2. The protonation pattern of proteins

Several among the 20 amino acids, which are the building blocks of all proteins, have a side chain that can be protonated or not, depending on the pH. These amino acids are referred to as titratable amino acids and usually the acidic groups glutamic acid, aspartic acid and cysteine as well as the basic groups arginine, lysine and histidine are considered to belong to this group of amino acids. The side chains of these amino acids exist in an acid-base equilibrium in solution. The set of titratable groups of a protein is completed by the amino group of the first and the carboxyl group of the last amino acid of a chain. They are not involved in a peptide bond and therefore also capable of protonation and deprotonation.

Each of the mentioned titratable groups has a well understood standard behavior concerning the acid-base equilibrium in aqueous solution. This is defined via the pK_a value and can be found in biophysical handbooks.[10,11]

The acid-base properties of a titratable amino acid side chain in a protein can differ significantly from its aqueous solution standard value. Each individual titratable group may show its own shifted behavior in correspondence to the local structural situation within the protein. This is due to the interaction of the titratable group with the partial charges of the protein atoms and the change of the dielectric properties when going from aqueous solution to the protein interior. Moreover the interaction between two charged titratable groups influences the energetics of the corresponding acid-base equilibria mutually.

The knowledge of the electrostatic properties, including the positions of possible ionic charges is crucial in understanding the function of proteins and especially enzymes. In this chapter, it will be discussed how electrostatic models can be used to learn more about the protonation state of a titratable group inside a protein and how it is even possible to establish the complete protonation pattern of a protein. Therefore, we begin with a summary of the fundamental electrostatic concepts that are important in the description of macromolecules.

2.1. Theory of electrostatic interactions in macromolecules

The basis of electrostatics is formulated with the Poisson equation:

$$\bigtriangledown^2\phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}). \tag{2.1}$$

It relates the electrostatic potential ϕ at position $\mathbf{r}$ with the spatial charge density ρ. Both ϕ and ρ are variable in space.

In principle all models that are used to investigate electrostatic properties of macromolecules are derived from the Poisson equation. It is dependent from the considered problem how the equation is used: If the charge distribution of a given system is homogeneous with $\varepsilon = 1$ and can be described explicitly by point charges q_i, the solution of the Poisson equation becomes Coulomb's law:

$$\phi(\mathbf{r}) = \sum_i \frac{q_i}{|\mathbf{r} - \mathbf{r_i}|}, \tag{2.2}$$

where $\mathbf{r}$ is the position and q_i the magnitude of the ith point charge. If all charges of a system are represented explicitly, the system can be described by a homogeneous dielectric medium with $\varepsilon = 1$, such that all interactions can be considered to occur in free space and Coulomb's law can be applied straightforward to calculate for example the electrostatic energy of the system:

$$E_{el} = \sum_i \phi(\mathbf{r_i}) q_i = \frac{1}{2} \sum_{ij} \frac{q_i q_j}{|r_i - r_j|} \tag{2.3}$$

If the explicit representation of all charges and their positions is not feasible or not desired, the Poisson equation can be varied to adopt this situation. If, for example a charge distribution is present in an environment different from vacuum, the effect or response of the corresponding medium (*e. g.* water) on the electrostatic potential $\phi(\mathbf{r})$ has to be taken into account. The medium can then be represented via a dielectric constant $\varepsilon \neq 1$ and Eq. 2.1 becomes

$$\bigtriangledown^2 \phi(\mathbf{r}) = \frac{-4\pi\rho(\mathbf{r})}{\varepsilon}. \tag{2.4}$$

The corresponding Coulomb law is then

$$\phi(\mathbf{r}) = \sum_i \frac{q_i}{\varepsilon |\mathbf{r} - \mathbf{r_i}|}, \tag{2.5}$$

In a homogeneous medium like an aqueous solution, the dielectric constant can be considered to be the same everywhere in the solution. A more complex situation exists for example, when an aqueous solution of macromolecules shall be described. The system is then heterogeneous and the dielectric constant, *i. e.* the response of the medium to a charge will be different in the bulk water and inside the macromolecule. The dielectric constant varies through space. The Poisson equation has to be modified and becomes:

$$\bigtriangledown \cdot \varepsilon(\mathbf{r}) \bigtriangledown \phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}). \tag{2.6}$$

Coulomb's law cannot be formulated to adopt this situation and is not applicable when the dielectric constant varies within the system.

2.1.1. The choice of the dielectric constant

Each method, that includes the calculation of electrostatic energies has to include an appropriate description of the dielectric properties of the medium, in which the electrostatic interactions take place. The choice of the dielectric constant is crucial for all results as can easily be seen from Eqs. 2.5 and 2.6.

Three physical processes determine the dielectric behavior of a medium:

1. Electronic polarization, that describes the reorientation of the electronic cloud around a nucleus in the presence of an electric field.

2. Nuclear Polarization, *i. e.* the reorientation of permanent dipoles, *e. g.* the water molecules or the dipoles within a macromolecule that reorient in response to an electric field.

3. Redistribution of charges, *e. g.* the movement of mobile ions in ionic solutions.

The electrostatic models used today consider these effects, however sometimes only implicitly. In standard empirical force fields used for Molecular Dynamics or Monte Carlo simulations all atoms of a molecular system are represented in detail. All electrostatic interactions between the corresponding point charges are then calculated explicitly, such that the electrostatic energy is given by Eq. 2.1. If also all solvent molecules are represented explicitly the dielectric constant can be set to $\varepsilon = 1$ throughout the whole system. This procedure has its disadvantages in high computational costs, because many atom pair interactions have to be evaluated. To overcome this problem, cutoffs are introduced that limit the distance up to which electrostatic interactions are evaluated. This reduces the number of atom pairs, but causes also new problems as electrostatic interactions have a long range nature due to which the usage of cutoffs has to be accompanied by a treatment of long range interactions, that go beyond the chosen cutoff.

The electronic polarization is not handled explicitly in a treatment that ascribes each atom a point charge as is done in most force fields until today. The electronic polarization is however accounted for within the reorientation of permanent dipoles by adjusting the values of the permanent dipoles by the force field parameters. The redistribution of mobile ions is in general neglected in force field simulations.

Using an explicit representation of all atoms causes problems in describing a macromolecule in solution. The number of solvent molecules that can be incorporated into the solvent sphere or box is naturally limited, such that boundary conditions have to be introduced, whose effects have to be considered carefully.

Sometimes it is not possible to describe each solvent atom with its explicitly. Then the dielectric constant is not unity anymore and in most cases it will not be constant throughout the whole system. Water has a dielectric constant of 80 at room temperature. Experimental and theoretical investigations suggest that proteins have an average dielectric response that can be approximated with a dielectric constant of about 4. Therefore two dielectric constants have to be used. In force field simulations this is done, when the solvent or a part of the solvent is treated as a continuum.[12,13] The usage of different dielectric constants within a molecular system is also applied in electrostatic continuum models like the Poisson-Boltzmann model that will be described in the next section.

2.1.2. The Poisson-Boltzmann equation

If one is not interested in the electrostatic properties of a large number of molecular configurations, that have to be generated by a Molecular Dynamics (MD) or a Monte Carlo (MC) approach, it is possible to use a dielectric continuum model, as for example the Poisson Boltzmann equation. The computational costs of evaluating the Poisson-Boltzmann equation of a macromolecule preclude its usage within an MD or MC simulation. In the Poisson-Boltzmann approach the solvent as well as mobile ions are not treated explicitly. Instead a well chosen dielectric constant that accounts for the less detailed description and an expression for the ionic strength is used. The entropic and electrostatic contribution to the chemical potential of an ion in solution at a point $\mathbf{r}$ are $kT\,ln\,c(\mathbf{r})$ and $q\phi(\mathbf{r})$ respectively, where $c(\mathbf{r})$ is the local concentration, q its charge and $\phi(\mathbf{r})$ the electrostatic potential. The ionic concentration can be described with a Boltzmann expression:

$$c(\mathbf{r}) = c^{bulk} exp\left(\frac{q\phi(\mathbf{r})}{kT}\right) \tag{2.7}$$

where k is the Boltzmann constant and T the absolute temperature. This expression can be incorporated into the Poisson equation (Eq. 2.6) leading to the Poisson-Boltzmann equation:

$$\bigtriangledown \cdot \varepsilon(\mathbf{r}) \bigtriangledown \phi(\mathbf{r}) = -4\pi \left(\rho(\mathbf{r}) + c^{bulk} q exp \left(\frac{-q\phi(\mathbf{r})}{RT} \right) \right) \tag{2.8}$$

R is the gas constant. For small electrostatic potentials $(e_0\phi(\mathbf{r})/RT < 1)$ Eq. 2.8 can be written in its linearized form:

$$\begin{aligned} &\sum_{i=1}^{K} c_i^{bulk} Z_i e_0 - exp \left(\frac{-Z_i e_0 \phi(\mathbf{r})}{RT} \right) \\ &\cong \sum_{i=1}^{K} c_i^{bulk} Z_i e_0 - \sum_{i=1}^{K} c_i^{bulk} Z_i^2 e_0^2 \frac{\phi(\mathbf{r})}{RT} \end{aligned} \tag{2.9}$$

where Z_i is the value of each charge and e_0 is the elementary charge. The first term in Eq. 2.9 vanishes because of the electroneutrality of the ionic solution. The expression can be simplified by defining ionic strength I (Eq. 2.10) and the the inverse Debye length κ (Eq. 2.11):

$$I = \frac{1}{2} \sum_{i=1}^{K} c_i^{bulk} Z_i^2 \tag{2.10}$$

$$\kappa^2(\mathbf{r}) = \frac{8\pi e_0^2 I(\mathbf{r})}{RT} \tag{2.11}$$

The linearized Poisson-Boltzmann equation (LPBE) is then:

$$\bigtriangledown \cdot \varepsilon(\mathbf{r}) \bigtriangledown \phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) + \kappa^2 \phi(\mathbf{r}) \tag{2.12}$$

With this formulation at hand it is possible to calculate the three dimensional shape of the electrostatic potential even of large molecules in solution. A prerequisite for this is to know the structure of the molecule. However, analytical solutions of Eq. 2.12 are known only for systems with an ideal spherical shape. This condition is not fulfilled by biological molecules. Therefore the Poisson-Boltzmann equation has to be solved numerically, when dealing with proteins or nucleic acids.

Several methods exist to obtain results of the LPBE. In most cases the finite difference method is applied to solve the LPBE.[14,15] In this method, the space is divided into a regular grid and derivatives are approximated as differences of the electrostatic potentials between neighbor grid points. Alternatively finite element methods,[16,17] boundary element methods[18,19] or multigrid based methods[20] can be applied.

In this work the LPBE is used to calculate electrostatic potentials that were used afterwards to establish the protonation pattern of proteins. The protonation-deprotonation equilibrium of a titratable group in a protein differs from the standard behavior in solution due to electrostatic interactions. The next section describes the individual electrostatic contributions to this shifted equilibria.

2.2. Acid-base behavior in solution and in proteins

The protonation equilibrium of a single titratable group is described by Eq. 2.13:

$$HA \rightleftharpoons A^- + H^+ \tag{2.13}$$

The equilibrium constant K_a is defined as

$$K_a = \frac{[A^-][H^+]}{[HA]} \tag{2.14}$$

The Henderson Hasselbalch equation (Eq. 2.15) describes the pH dependence of the protonation equilibrium:

$$pH = pK_a + log\frac{[A^-]}{[HA]} \tag{2.15}$$

The standard reaction free energy G_a° is related to the pK_a value by the following expression:

$$G_a^\circ = -RT\,pK_a ln10, \tag{2.16}$$

The probability, that a group is protonated is given by

$$\langle x \rangle = \frac{[HA]}{[HA]+[A^-]} \tag{2.17}$$

The pH dependent protonation probability reads then

$$\langle x \rangle = \frac{exp(-ln10(pH - pK_a))}{1+exp(-ln10(pH - pK_a))} \tag{2.18}$$

To understand the different acid-base behavior of a titratable group in aqueous solution and in a protein it is necessary to consider the energetics of the acid-base reactions in different media. If one considers the pH dependence of the protonation energy of one titratable group in a protein only, the thermodynamic cycle in Fig 2.1 contains all relevant energetic contributions. The Figure shows the protonation equilibria in the gas phase, in aqueous solution and in the protein by the horizontal equations. Additionally the free energy values of transferring the protonated or deprotonated species from one medium to another are indicated by the vertical arrows. As the energy of the protonation equilibrium in the protein is not directly accessible, on has to use a thermodynamic cycle, starting from the experimentally or theoretically known energetics of the reaction in solution or gas phase and calculate the energetics of the various solvation processes like transferring the protonated and unprotonated group from aqueous solution to protein.

In a protein in general not only one but typically 100 or more titratable groups have to be treated at the same time. As the interaction between two titratable groups in a protein is also pH dependent, it is necessary to evaluate the acid base equilibria of all titratable groups simultaneously. In such a situation the simple Henderson-Hasselbalch description (Eq. 2.15) of the pH dependent protonation behavior is no longer valid, because it is not possible to ascribe a unique pK_a value to a titratable group and the protonation probability of one group depends also on the protonation state of the other groups. So the whole protonation pattern of the protein has to be established to obtain a meaningful energetic description.

Each titratable group has two possible titration states: protonated or unprotonated. The total number of possible titration states of a protein with N titratable groups sums up to 2^N. A useful description of the protein's protonation state is then given by an N-component vector $\vec{x} = (x_1, x_2, \ldots x_N)$. The components x_μ adopt the value 1 or 0 denoting the protonated and

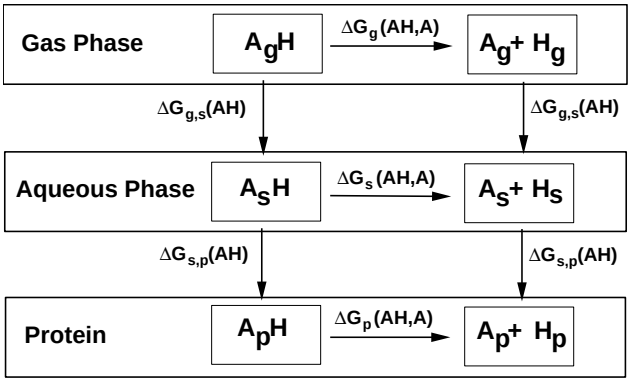

Figure 2.1.: Thermodynamic cycle to calculate the protonation energy of a titratable group in solution and in a protein from gas phase properties. The deprotonation reaction is shown in three media: the gas phase (g), aqueous solution (s) and the protein (p). The vertical arrows denote solvation processes (from gas phase to solution (g,s) and from solution to protein (s,p)). The horizontal arrows denote the protonation processes, which have a different energy in the various environments.

unprotonated state respectively. The protonation probability of the a single group x_μ is given by the following thermodynamic average:

$$\langle x_\mu \rangle = \frac{\sum_{i=1}^{2^N} x_\mu^i exp(-\beta G^i)}{\sum_{i=1}^{2^N} exp(-\beta G^i)} \tag{2.19}$$

where $\beta = (k_B T)^{-1}$ and i runs over all possible protonation states. G^i is the energy of the ith protonation state.

2.3. Protonation state energies from electrostatic potentials

In many cases electrostatic interactions are predominantly responsible for the shift in protonation energy of a titratable group in a protein compared to the corresponding value in aqueous solution. Fig. 2.2 shows a model compound of a titratable group in solution and in the protein. In both environments this model compound has different electrostatic interactions.

The linearized Poisson Boltzmann equation is an ideal means to calculate the electrostatic potentials at the individual titratable groups. These potentials can be used to calculate the energies of in principle all protonation states of a protein. Due to the additivity of the solutions of the LPBE it is possible to treat the individual contributions to the protonation energy of a titratable group independently. The protonation energy of a titratable group in a protein can be considered to consist of four parts: The first contribution is the chemical process to protonate or deprotonate an isolated titratable group. This energy is represented by the pK_a value of the amino acid or an appropriate model compound in aqueous solution via Eq. 2.16. A selection of experimental pK_a^{model} values is given in the appendix E. Transferring this amino acid from aqueous solution into a protein (Fig 2.2) will alter the protonation energy.

As the dielectric properties of the protein differ from that of aqueous solution, the so called Born energy will also be different. This difference, $\Delta\Delta G_\mu^{Born}$, arises from the interaction of the

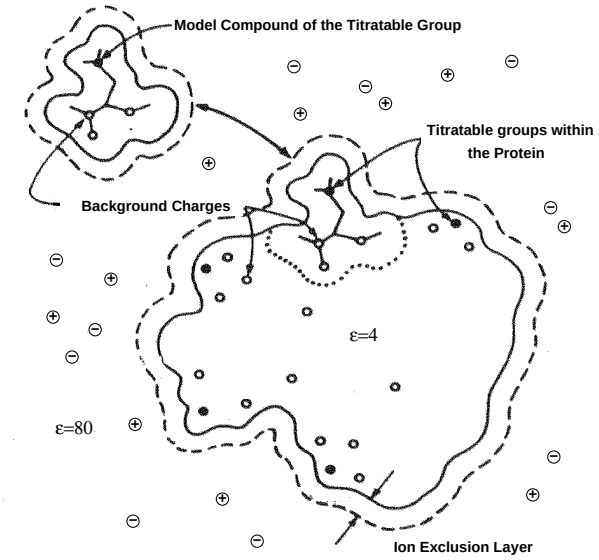

Figure 2.2.: A model compound of a titratable group in aqueous solution and in a protein. When the model compound is transferred from aqueous solution into the protein its electrostatic interactions are altered. The dielectric constant and the ionic strength differ in both media. Moreover, in the protein the titratable group interacts with other titratable groups that maybe charged and with a large number of background charges of all other amino acids of the protein.

partial charges of the titratable group μ with its reaction field, which depends on the dielectric medium. This energy difference can be expressed by the corresponding electrostatic potentials in both media, which are accessible via the LPBE:

$$\Delta\Delta G_{\mu}^{Born} = \frac{1}{2}\sum_{i=1}^{N_{\mu}}\{Q_{i,\mu}^{h}[\phi_p(\mathbf{r}_i;Q_{\mu}^{h}) - \phi_m(\mathbf{r}_i;Q_{\mu}^{h})] - Q_{i,\mu}^{d}[\phi_p(\mathbf{r}_i;Q_{\mu}^{d}) - \phi_m(\mathbf{r}_i;Q_{\mu}^{d})]\} \tag{2.20}$$

The sum runs over the N_{μ} atoms of the titratable group μ that have different charges in the protonated (h)($Q_{i,\mu}^{h}$) and deprotonated (d) ($Q_{i,\mu}^{d}$) form. The terms $\phi_m(\mathbf{r}_i;Q_{\mu}^{h})$, $\phi_m(\mathbf{r}_i;Q_{\mu}^{d})$, $\phi_p(\mathbf{r}_i;Q_{\mu}^{h})$ and $\phi_p(\mathbf{r}_i;Q_{\mu}^{d})$ denote the electrostatic potentials at the position $\mathbf{r}$ of atom i in the protein (p) and in aqueous solution for the model compound (m) respectively. Eq. 2.20 reflects the double difference of the changing born energy of the charged and uncharged group in aqueous solution and in the protein respectively.

The next energy contribution to the shift in the protonation energy arises from the interaction of the charges $Q_{i,\mu}$ of the titratable group μ with charges of the non titrating groups, and with the charges of the uncharged form of all other titratable groups in the protein. Both groups are referred to as the so called "background charges".

$$\Delta\Delta G_{\mu}^{back} = \sum_{i=1}^{N_p} q_i[\phi_p(\mathbf{r}_i;Q_{\mu}^{h}) - \phi_p(\mathbf{r}_i;Q_{\mu}^{d})] - \sum_{i=1}^{N_m} q_i[\phi_m(\mathbf{r}_i;Q_{\mu}^{h}) - \phi_m(\mathbf{r}_i;Q_{\mu}^{d})] \tag{2.21}$$

The first sum in Eq. 2.21 runs over the N_p charges of the protein belonging to atoms of non-titratable groups or to atoms of titratable groups ($\mu \neq \nu$) in their uncharged protonation state. The second summation runs over the N_m charges of the atoms of the model compound, whose charges do not differ in both protonation states. The charges q_i are the ones of non-titratable groups and of titratable groups different from μ which are in their uncharged protonation form.

The pK_a value of a model compound, $\Delta\Delta G_{\mu}^{Born}$ and $\Delta\Delta G_{\mu}^{back}$ can be combined to yield the so-called intrinsic pK_a value:

$$pK_{a,\mu}^{intr} = pK_{a,\mu}^{model} - \frac{\beta}{ln\,10}(\Delta\Delta G_{\mu}^{Born} + \Delta\Delta G_{\mu}^{back}) \tag{2.22}$$

The fourth part responsible for the energy shift of a titratable group that is transferred from aqueous solution into a protein is the interaction between two titratable groups μ and ν in their charged form. It is defined by

$$W_{\mu\nu} = \sum_{i=1}^{N_{\mu}}[Q_{\mu,i}^{h} - Q_{\mu,i}^{d}][\phi_p(\mathbf{r}_i,Q_{\nu}^{h}) - \phi_p(\mathbf{r}_i,Q_{\nu}^{d})] \tag{2.23}$$

With the definition of these four terms the energy of a protonation state n of a protein can be described.

The intrinsic pK_a value is the pK_a that the titratable group μ in a protein would have if all other titratable groups are in their uncharged protonation form. Together with the interaction energy $W_{\mu\nu}$ from Eq. 2.23 the energy of the protonation state n of a protein is

$$G^n = \sum_{\mu=1}^{N}((x_\mu^n - x_\mu^0)\beta^{-1} ln\,10(pH - pK_{a,\mu}^{intr})) + \frac{1}{2}\sum_{\mu=1}^{N}\sum_{\nu=1}^{N}(W_{\mu\nu}(x_\mu^n + z_\mu^0)(x_\nu^n + z_\nu^0)) \quad (2.24)$$

where the x_μ^n are 1 or 0 denoting that group μ is protonated or not and z_μ^0 is the unitless formal charge of the deprotonated form of group μ: -1 for acids and 0 for bases. The sums run over all N titratable groups. The additional x_μ^0 term in the first sum refers to the uncharged protonation state. By this expression it is accomplished that G^n vanishes for the uncharged state, which is accordingly the zero point of the system. The state, where all titratable groups are in their uncharged protonation state is the reference state to which all electrostatic energies refer. Eq. 2.24 is used in principal in all applications, that calculate protonation patterns of proteins by solving the LPBE to get the electrostatic potentials. With the shown set of equations the energies of all 2^N protonation states of a protein are accessible, without solving the LPBE 2^N times. For each titratable group 4 numerical solutions of the LPBE are required: In the protein and as a model compound in solution the electrostatic potentials have to be calculated for the protonated and the unprotonated state. The energy of Eq. 2.24 can be used in Eq. 2.19 to calculate the protonation probability of each titratable group μ. However, the exact summation of the thermodynamic average in Eq. 2.19 is in practice not feasible due to the large number of protonation states. For this reason an approximation method has to be applied as introduced in section 2.5. This is necessary especially, when also several protein conformations are considered. This case is discussed in the following section.

2.4. Protonation energies in different protein conformations

In many cases, it is not sufficient to calculate the protonation pattern of a protein only in one conformation. Especially if one is interested in the protonation behavior over a wider pH range, it is probable that the conformation of the protein varies with the pH, which has to be considered in the titration calculations. Moreover, in principle a protein can also adopt two conformations at one pH: the association between proteins or between a protein and a ligand can be viewed as a conformational change, where the isolated protein and the isolated ligand are one conformation and the the protein-ligand complex is the second conformation. The same holds for the association of two protein monomers. The description of the energy of the protonation states (Eq. 2.24) has to be modified to account also for several conformations:

$$G^{n,l} = \sum_{\mu=1}^{N}((x_\mu^{n,l} - x_\mu^{0,l})\beta^{-1} ln\,10(pH - pK_{a,\mu}^{intr,l})) + \frac{1}{2}\sum_{\mu=1}^{N}\sum_{\nu=1}^{N}(W_{\mu\nu}^{l}(x_\mu^{n,l} + z_\mu^0)(x_\nu^{n,l} + z_\nu^0)) + \Delta G_{conf}^{l} \quad (2.25)$$

where $\Delta G_{conf}^{l} = G_{conf}^{l} - G_{conf}^{r}$ is the energy difference between a fixed but arbitrary reference conformation r and the actual conformation l. Eq. 2.25 requires the calculation of the relative conformational energy ΔG_{conf}^{l} in addition to the terms required for the calculation of the protonation pattern of a single conformation. This relative conformational energy consists of three parts:

$$\Delta G_{conf}^{l} = \Delta G_{S}^{l} + \Delta G_{FF}^{l} + \Delta G_{NE}^{l} \quad (2.26)$$

In Eq. 2.26 ΔG_S^l denotes the electrostatic contribution to the solvation energy difference of conformation l relative to the reference conformation r. ΔG_{NE}^l is the non electrostatic contribution to this solvation energy difference. In ΔG_{FF}^l the Coulomb energy differences between conformation l and r corresponding to a classical molecular mechanics force field are summarized. For the description of a molecular mechanics force field see appendix C. The theoretical basis for Eq. 2.26 is the thermodynamic cycle in Fig 2.3. The calculation of the electrostatic contribution to the

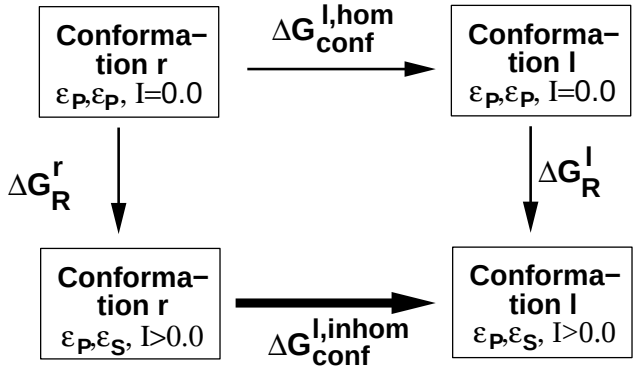

Figure 2.3.: Thermodynamic cycle to calculate the energy difference between the conformations r and l. The conformation r is the reference conformation. Both conformations exist in a homogeneous dielectric medium and in an inhomogeneous dielectric medium. The transfer of each conformation between the media with different dielectric constants is indicated by the vertical arrows. The horizontal arrows show the conformational change in one dielectric medium.

solvation energy of the various conformations can be done from numerical results of the LPBE. It is the energy required to transfer the protein in each conformation from a medium with the dielectric of the protein and an ionic strength equal to zero to a medium with the dielectric constant of water and an ionic strength $\neq$ zero. As in the protein the dielectric constant stays the same, the first state is referred to as homogeneous dielectric and the second case as inhomogeneous dielectric. The corresponding energy is given by

$$\Delta G_R^l = \frac{1}{2}\sum_{i=1}^{N_p} q_{i,p}(\phi_p^{inhom}(\mathbf{r}_i^l, q_p) - \phi_p^{hom}(\mathbf{r}_i^l, q_p)) \tag{2.27}$$

This energy represents the interaction of the protein charges with their own induced reaction field in the corresponding medium. The term $\phi_p^{inhom}(\mathbf{r}_i^l, q_p)$ denotes the electrostatic potential at position $\mathbf{r}_i^l$ in the case of an inhomogeneous dielectric, whereas $\phi_p^{hom}(\mathbf{r}_i^l, q_p)$ is the electrostatic potential in the case of a homogeneous dielectric. In the homogeneous as well as in inhomogeneous dielectric system the electrostatic potentials are calculated from all N_p protein charges $q_{i,p}$, when all titratable groups are in the uncharged protonation state. G_R^r is calculated analogously. The nonpolar contribution ΔG_{NE}^l in Eq.2.26 to the solvation energy of both conformations is assumed to be proportional to the solvent accessible surface:[21, 22, 23, 24]

$$\Delta G_{NE}^l = \gamma(A^l - A^r) \tag{2.28}$$

where A^l and A^r are the solvent accessible surfaces of the reference conformation r and conformation l respectively. The parameter γ has to be determined empirically.[25]

2.5. Monte Carlo sampling of protonation states

As stated before, the exact evaluation of Eq. 2.19 is not possible due to the large number of possible protonation states. For a protein with N titratable groups 2^N terms would have to be summed up. In the case of different conformations this number increases to $L \times 2^N$, where L is the number of different conformations.

To reduce the computational expense several approximate methods have been developed, which all avoid the exact summation of Eq. 2.19: These methods include the Tanford-Roxby approximation,[26] the reduced-site approximation[27] and a hybrid statistical mechanical/Tanford-Roxby approximation.[28,29] An overview and a discussion about these different methods can be found in a review by Ullmann and Knapp.[1]

A further method to calculate the protonation pattern of a protein from Eq. 2.19 without summing up the terms of all states, is the so called Monte Carlo (MC) titration. I used this method in my work and so it will be explained here in more detail. In the MC method, which was developed by Beroza et al.,[30] protonation states are sampled with the probability with which they occur. This procedure is referred to as importance sampling. The protonation probability $\langle x_\mu \rangle$ of a group μ is then obtained by averaging x_μ over all sampled states.

To start the MC sampling the initial protonation state vector $\vec{x}$ is generated randomly. The protonation of one randomly chosen group is then changed. This is defined as one MC move. The energy change corresponding to one MC move is obtained from

$$\Delta G_\mu = \Delta x_\mu [\beta^{-1} ln\, 10 (pH - pK_{a,\mu}^{intr}) + \sum_{\nu=1}^{N} W_{\mu\nu}(x_\nu + z_\nu^0)] \tag{2.29}$$

where $\Delta x_\mu = x_\mu^{new} - x_\mu^{old} = \pm 1$ is the change in the protonation of group μ. The Metropolis criterion[31] is applied then to decide whether this MC move is accepted: if $\Delta G_\mu \leq 0$ the move is always accepted, if $\Delta G_\mu > 0$ the move is accepted with the probability $exp(-\Delta G_\mu \beta)$. When in the statistical average the titration state of each group is changed once, then one MC scan is complete. In the next section the definition of an MC scan will be given more elaborated.

If one applies a large enough number of MC scans and the sampling efficiency (*i. e.* the number of accepted moves) is acceptable, a Boltzmann weighted ensemble is generated.

This method is easily extended to the sampling of titration states of a protein in different conformations: In addition to the titration moves, attempts to switch from one conformation to another have to be inserted. These conformation moves are then accepted or refused by the same criteria as the titration moves. In general one or two conformation moves are made per one MC scan.

2.5.1. Methods to improve the sampling efficiency

Double and triple moves

A problem with the Monte Carlo Metropolis procedure applied to the sampling of protonation states can arise, if two or three titratable groups do not change their protonation state independently but only strongly correlated. The change of the protonation state of just one of the two or three groups will then always be rejected, whereas the simultaneous change of the states of all groups may succeed. For this reason also double and triple moves are made to treat strongly coupled groups correctly (Fig. 2.4). The decision, for which groups double or triple moves are necessary, is made based on the interaction energy of their charged protonation state.

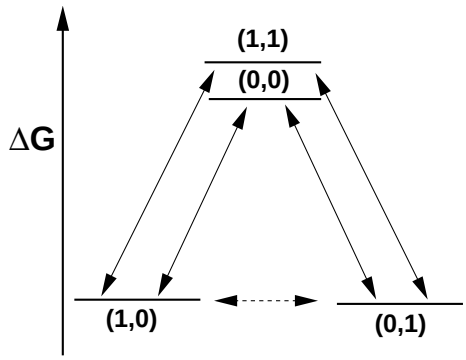

Figure 2.4.: The protonation states (1,0) and (0,1) of two coupled titratable groups are separated by a high energy barrier ((1,1) and (0,0)). A change of the protonation state of just one group would lead to a state with high energy. In a double move the protonation state of both groups is changed at the same time.

One MC scan is defined as the number of MC moves necessary to change the protonation state of each titratable group in a protein in average once. For a protein with N titratable groups this is N. A scan should also include all necessary double and triple moves, such that an MC scan comprises in general more than N moves. The protonation state of all scans are accumulated to obtain the protonation probability $\langle x_\mu \rangle$ of each group μ.

Parallel tempering

Sometimes the system will get stuck in a region of the protonation pattern landscape or at one conformation due to large energy barriers. Only small parts of the total phase space of protonation patterns and conformations are then sampled and the protonation probability of each group as well as the distribution between several conformations cannot be calculated accurately. This problem can be overcome with the parallel tempering algorithm.[32]

In the parallel tempering approach an artificial system built up of N non interacting copies of a molecule is constructed. In our case a molecule is a protein with a large number of possible protonation states that probably exists in several conformations. Each copy exists at a different temperature T. A *state* of the artificial system is described by $S = \{C_1, C_2, \ldots, C_N\}$, where each C_i is a configuration of the real system, that describes the protonation state and the conformational state. The N copies of the system do not interact. Therefore one can assign a weight w to a state S of the compound system:

$$w(S) = exp(-\sum_{i}^{N} \beta_i G(C_i)) \tag{2.30}$$

One can assume that $\beta_1 < \beta_2 < \cdots < \beta_N$. A numerical simulation of the system has to yield the corresponding equilibrium distribution of the total artificial system. This can be accomplished by the following two sets of moves.

1. standard MC moves that effect only the ith copy. These moves are called local moves, because they change one coordinate (the protonation state of one group or the conformation) of the configuration solely in one copy. Since the copies are not interacting, the transition probability of such a local move depends only on the change of the potential energy in

the ith copy. Such local MC moves are accepted or rejected according to the Metropolis criterion with the probability $exp(-\Delta G_i \beta_i)$ (see section 2.5).

2. Exchange of configurations between two copies i and $j = i+1$

$$\begin{aligned} C_i^{new} &= C_j^{old} \\ C_j^{new} &= C_i^{old} \end{aligned} \tag{2.31}$$

Such an exchange is called a global move in the sense that for the ith copy (and the jth copy) the whole configuration changes. Since this move introduces configurational changes in two copies of the molecule it follows from Eq. 2.30 that the exchange is accepted or rejected according to the Metropolis criterion with probability:

$$\begin{aligned} w(S^{old} \rightarrow S^{new}) &= min(1, e^{-\beta_i, G(C_j) - \beta_j G(C_i) + \beta_i, G(C_i) + \beta_j G(C_j)}) \\ &= min(1, e^{(\beta_j - \beta i)(G(C_j) - G(C_i))}) \\ &= min(1, e^{\Delta\beta\Delta G}) \\ &= min(1, e^{\Delta}) \end{aligned} \tag{2.32}$$

where $\Delta = \Delta\beta\Delta G$, $\Delta\beta = \beta_j - \beta_i$ and $\Delta G = G(C_j) - G(C_i)$

It is not necessary to restrict the exchange to pairs of copies associated with neighboring inverse temperatures β_i and β_{i+1}. But this choice would be optimal, since the acceptance ratio will decrease exponentially with the difference $\Delta\beta = \beta_j - \beta_i$. Parallel tempering realizes for each copy of the real system a canonical simulation at corresponding temperature T_i. The exchange of configurations is an improved move which decreases the correlations between the conformations and increases the thermalization of the canonical simulation for each copy. This means that each copy will reach its equilibrium distribution faster than without global moves.