of the heme, as well as the one histidine and three tyrosine residues were included in the simulation. Tyrosine residues although, basic with pK_a 9.6 [23] were considered as well, due their special location in the protein. All three of them are located within the hydrophobic core and TYR 66 is very close to the heme group. This may lead to large shifts of their pK_a's upon charge dislocation as shown in [44]. Since tyrosine has a hydroxyl functional group, there is only one oxygen that can undergo deprotonation. Therefore only a two state model was used to simulate the behaviour of the residue. All other acidic residues and the histidine, were modelled by the four state model as described in the Theory and Methods Chapter.

The overall simulation setup of the system can be seen on Fig.4.5. Again, all titration coordinates were coupled to restrained water molecules to keep the overall charge of the system neutral. There were 19 groups that were included in the simulation and 19 coupled water molecules. The closest distance among interacting partners, such as protein atoms and coupled water molecules, as well as their periodic images, was larger than 15Å. As shown in earlier in this chapter, this will not lead to large differences in the

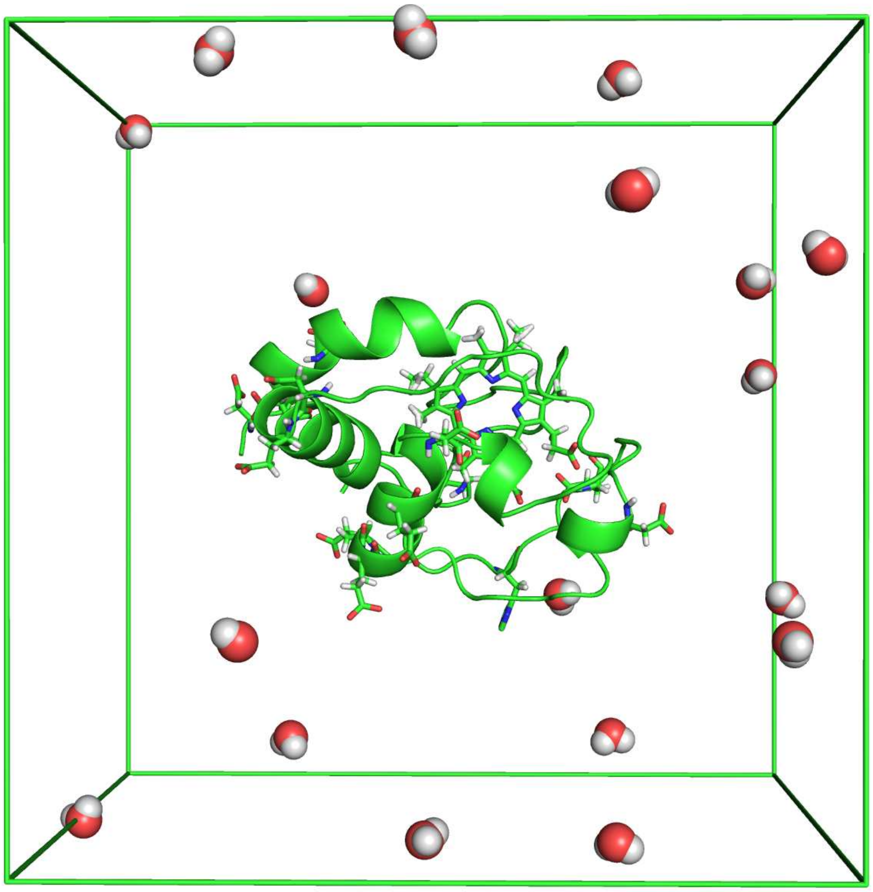


Figure 4.5: The constant pH MD simulation setup of Cytochrome C. The protein and the coupled water molecules in the simulation box.

free energy of deprotonation, due to the different distance between the coupled partners in the parametrization and the constant pH MD simulations.

For each of the two oxidation states, ten separate simulations were performed with length ranging from 4 to 8 nanoseconds. Thus, the overall length of simulations for each form was approximately 70 nanoseconds and all of the analysis was performed on the concatenated trajectory of the respective form.

4.5 Data Analysis and Results

The first observation from the trajectories of lambda particles, is the overall shift down of the pK_a's of almost all acidic groups. The reference value for the pK_a's of 3.95 for ASP [23] [24] and 4.4 for GLU [23] [24], should give relatively high protonation fraction for GLU and noticeable for ASP. If following HendersonHasselbalch behaviour the deprotonated form of the ASP residues should be only ten times more populated and the protonated GLU ones only four. However, almost no population of the protonated form was observed for most of the residues in both the oxidized and the reduced form, except GLU 43 and ASP 92. This effect can be explained by the presence of large number of postive lysine and arginine residues in the protein. Another interesting observation is the quite high population of deprotonated fraction of HIS 38 at such a low pH. For better characterization of the protonation behaviour, a PCA analysis (Fig.4.6) on the lambda coordinates of the titratable groups was performed. As it can be seen from the plot, the overall protonation pattern of the protein is guided mainly by the behaviour of three residues - GLU 43, ASP 92 and HIS 38. On Fig.4.7, the transitions among the states in the PCA can be seen in projection of the current state protonation vector on to the first and second protonation eigenvector as a function of time. The large number of transitions among the states is an indicator for reasonable sampling of the free energy landscape of the system. Fig 4.8 shows the individual protonation-deprotonation behaviour of the three groups. Each of them has sampled both states, and as it can be seen there has been large number of transitions even if one of states is much more favourable than the other.

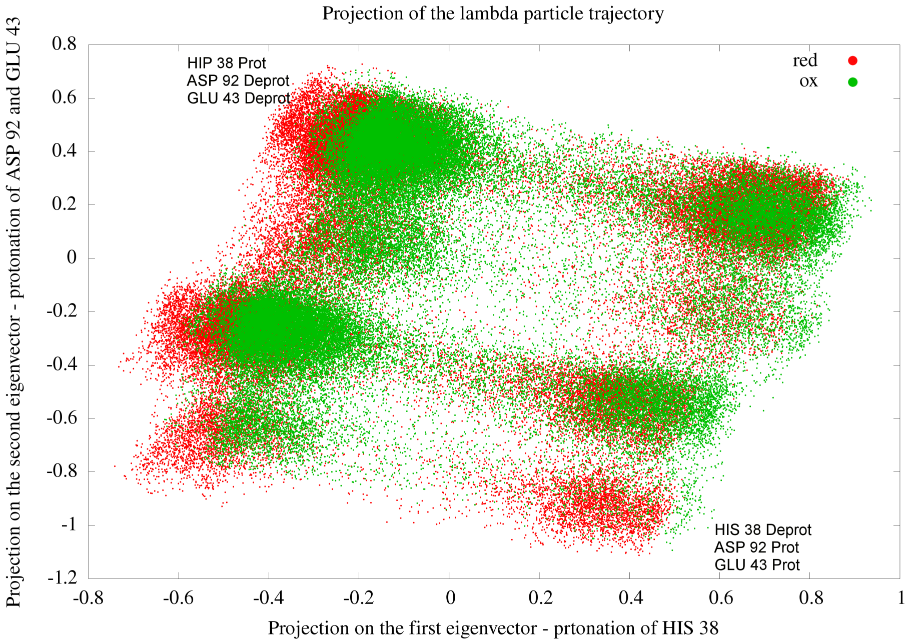


Figure 4.6: The PCA plot of the lambda particle coordinates. First two eigenvectors of the covariance matrix of the lambda trajectories represent the deprotonation - protonation of HIS 38 (first eigenvector) and GLU 43, ASP 92 (second eigenvector)

4.6 Discussion

Fig.4.8 shows that the individual deprotonation-protonation ratios of the three residues, that undergo transitions is not changed. However, from the PCA plot on Fig.4.7, it can be seen that there is a change of the occupancy of the different collective protonation states in the reduced and oxidized form of the protein. In both them, most populated state is the one with protonated HIS 38 and deprotonated ASP 92 and GLU 43. Although its occupancy is almost the same in both forms, the population of all the others is changed. A particular interest is the state with protonated ASP 92 and GLU 43 and deprotonated HIS 38, which is completely unpopulated in the oxidized form.

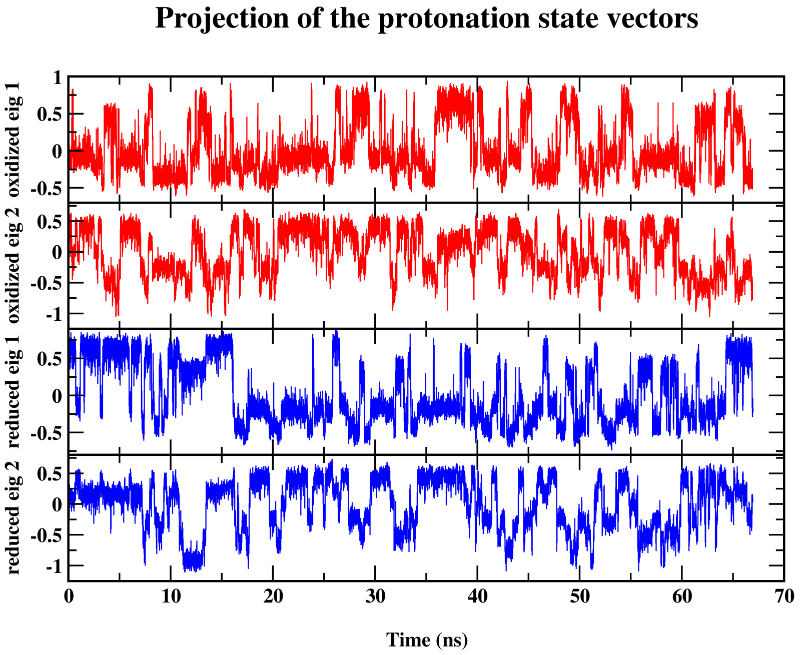


Figure 4.7: Projection of the protonation vector of the reduced(blue) and oxidized(red) form of Cytochrome C onto the first and the second eigenvector of the covariance matrix of the lambda coordinates of the reduced form of the protein.

The significance of these shifts is to be estimated quantitatively, by free energy calculations with thermodynamic cycle, in which the protein is oxidized with and without change of the protonation pattern.

The large number of positive residues and their electrostatic interaction can explain the low occupancy of the deprotonated forms of the acidic residues, despite the relatively low pH at which the simulations were performed. This suggests that in some regions with high density of positively charged residues, there might be a drop in the pK_a of the basic residues and especially lysine. In Poisson-Boltzmann calculation (data not shown) LYS

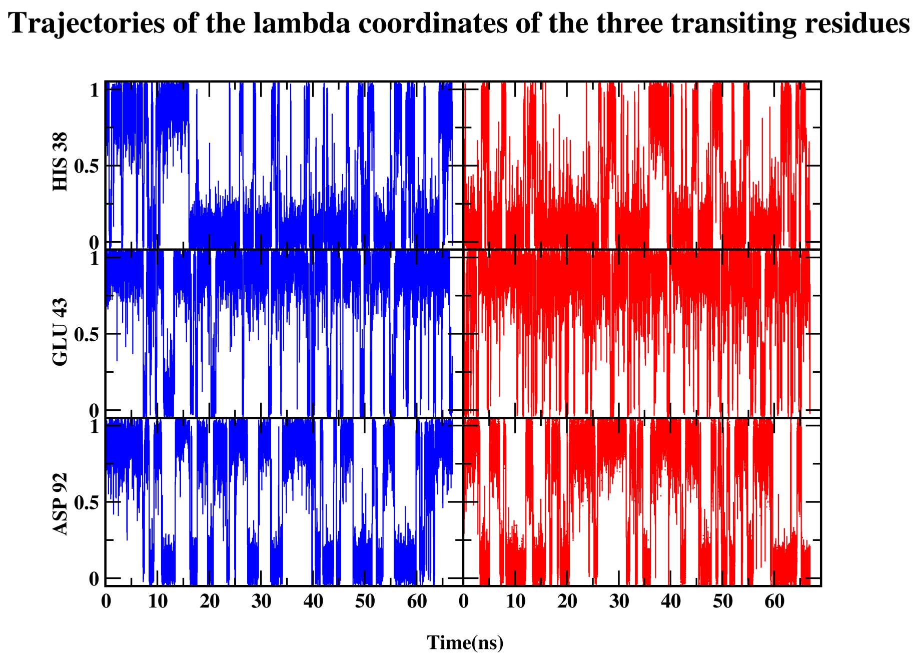


Figure 4.8: The trajectories of the lambda particles of the three transiting residues in the reduced (blue) and the oxidized (red) form.

24 has pK_a close to the pH region, in which the simulations were performed. Therefore it is possible to deprotonate and thus facilitates the protonation of some of the acidic residues. To test that, further simulations will be performed with carefully selected basic and acidic residues.

Chapter 5

Conclusions

Theoretical pK_a calculations of ionizable groups in proteins is challenging task and despite the great advance made in the field, there are still systems for which the existing approaches are not sufficiently accurate. Molecular dynamics simulations, although being able to give atomistic details for the dynamics of biological macromolecules, rely on force field description of the interactions among the particles, which assigns fixed protonation states of all titratable groups. However, due to possible incorrect assignment at the beginning of the simulation or due to conformational change in the protein, the fixed protonation state may not represent the electrostatic properties of system correctly. Furthermore there are cases in which given protonation state triggers response in the protein molecule [49], which would be never seen if the wrong protonation state is assigned beforehand.

The primary goal of this work was the implementation of explicit solvent molecular dynamics approach for calculating protonation behaviour of ionizable groups in electron transfer proteins. Additionally, the dynamics and the electrostatics of our model, were tested on protein systems with the according properties and the experimentally known pK_a values agree fairly with the data obtained from our simulations. The average deviation from the experimental values was 0.3 pK_a units with largest difference of 0.65 pK_a units. The latter was found in the flexible C-terminus region of the Murine Epidermal Growth Factor. The pK_a value we calculated in this case is shifted down with respect to experiment, and that can be attributed to metastable initial states of the titrated ASP residue, interacting with nearby ARG residues in some of the initial structures. Thus, long simulation time would be necessary to leave that state and this leads to insufficient sampling.

Unfortunately, the current implementation of the λ-dynamics approach is still very expensive. The force acting on each λ-particle has to be calculated. That leads to as many calculation of the derivative of the particles' Hamiltonians with respect to their reaction coordinates, as many group there are. Thus, the performance scales down linearly, with the number of groups included for constant pH MD. Partial solution to that problem is already available[50] and it based on calculating the derivative of each lambda on a separate CPU. However, for this approach, a large number of well scaling CPUs is required.

Much more effective approach, requires new way of calculating the long-ranged electrostatic interactions. Such an implementation will make the speed of the constant pH MD independent on the number of groups, and its performance should be as fast as regular free energy calculation. However, this approach is technically much more difficult for realisation and would require much longer time to be implemented.

The importance of efficient sampling, can be seen in the case of the Cardiotoxin V system. The protein has only three titratable residues and thus its constant pH MD simulation is much more efficient. The results show good agreement with experiment, and particularly for ASP 59 whose value in protein is 1.6 pK_a units lower then its reference value. It seems that the electrostatics of the force field models correctly the interactions among charged atoms and the low pK_a value of this residue is estimated reasonably well.

Since the results for the two test proteins were in fair agreement with experiment, the next step in testing our method, would be to predict pK_a values of a system that then can be experimentally titrated. That was the primary intention behind the simulation of the synthetic Gly-His-Ala-His-Gly pentapeptide. As it can be seen from the sequence, the two histidine residues are located closely and will interact with each other. Thus, this is an important control system, in which we can test the accuracy of our calculations in the most complicated residue, with two titrating sites, with different pK_a values. The calculated pK_a's are shifted down from their reference values due to the appearance of two positive charges next to each other. In that sense, the behaviour of the system is as one can expect. When the experimental results are available, they will be valuable control case for the method.

After the method was tested it was implemented in the primary goal of this work, namely electron transfer proteins. The interest in that area was motivated from the importance of the electron transfer as it is at the base of energy conversion in the living systems and thus is one of most fundamental processes in the biosphere. Due to the large charge change in the low dielectric environment of the protein, the protonation state of the ionizable groups around the electron carriers will have large impact on the thermodynamics of the process. To address some of these questions, we implemented our method on relatively simple redox system namely soluble Cytochrome C from *Rhodopseudomonas viridis*.

The results show large number of transitions between protonated and deprotonated state in most of the residues, even if they don't spend much time in the state that is unfavourable for them. From that one can assume that the free energy landscape is sampled well and the protonation pattern is equilibrated. This can be seen also in the significant number of transition among the different states found in the PCA. The most interesting observation that can be made, from the data available so far, is that despite the average protonation of the residues contributing mostly to the protonation behaviour of the protein are not changed in the two forms, their collective response is. That suggests that the protein environment is responding collectively to the oxidation. For further estimation of this effect, free energy calculations in thermodynamic cycle build on different redox states and different protonation patterns can be used.

Methods, that rely on continuum electrostatics models, have been most widely used for pK_a calculation in the field of molecular modelling. Despite their success in many cases, they rely on static models and do not provide dynamic description of the modelled system.

Molecular dynamics on the other hand, is a powerful tool to study the structural changes of the protein with atomistic details. However, the fixed charges of the ionizable residues in MD, can not describe correctly the change of their protonation in the course of the simulation.

For that reason, a dynamic protonation state models have been implemented in implicit solvent MD. However, the implicit treatment of the solvent, misses important interactions and entropy contributions from the solvent molecules.

In this work, for the first time, we applied explicit solvent constant pH

MD method in protein simulations. The method was first tested on small prototypic proteins with experimentally known pK_a values. The results from their titration indicate, that our method, based on explicit solvent force filed representations of the systems, models the dynamics and electrostatics of the titratable residues with sufficient accuracy, to allow dynamic treatment of their protonation state. With this approach, it is now possible to obtain more reliable information on both, the atomic and protonation configuration space.

Electron transfer, as one of the most fundamental processes in nature, has been modelled extensively. Due to the fact, that shifts in the charge distribution, is its main feature, many studies have been carried, using continuum electrostatics approaches. However, it has been shown, that the dynamics and particularly the frequency of vibration of specific residues is crucial for the electron transfer process[51].

Moreover, hydrogen bonding and explicit treatment of the solvent are often important for correct estimation of the charge interactions in protein, which are crucial in proton coupled electro transfer. These contributions however, are not included in continuum electrostatic models.

With explicit solvent constant pH MD, we include in the model of electron transfer both of the features - dynamics of the system with atomistic details, as well as dynamic protonation states, to account for the interactions between the electron carriers and the ionizable groups in the protein. Thus, we brought two of the most important characteristics of the electron transfer in one model.

However, despite the good agreement with the experimental data and fast convergence in the case of cytochrome C, the systems that were simulated were relatively small with ionizable resudes exposed in solution. For larger systems, with residues located in the interior of the proteins, longer simulation times are needed for a structural reorganization to take place. Therefore, in order to have correct sampling in those cases, a better technical implementation of the constant pH MD code is required. Such implementation is possible and there are solutions currently developed in our Department of Theoretical and Computational Biophysics.

Acknowledgements

Many thanks to Prof. Helmut Grubmüller for the guidance and support during my work, to Serena Donnini who is the developer of the constant pH MD code, Gerrit Groenhof who was a second supervisor of my work, our GROMACS expert Carsten Kutzner, our system administrators Ansgar Esztermann and Martin Fechner for the technical support, and the whole Department of Theoretical and Computational Biophysics at Max Planck Institute for Biophysical Chemistry Göttingen.

Bibliography

[1] D. Bashford, K. Gerwert (1992) J. Mol. Biol. vol. 224 pp. 473-486

[2] Donald Bashford, volume 1343 of Lecture Notes in Computer Science, pages 233–240, Berlin, 1997. ISCOPE97, Springer.

[3] Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B.(1993) Proteins, 15(3)252-65

[4] Lee, M. S., Salsbury, F. R., Jr., and Brooks, C. L., III (2004), Proteins 56, 738-752.

[5] Khandogin J, Brooks CL 3rd., Biophys J. 2005 Jul;89(1):141-57. Epub 2005 Apr 29.

[6] Mongan, J.; Case, D. A.; McCammon, J. A. J., Comput. Chem. 2004, 25, 20382048

[7] Donnini S, Tegeler F, Groenhof G, Grubmller H., J. Chem Theory and Comp 7: 1962-1978 (2011)

[8] Singhal AK, Chien KY, Wu WG, Rule GS, Biochemistry. 1993 Aug 10;32(31):8036-44.

[9] Montelione GT, Wthrich K, Burgess AW, Nice EC, Wagner G, Gibson KD, Scheraga HA., Biochemistry. 1992 Jan 14;31(1):236-49.

[10] Kohda D, Sawada T, Inagaki F., Biochemistry. 1991 May 21;30(20):4896-900.

[11] Chiang CM, Chien KY, Lin HJ, Lin JF, Yeh HC, Ho PL, Wu WG., Biochemistry. 1996 Jul 16;35(28):9167-76.

[12] http://pka.engr.ccny.cuny.edu/

[13] Helmut Grubmüller: Proteins as Molecular Machines: Force Probe Simulations, Computational Soft Matter: From Synthetic Polymers to Proteins, NIC Series, Vol. 23, ISBN 3-00-012641-4, pp. 401-422, 2004

[14] R. W. Hockney and S. P. Goel: Quiet high-resolution computer models of a plasma, Journal of Computational Physics, 14:148–158, 1972

[15] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije: LINCS: A linear constraint solver for molecular simulations, Journal of Computational Chemistry, 18:1463–1172, 1997

[16] Hornak et. al (2006). Proteins 65, 712-725

[17] Sorin, Pande (2005), Biophysical Journal, 88, 2472-2493

[18] DePaul, Thompson, Patel, Haldeman, Sorin (2010), Nucleic Acids Research, 38, 4856-4867

[19] Feenstra, K.; Hess, B. , Berendsen, H. Improving efficiency of large timescale molecular dynamics simulations of hydrogen-rich systems Journal of Computational Chemistry, Citeseer, 1999, 20, 786-798

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

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

[22] H-NMR study on the tautomerism of the imidazole ring of histidine residues: I. Microscopic pK values and molar ratios of tautomers in histidine-containing peptides, Masaru Tanokura, Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology, Volume 742, Issue 3

[23] Nozaki, Y. Tanford, C. 1967. Examination of titration behavior. Methods Enzymol. 11: 715734.

[24] Proteins. Structure and molecular properties. Second edition. Thomas E. Creighton

[25] Erik Lindahl, Berk Hess, and David van der Spoel: GROMACS 3.0: a package for molecular simulation and trajectory analysis, Journal of Molecular Modeling, 7:306–317, 2001

[26] Gaussian 03, Revision C.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A.

D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, Gaussian, Inc., Wallingford CT, 2004.

[27] Matteo Ceccarelli, Piero Procacci, Massimo Marchi An *Ab initio* Force Field for Cofactors of Bacterial Photosynthesis, J Comput Chem 24: 129-142, 2003

[28] The original SPC/E reference is H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, The missing term in effective pair potentials, J. Phys. Chem. 91 (1987) 6269-6271.

[29] Nose, S. A molecular dynamics method for simulations in the canonical ensemble. Mol.Phys. 52:255268, 1984.

[30] Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. Phys. Rev. A 31:16951697, 1985.

[31] The Andersen thermostat in molecular dynamics, Weinan E1, Dong Li2. Communications on Pure and Applied Mathematics 61(1) 96136, 2008.

[32] Dynamic Protonation of Titratable Groups in Biomolecules for Molecular Dynamics Simulations. Diploma thesis: Florian Tegeler, Helmut Grubmller

[33] The PyMOL Molecular Graphics System, Version 1.3, Schrdinger, LLC

[34] Essential dynamics of proteins. Andrea Amadei, Antonius B. M. Linssen, Dr. Herman J. C. Berendsen, Proteins: Structure, Function, and Bioinformatics, 17, 412425, (1993)

[35] Pattern Classification, Duda, R., Hart, P., Stork, D. (Wily-Interscience, 2001).

[36] Epidermal Growth Factor and transformation growth factor alpha, Burgess AW, Br Med Bull. 1989 Apr, 45(2): 401-24. Review

[37] Solution structure of murine epidermal growth factor determined by NMR spectroscopy and refined by energy minimization with restraints, Gaetano T. Montelione, Kurt Wuethrich, Antony W. Burgess, Edward C. Nice, Gerhard Wagner, Kenneth D. Gibson, Harold A. Scheraga, Biochemistry, 1992, 31 (1), pp 236249

[38] Kohda D, Sawada T, Inagaki F., Biochemistry. 1991 May 21;30(20):4896-900.

[39] http://pka.engr.ccny.cuny.edu/

[40] Singhal AK, Chien KY, Wu WG, Rule GS, Biochemistry. 1993 Aug 10;32(31):8036-44.

[41] Fusion of sphingomyelin vesicles induced by proteins from Taiwan cobra (Naja naja atra) venom. Interactions of zwitterionic phospholipids with cardiotoxin analogues. KY Chien, WN Huang, JH Jean and WG Wu, J. Biol. Chem., 266, 5, 3252-3259, 02, 1991

[42] Chiang CM, Chien KY, Lin HJ, Lin JF, Yeh HC, Ho PL, Wu WG., Biochemistry. 1996 Jul 16;35(28):9167-76.

[43] (5 Ed.). H. Lodish, A. Berk, P. Matsudaira, C.A., Kaiser, M. Krieger, M.P. Scott, S.L. Zipursky, J. Darnell. Arhen. Editorial Mdica Panamericana. Buenos Aires. (2004)

[44] Light-Induced Structural Changes in a Photosynthetic Reaction Center Caught by Laue Diffraction. Annemarie B. Whri, Gergely Katona, Linda C. Johansson, Emelie Fritz, Erik Malmerberg, Magnus Andersson, Jonathan Vincent, Mattias Eklund, Marco Cammarata, Michael Wulff, Jan Davidsson, Gerrit Groenhof, Richard Neutze, SCIENCE, VOL 328, 2010

[45] Refined crystal structure of ferrocytochrome c2 from Rhodopseudomonas viridis at 1.6 A resolution. Sogabe, S., Miki, K.,J.Mol.Biol. 252: 235-246, 1995

[46] Crystal structure of the oxidized cytochrome c(2) from Blastochloris viridis. Sogabe, S., Miki, K., FEBS Lett. 491: 174-179, 2001

[47] Classical force field parameters for the heme prosthetic group of cytochrome C. Felix Autenrieth1, Emad Tajkhorshid, Jerome Baudry1, Zaida Luthey-Schulten, Journal of Computational Chemistry Volume 25, Issue 13, pages 16131622, October 2004

[48] An ab initio force field for the cofactors of bacterial photosynthesis. Matteo Ceccarelli, Piero Procacci, Massimo Marchi, Journal of Computational Chemistry Volume 24, Issue 2, pages 129142, 30 January 2003

[49] Niemeyer et al., Journal of Physiology April 1, 2009; 587 (7)pp. 13871400.

[50] http://www.mpibpc.mpg.de/home/grubmueller/projects/Methods/ConstpH/usage/index.html

[51] Dispersed Polaron Simulations of electron Transfer in Photosynthetic Reaction Centers, A. Warshel, Z.T. Chu, W. W. Parson, Science, Vol 246

Curriculum Vitae

Personal information

Plamen Dobrev

Born 17.11.1983 in Harmanli (Bulgaria)

Nationality: Bulgarian

Education

05.2012	Doctoral exam for the award of the degree “doctor rerum naturalium”, Division of Mathematics and Natural Sciences of the Georg-August-Universitat Göttingen
10.2008–04.2012	Ph.D. Thesis in the group of Prof. Helmut Grubmüller, Max Planck Institute for Biophysical Chemistry, Göttingen, on topic: “Protonation patterns in reduced and oxidized forms of electron transfer proteins”
since 01.2009	Member of the Göttingen Graduate School for Neurobiology and Biomolecular Biosciences (GGNB)
10.2008–04.2012	Master Thesis in the group of Prof. Helmut Grubmüller, Max Planck Institute for Biophysical Chemistry, Göttingen, on topic: Molecular dynamics simulations of GPCRs - structure and stability
10.2006-09.2008	Master program Molecular Bioengeneering at BIOTEC at TU-Dresden
10.2002-09.2006	Bachelor in Molecular Biology at Sofia University “St. Kliment Ohridski”