Shaping electrostatic energy computations in proteins: The ClC-type proton-chloride antiporter function

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von

Gernot Kieseritzky

aus Berlin

2011

Die vorliegende Arbeit wurde unter Anleitung von Prof. Dr. E. W. Knapp im Zeitraum 01.10.2005-31.12.2010 am Institut für Chemie / Kristallographie der Freien Universität Berlin im Fachbereich Biologie, Chemie und Pharmazie durchgeführt.

1. Gutachter: Prof. Dr. Ernst-Walter Knapp, Freie Universität Berlin
2. Gutachter: Prof. Dr. Markus Wahl, Freie Universität Berlin

Tag der Disputation: 23.02.2011

Preamble

This is a resume of my scientific work I have been carrying out in the past years focusing on the development of improved methods to perform pK_A calculations and model proton transfer reactions in proteins. My doctoral thesis is based on three articles published in peer-reviewed scientific journals, namely

1. Kieseritzky G and Knapp EW
   Optimizing pK_A computation in proteins with pH adapted conformations
   Proteins: Structure, Function, Bioinformatics 71: 1335-1348, 2008
   http://dx.doi.org/10.1002/prot.21820
2. Kieseritzky G and Knapp EW
   Improved pK_A prediction: combining empirical and semi-microscopic Methods
   Journal of Computational Chemistry 29 (15): 2575-2581, 2008
   http://dx.doi.org/10.1002/jcc.20999
3. Kieseritzky G and Knapp EW
   Charge Transport in the ClC-type Chloride-Proton Anti-porter from *Escherichia coli*.
   The Journal of Biological Chemistry 286 (4): 2976-2986, 2011
   http://dx.doi.org/10.1074/jbc.M110.163246

During the same period I also conducted and participated in related research projects that lead to the following publications:

1. SO, Kieseritzky G, Knapp EW and Labahn A
   Role of tetraheme cytochromes
   in prep.
2. Langen G, Imani Jafargholi, Altincicek B, Kieseritzky G, Kogel K H and Vilcinskas A
   Transgenic expression of gallerimycin, a novel antifungal insect defensin from the greater wax moth Galleria mellonella, confers resistance to pathogenic fungi in tobacco
   Biological Chemistry 387 (5), 549-557, 2006
3. Gámiz-Hernández A P, Kieseritzky G, Galstyan A S, Demir-Kavuk O and Knapp EW
   Understanding Properties of Cofactors in Proteins: Redox Potentials of Synthetic Cytochrome b
   ChemPhysChem 11 (6), 1196-1206, 2010
4. Gámiz-Hernández A P, Kieseritzky G, Ishikita H, and Knapp EW
   Rubredoxin Function: Redox Behavior from Electrostatics
   Journal of Chemical Theory and Computation 7 (3), 742-752, 2011

My thesis first explains the motivation behind this body of work. In the second chapter I give a mini review of the chemistry of acid-base equilibria and the various computational approaches to calculate ionization constants. Furthermore, the results of each paper in the upper list are shortly discussed in the third chapter. Finally, I conclude my dissertational work discussing the general problems affecting the accuracy electrostatic pK_A and redox computations and some ideas on how to solve them.

Contents

List of figures

1 Motivation

Electrostatic interactions play an important role in protein function (Perutz 1978; Warshel 1981; Sharp and Honig 1990). The tertiary structure of a protein for example is defined by specific hydrogen bonds between polar and charged residues, and it is ionization of internal titratable residues that is ultimately responsible for pH dependent denaturation of proteins (Perutz 1978). Consequently, it is possible to compute absolute folding free energies quite accurately by calculating the ionization energies of the protein's titratable groups (Warshel, Sharma et al. 2006; Rocaa, Messera et al. 2007). Frequently, one is interested in structure-function relationships, i.e. understanding protein function based on protein structure (Warshel 1981). Enzymes for example work by lowering the activation energy of a reaction. Usually, the corresponding protein achieves that by placing charged residues close to the active site if the activation complex is charged or contains a strong dipole. If the reaction involves a proton transfer the catalytic site contains one or more titratable residues of which the pK_A is often perturbed by the protein environment compared to its value in aqueous solution.

A classic example is HEWL (Hen Egg-White Lysozyme) catalyzing the hydrolysis of β-1,4-gylcosidic bonds in peptidoglycans of Gram negative bacteria (Warshel and Levitt 1976; Kuramitsu, Ikeda et al. 1977; Post and Karplus 1986; Benkovic and Hammes-Schiffer 2003). Lysozyme contains a glutamate at sequence position 35 (E35) and an aspartate at position 52 (D52). In the first step of the reaction E35 protonates the oxygen between the sugar rings (Figure 1). E35 stands out among the other acidic residues inside lysozyme because of its unusually high pK_A of about 6 although in aqueous solution glutamate has a pK_A of 4.4. Following the protonation step the two carboxylates E35 and D52 stabilize the transient oxocarbonium cation.

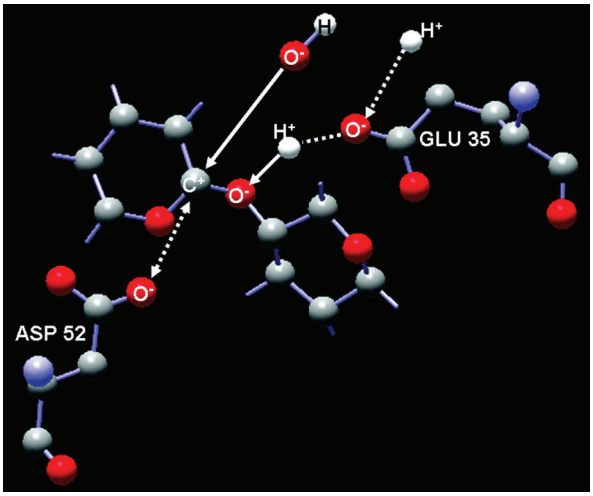

Figure 1 Schematic picture of transition-state stabilization in lysozyme due to electrostatic interactions between the catalytically active residues D52 and E35 and the transient oxocarbenium ion in the saccharide substrate. Taken from (Benkovic and Hammes-Schiffer 2003).

Electrostatic interactions also predominantly control biological electron transfer processes such as those in the light reactions of photosynthesis: In both the bacterial reaction center and the plant photosystems e.g. it has been repeatedly demonstrated that protein electrostatics determines the redox potentials of the relevant cofactors in the electron transfer chain in the same manner it influences the acidity of protonatable groups (Ishikita and Knapp 2003; Ishikita and Knapp 2004; Ishikita and Knapp 2005; Ishikita and Knapp 2005; Ishikita, Loll et al. 2005; Ishikita, Galstyan et al. 2007). This dissertation, too, demonstrates how useful electrostatic pK_A computations are in understanding protein function. The ClC-type proton-chloride anti-porter from *Escherichia coli* (EClC) presented below is able to pump either protons or chlorides against a concentration gradient by utilizing the energy stored in the counter ion's osmotic gradient. A special glutamate inside the transporter undergoes a dramatic pK_A shift upon binding of chloride transforming the usually acidic residue into a base. As a consequence, binding of chloride is associated with protonation of this

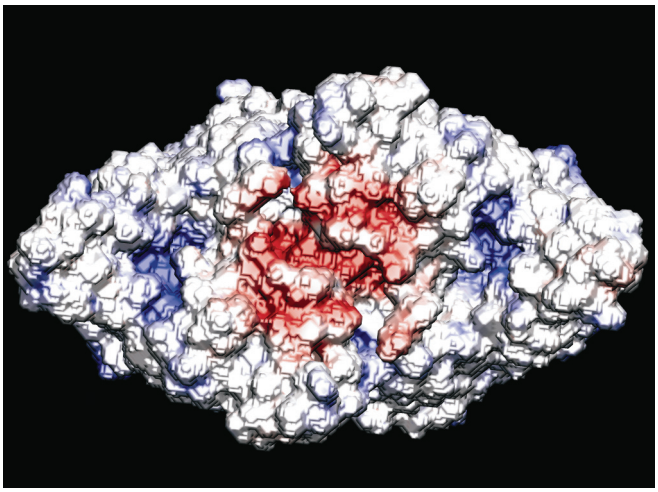

Figure 2 Periplasmic view on the ClC-type chloride-proton antiporter from Escherichia coli. The electrostatic potential as calculated without presence of chloride anions was mapped onto the molecular surface of the protein at isocontour level 10 kT/e_c. The blue spots besides the central red area coincide with the two periplasmic chloride binding sites. This is yet another example of how electrostatic interactions establish a structure-function relationship.

residue and, in turn, deprotonation coincides with the release of chloride. In this case electrostatic pK_A computations successfully probed the mechanism of how EClC is coupling the proton with the chloride transport process.

Computational models that are able to accurately calculate pK_A values or redox potentials not only enable us to identify catalytic residues and understand their properties in the context of the protein structure, they have become essential tools in enzymology helping to clarify the catalytic mechanism underlying enzyme activity. Furthermore, these models have potentials beyond analyzing protein function. Together with better structure prediction tools they could be used in designing synthetic enzymes catalyzing reactions for which no biocatalysts are known to exist because they enable us to predict the effect on pK_As or redox potentials induced by changes in the protein tertiary structure. Early proofs of principle have been published recently (Kaplan and DeGrado 2004; Tynan-Connolly and Nielsson 2007; Jiang, Althoff et al. 2008; Röthlisberger, Khersonsky et al. 2008). Obviously, protein engineering would have many potential applications in biotechnology, medical therapies and industrial processes and will keep electrostatic pK_A and redox potential computations quite relevant for the foreseeable future.

2 Introduction

2.1 What is a pK_A?

Chemists like to characterize compounds using simple, intuitive numbers conferring complex physical and chemical properties at a glance. Many of them can be considered conceptual, i.e. not based on a rigorous theory, such as the atomic radius or electronegativity. Consequently, many different definitions for these parameters have been developed over time and interpretations depend on the definition used to calculate these values. Others are derived from fundamental thermodynamics including the pK_A, a unitless measure of the free energy of deprotonation of an acid, and the redox potential E^0 representing the free energy of oxidation per electronic charge of a redox-active compound. Both quantities can be derived formally from equilibrium properties of the underlying chemical reactions as shall be explained in the following.

2.1.1 Acid-base equilibria

We start by considering the simplified chemical equilibrium of the deprotonation of a generic acid, namely 'HA', in aqueous solution:

$$\mathrm{HA_{(aq)} \rightleftharpoons H^+_{(aq)} + A^-_{(aq)}}.$$

The neutral, protonated acid dissociates a proton and the conjugated anionic base of the acid. Application of the law of mass action yields an expression for the equilibrium constant $\mathrm{K_A}$

$$\mathrm{K_A} = \frac{\mathrm{a(H^+)\,a(A^-)}}{\mathrm{a(HA)}} \approx \frac{\mathrm{c(H^+)\,c(A^-)}}{\mathrm{c(HA)}}$$

where the equilibrium activities $\mathrm{a(H^+), a(A^-)}$ and $\mathrm{a(HA)}$ of the various species were replaced by their corresponding concentrations assuming an ideal solution in which interactions between the solutes are negligible. The value of the equilibrium constant tells us something about the balance point of the equilibrium, i.e. whether there is more of the neutral acid or more of the dissociated species to be found when both forward and backward reactions proceed with equal rate. Correspondingly, "strong acids" for example will cause $\mathrm{K_A}$ to take on values larger than one. On the other hand, "weak acids" will cause $\mathrm{K_A}$ to take on smaller values than one. By taking the negative decadic logarithm of the expression of $\mathrm{K_A}$ we obtain the famous Henderson-Hasselbalch equation:

$$\mathrm{pK_A} = \mathrm{pH} - \log\left(\frac{\mathrm{c(A^-)}}{\mathrm{c(HA)}}\right). \qquad (1)$$

It turns out that the pK_A is nothing but the value of the equilibrium constant of the deprotonation reaction on a unitless logarithmic scale. According to eq. (1) the pK_A of an acid could be determined by a simple measurement of the pH of its solution if the equilibrium concentrations of the acid and its conjugated base were known. Typically, this is the case at the equivalence point of a titration where $\mathrm{c(A^-)} \approx \mathrm{c(HA)}$.

Besides serving as a definition of the pK_A, the Henderson-Hasselbalch equation is a handy formula to calculate the pH of a solution from tabulated values of the pK_A of the dissolved acid and has been used by generations of chemists to adjust the pH of buffered solutions. The Henderson-Hasselbalch equation can be reformulated in terms of the dissociation grade $\alpha = \frac{\mathrm{c(A^-)}}{\mathrm{c(A^-)+c(HA)}}$:

$$pK_A = pH - \log\left(\frac{\alpha}{1-\alpha}\right).$$

In this form it becomes obvious that the Henderson-Hasselbalch equation is invalid for two limiting cases: 1) In case of $\alpha = 1$ for very strong acids which dissociate completely in solution state and 2) in case of $\alpha = 0$ for very weak "acids" which do not dissociate at all. We can transform the last equation once again by substituting $\alpha = 1 - \langle x \rangle$ where $\langle x \rangle$ is the protonation probability, i.e. the degree of protonation of the acid HA:

$$pK_A = pH - \log\left(\frac{1-\langle x \rangle}{\langle x \rangle}\right).$$

We can solve the above equation for $\langle x \rangle$ and convert the decadic into the natural logarithm to obtain an expression for the protonation probability:

$$\langle x \rangle = \frac{e^{-\ln(10)[pH-pK_A]}}{1 + e^{-\ln(10)[pH-pK_A]}}. \qquad (2)$$

A variation of this equation is used to determine the pK_A of a compound if its pH dependent protonation probability can be measured directly in laboratory experiments. To determine the pK_A of titratable residues inside proteins one typically records nuclear magnetic resonance (NMR) spectra correlating backbone hydrogen and nitrogen resonances at different pH values. For many of the spectral peaks one observes a pH dependent shift caused by the titration of residues like aspartate or glutamate. The different chemical shifts corresponding to the same residue are then tabulated together with their corresponding pH values and fitted against eq. (2) treating the pK_A as a free parameter. However, due to presence of more than one titratable residue inside the protein and their mutual interaction, one frequently observes a deviation from the ideal Henderson-Hasselbalch behavior so that the titration curve appears squeezed or stretched along the x-axis. To account for such cooperative effects an additional factor (the Hill coefficient) is introduced inside the exponentials of eq. (2).

In general, the equilibrium constant K_A is related to the standard reaction free energy ΔG^0 by

$$K_A = e^{\frac{-\Delta G^0}{RT}}.$$

It follows the pK_A is related to the standard Gibbs free energy of deprotonation $\Delta G^0(HA \longrightarrow A^-)$:

$$\Delta G^0(HA \longrightarrow A^-) = \ln(10)\ RT\ pK_A \approx 5.74\ \text{kJ mol}^{-1}\ pK_A.$$

The above equation establishes pK_A also as an energy measure with one pK_A unit being equivalent to about 5.74 kJ/mol. Correspondingly, the Gibbs free energy change under non-equilibrium conditions is pH-dependent and given by:

$$\Delta G(HA \longrightarrow A^-) = -\ln(10)\ RT\ (pH - pK_A). \qquad (3)$$

By plugging eq. (3) into eq. (2) one obtains an expression enabling us to compute the protonation probability of a compound from the corresponding protonation energy $\Delta G(A^- \longrightarrow HA) = -\Delta G(HA \longrightarrow A^-)$:

$$\langle x \rangle = \frac{e^{-\frac{\Delta G(A^- \longrightarrow HA)}{RT}}}{1 + e^{-\frac{\Delta G(A^- \longrightarrow HA)}{RT}}}. \qquad (4)$$

This equation is more useful than eq. 2 for our purposes as it is easier to calculate the protonation probability $\langle x \rangle$ than the pK_A value directly, as we shall see.

2.1.2 Redox equilibria

The description of redox reactions can be done in analogy to acid-base equilibria:

$$A_{red,(aq)} \rightleftharpoons A_{ox,(aq)} + e^-_{(aq)}$$

where A_{red} represents the reduced and A_{ox} the oxidized species. Again we can apply the law of mass action and obtain an expression for the equilibrium constant:

$$K_{ox} \approx \frac{c(A_{ox})\, c(e^-)}{c(A_{red})}.$$

This time we evaluate the negative natural logarithm on both sides of the equation

$$-\ln(K_{ox}) = -\ln[c(e^-)] - \ln\left[\frac{c(A_{ox})}{c(A_{red})}\right]$$

and substitute $\ln(K_{ox}) = -\frac{\Delta G^{0\prime}}{RT}$ because the equilibrium constant is related with the standard free energy of oxidation

$$\frac{\Delta G^{0\prime}}{RT} = -\ln[c(e^-)] - \ln\left[\frac{c(A_{ox})}{c(A_{red})}\right].$$

Next, we divide both sides with the Faraday constant F being the molar charge of an electron, and identify $\Delta G^{0\prime}/F$ as the standard redox potential E^0 of the redox pair A_{ox}/A_{red}:

$$E^0 = -\frac{RT}{F}\ln[c(e^-)] - \ln\left[\frac{c(A_{ox})}{c(A_{red})}\right]\frac{RT}{F}.$$

The term $-\frac{RT}{F}\ln[c(e^-)]$ is called solution redox potential and measures the solution electron concentration on a logarithmic scale in analogy to the pH in eq. (1). We substitute $E = -\frac{RT}{F}\ln[c(e^-)]$ and obtain:

$$E^0 = E - \ln\left[\frac{c(A_{ox})}{c(A_{red})}\right]\frac{RT}{F}. \qquad (5)$$

This is the redox-equivalent of the Henderson-Hasselbalch equation and is better known in its rearranged form

$$E = E^0 - \ln\left[\frac{c(A_{red})}{c(A_{ox})}\right]\frac{RT}{F} \qquad (6)$$

under the name of “the Nernst equation”. We can express the reaction quotient in eq. (6) in terms of the oxidation probability $\langle y \rangle = \frac{c(A_{ox})}{c(A_{red})+c(A_{ox})}$:

$$E = E^0 - \ln\left(\frac{1-\langle y \rangle}{\langle y \rangle}\right)\frac{RT}{F}$$

and can then calculate the oxidation probability from the free energy released upon oxidation of $A_{red}$by

$$\langle y \rangle = \frac{e^{-\frac{F\left(E-E^0\right)}{RT}}}{1 + e^{-\frac{F\left(E-E^0\right)}{RT}}} \tag{7}$$

in analogy to eq. (4).

2.2 Electrostatic theory

While there are different ways to model electrostatic interactions in a protein, in this work we will focus on the well-established Poisson-Boltzmann continuum electrostatics description. A short overview of available alternative approaches and a discussion of their comparative advantages and disadvantages over Poisson-Boltzmann theory are given in the next chapter.

2.2.1 The Poisson-Boltzmann equation

As we want to be able to compute pK_A and redox potentials in proteins, we need a reliable mathematical description of biopolymers. In 1923, Debye and Hückel published a theory (Debye and Hückel 1923; Debye and Hückel 1923) that dealt with the energetics of ionic solutions. They found a formula to calculate the activity coefficients of ions which is now known under the name of "Debye-Hückel limiting law". According to the Debye-Hückel model ions in aqueous solutions are surrounded by counter ions and solvent molecules which form "ionic clouds" representing the true charge carrying species in electrolytes. Mathematically, Debye and Hückel combined a Poisson equation describing the interaction between ionic particles with a statistical Boltzmann term describing the charge distribution of ions in the electrolyte and, hence, invented the Poisson-Boltzmann equation (PBE) which they were able to solve analytically. In the following we develop a Poisson-Boltzmann electrostatics based description of proteins dissolved in electrolytes.

Electrostatic theory is ultimately based on Gauß' law being one of the four fundamental Maxwell equations of classical electromagnetism. All Maxwell equations are axiomatic, i.e. they were obtained empirically and can not be derived from first principles. Intuitively, Gauß' law states that a charge is the source of an electric field. Its mathematical formulation

$$\oint \vec{\mathrm{E}}\,\mathrm{d}\vec{\mathrm{A}} = \frac{\mathrm{q}}{\varepsilon_0} \tag{8}$$

relates the electric flux $\phi = \oint \vec{\mathrm{E}}\,\mathrm{d}\vec{\mathrm{A}}$, which is roughly a measure of the number of field lines of the electric field $\vec{\mathrm{E}}$ penetrating a surface $\vec{\mathrm{A}}$, with the total charge q. The fundamental constant ε_0 is called permittivity of free space and vanishes in the cgs (centimeter, gram, seconds) unit system:

$$\oint \vec{\mathrm{E}}\,\mathrm{d}\vec{\mathrm{A}} = 4\pi\,\mathrm{q}\,.$$

The surface integral form can be converted into a volume integral using the divergence theorem:

$$\oint \vec{\mathrm{E}}\,\mathrm{d}\vec{\mathrm{A}} = \int \vec{\nabla}\cdot\vec{\mathrm{E}}\;\mathrm{dV} = 4\pi\,\mathrm{q}.$$

By subsequent differentiation of both sides of the equation with respect to the volume V we can convert the original integral equation into the partial differential equation

$$\vec{\nabla}\cdot\vec{\mathrm{E}} = 4\pi\,\frac{\mathrm{dq}}{\mathrm{dV}}$$

which can be simplified further by substituting the derivative $\frac{\mathrm{dq}}{\mathrm{dV}}$ with a quantity called charge density or charge distribution function $\rho(\vec{\mathrm{r}})$ and $\vec{\mathrm{E}}$ with the negative spatial derivative of the electrostatic potential $-\vec{\nabla}\Phi(\vec{\mathrm{r}})$:

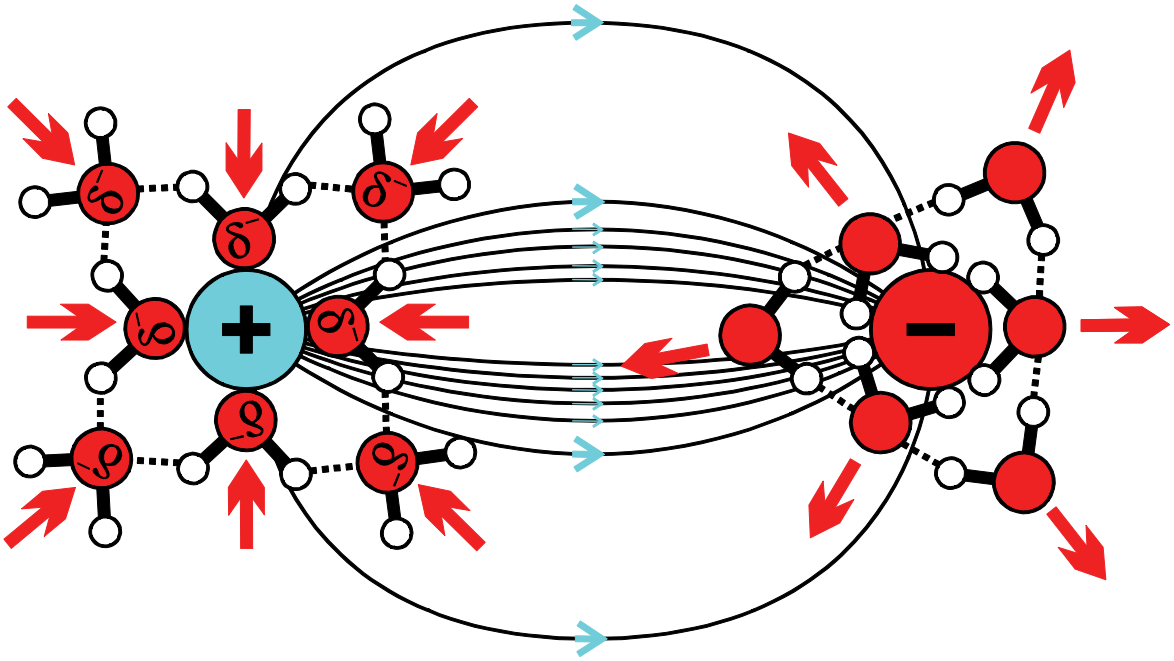

Figure 3 Illustration of the dielectric screening effect. Two point charges are immersed in water and attract each other. Field lines originating at the positive charge are flowing to center of the negative charge. In the environment of the point charges water molecules reorient as to align their dipole field in an anti-parallel fashion and, thus, weakening the total field.

$$\vec{\nabla} \cdot \vec{\nabla}\Phi(\vec{r}) = -4\pi\,\rho(\vec{r})\,. \tag{9}$$

Eq. (9) is a second-order partial differential equation in Cartesian space and runs under the name of Poisson's equation. In dielectric media other than vacuum the electric field in eq. (8) is replaced with the dielectric displacement field $\vec{D} = \varepsilon\,\vec{E}$ due to the presence of induced dipoles in the material reducing the effective field $\vec{E}$ transmitted inside of it. The effect is called "dielectric screening" and is illustrated in Figure 3. The symbol ε is referred to as the "dielectric constant" and usually a scalar value characterizing the polarity of a dielectric medium assuming uniform polarizability of each atom in the material. For instance, in bulk water which is a very polar solvent ε has a relatively high value of 80 at least at temperatures less than or equal to room temperature. If we consider an inhomogeneous medium such as a protein volume surrounded by bulk water, ε becomes a spatial function $\varepsilon(\vec{r})$ that adopts a value of 80 outside of the protein but a significantly smaller value inside the protein. The corresponding Poisson equation is then given by

$$\vec{\nabla}\varepsilon(\vec{r}) \cdot \vec{\nabla}\Phi(\vec{r}) = -4\pi\,\rho(\vec{r}). \tag{10}$$

This equation could already serve as reasonable description of a solvated protein (Figure 4): The polymer is represented by a dielectric volume characterized by $\varepsilon = \varepsilon_p = 4$ embedded in a dielectric continuum with $\varepsilon = \varepsilon_w = 80$. The information about the protein shape is contained in the functional form of $\varepsilon(\vec{r})$. Furthermore, the protein volume is populated by a set of point charges q_i representing the different polar groups, i.e. non-titratable amino-acid residues (G, A, M, S, T, V, N, Q, I, L, P, F and W), the charged titratable amino acid residues (D, E, C, H, K, R, Y), titratable N- and C-termini as well as titratable co-factors, which contribute to the charge density function $\rho(\vec{r}) = \sum_i \delta(\vec{r} - \vec{r}_i)\,q_i$. By solving eq. (10) for $\Phi(\vec{r})$ which can typically be done only numerically one can compute the total electrostatic potential energy of the system. However, that would be the potential energy of the protein in distilled water, as the model so far does not account for the presence of salt. Since protein stability often depend on the presence of salt, this is a crucial point.

By incorporating elements of the Debye-Hückel theory we can extend the charge density on the right side of eq. (10) so that it also includes the density corresponding to mobile ions dissolved together with the protein. In doing so we obtain a non-linear partial differential equation called the Poisson-Boltzmann equation (PBE) given by

$$\vec{\nabla}\varepsilon(\vec{r}) \cdot \vec{\nabla}\Phi(\vec{r}) = -4\pi \left[\rho(\vec{r}) + \kappa^2 \frac{kT}{e_c} \upsilon(\vec{r}) \sinh\left(\frac{e_c \Phi(\vec{r})}{kT}\right)\right] \quad (11)$$

where

$$\kappa = \sqrt{\frac{8\pi N_A e_c^2 I_s}{kT}}$$

is the so-called Debye-Hückel parameter. The symbol $I_s = \frac{1}{2}\sum_i c_i z_i^2$ is the ionic strength of the solution and depends on the concentration c_i of all electrolytes i and their respective formal charge z_i. The unit of κ is that of an inverse length. Correspondingly, $\frac{1}{\kappa} = r_D$ is called the Debye length, i.e. the average radius of the "ion cloud" around any mobile ion picked randomly from the solution. The factor

$$\sinh\left(\frac{e_c \Phi(\vec{r})}{kT}\right) = \frac{1}{2}\left[e^{\frac{e_c \Phi(\vec{r})}{kT}} - e^{\frac{-e_c \Phi(\vec{r})}{kT}}\right]$$

describes a Boltzmann distribution, hence the name, of the mobile ions around the charge distribution in response to the unknown electrostatic potential. The volume exclusion function $\upsilon(\vec{r})$ depends on the exact three-dimensional protein structure, and is equal to one in regions of the continuum which are ion-accessible and zero everywhere else. At low charge densities we can safely replace the PBE with a linear approximation based on the Taylor series expansion of $\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots$ and obtain the linearized PBE (LPBE):

$$\vec{\nabla}\varepsilon(\vec{r}) \cdot \vec{\nabla}\Phi(\vec{r}) + \kappa^2 \upsilon(\vec{r})\, \Phi(\vec{r}) = -4\pi\, \rho(\vec{r})\,. \quad (12)$$

In this work we are limited to the application of the LPBE due to constraints explained below.

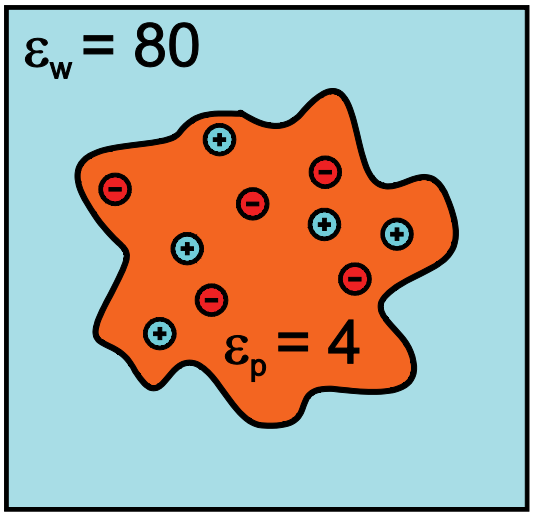

Figure 4 Implicit solvent description of a solvated protein using the Poisson equation. The blue area represents bulk water characterized by $\boldsymbol{\varepsilon = \varepsilon_w = 80}$. The orange area represents the protein volume characterized by $\boldsymbol{\varepsilon = \varepsilon_p = 4}$.

2.2.2 Numerical solution of the LPBE

For regular, spherical geometries such as an ion, the LPBE in equation 12 can be solved analytically:

$$\Phi(r) = \frac{z\, e_c}{4\pi\varepsilon_w\varepsilon_0}\frac{e^{-\kappa r}}{r}. \qquad (13)$$

Eq. (13) gives the (mean) electrostatic potential at distance r from the ion with its total charge $z\, e_c$. However, for complex geometries such as proteins no exact, analytical solutions exist because no analytic form of the dielectric "constant" $\varepsilon(\vec{r})$ can be specified in these cases. Luckily, there are quite a many fast and robust numerical methods to solve three-dimensional partial differential equations (PDEs). Amongst the most widely used recipes is the method of Finite Differences (FD). In this case the problem domain is discretized on a rectangular grid with a fixed grid constant. On a grid differential forms become finite differences and PDEs are reduced to sets of algebraic equations. More specifically, the LPBE (12) becomes a set of linear equations with the total number of unknowns being equal to the number of grid points.

The specific nature of linear equations arising in LPBE problems lends itself to an efficient treatment by iterative procedures such as the Gauss-Seidel or Jacobi method. A version of the former with faster convergence behavior is called "successive over-relaxation" (SOR) and is implemented in the electrostatics library called MEAD (Bashford and Karplus 1990). Gauss-Seidel or SOR can also be combined with the so-called "Multigrid" strategy in which an approximate solution is first obtained on a small, "coarsened" grid used to find a reasonable guess for the solution on the fine grid (the true solution) by interpolation. The "Adaptive-Poisson-Boltzmann Solver" (APBS) implements a Multigrid-FD solver (Baker, Sept et al. 2001) which has been used in this work.

Problems like the LPBE belong to the class of "boundary value" problems because a unique solution only exists if the boundary condition is completely specified, i.e. the solution (or the first derivative) is already known on all borders of the lattice. If the border is far away from the geometric center $\overrightarrow{r_0}$ of the charge distribution $\rho(\vec{r} - \overrightarrow{r_0})$ a reasonable approximation is to assume $\Phi(\vec{r} - \overrightarrow{r_0}) \approx 0$ at the border at the lattice because the electrostatic potential decays geometrically with increasing distance. However, using a "zero" boundary condition requires a very large grid to obtain a numerically stable solution. A better strategy is to obtain the boundary values from eq. (13) treating the protein as a spherical ion with a charge that is equal to its net charge, i.e. the sum of all atomic partial-charges specifying $\rho(\vec{r})$ in eq. (12). In this case, as a rule of thumb, the grid length does not need to be larger than four times the longest length of the protein (Rabenstein 2000).

In pK_A computations, however, one typically needs a resolution between 0.25 and 0.50 Å to obtain values with reasonable accuracy which still leads to large grids despite using the Debye-Hückel boundary condition. The chloride-proton anti-porter EClC which will be discussed later in this dissertation (see the fourth chapter) for example has the dimensions $74 \times 85 \times 61$ Å^3. EClC would require a lattice being at least $85 \times 4 = 340$ Å long. With a grid spacing of 0.50 Å the cubic lattice would comprise $680^3 \approx 300$ million grid points. This is why FD methods usually rely on the grid focusing technique to gradually zoom in on the regions of interest (see Figure 5): the solution is first obtained on a very coarse grid using Debye-Hückel boundary values. Typical resolutions range between 1 and 3 Å in this phase. New boundary values are obtained from the resulting solution for a second FD computation using a medium sized grid that is fully contained within the first one. In the second step the resolution is typically 1 Å and fully contains the protein. Finally, a third high resolution computation is performed in which the grid is centered on the titratable residue of

interest and contains only its immediate environment. In the last focusing step boundary values are initialized from the result of the second computation.

Note that the grid focusing procedure has nothing to do with the Multgrid method described earlier and can even be used together. In fact, APBS combines a Multigrid FD-solver with a grid focusing algorithm. More advanced discretization methods render grid focusing obsolete. The method of Finite Elements (FE) e.g. constructs an adaptive mesh of triangles around the problem domain in such a manner that the size and density of the triangles increases at the critical dielectric boundary. While the mesh spacing is small inside the protein, outsides the mesh can be much coarser and, therefore, saves computing resources for the more important parts of the problem domain. On the other hand, the mesh generation is much more expensive in FE than compared to FD. The Boundary Element (BE) (Boschitsch, Fenley et al. 2002) approach is a method which completely avoids the generation of a mesh. It works by transforming the partial-differential equation (12) into an equivalent surface integral equation that is solved instead. When certain criteria are met BE can be more efficient than both FD and FE but this depends highly on the specific problem.

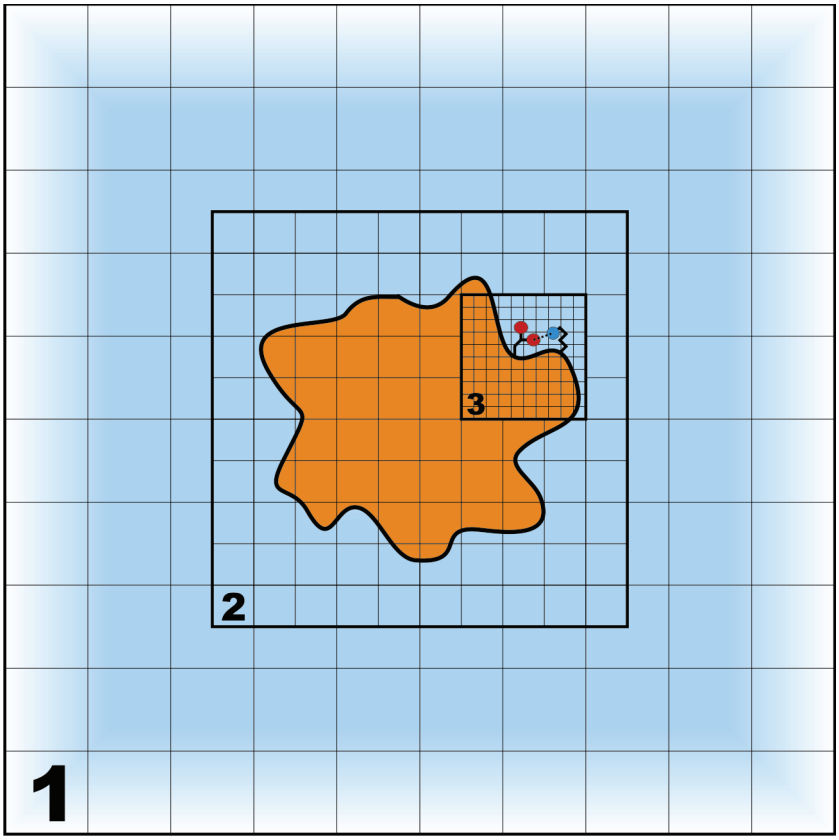

Figure 5 Grid focusing in FD strategies of solving the Poisson-Boltzmann-Equation. The orange object in the center represents the protein. The blue area represents the solvent. Initially, a rough solution is obtained on a coarse grid (#1) which is initialized with an approximate solution on the boundary. The solution obtained on grid #1 is then used as a boundary condition used to calculate the solution on a second grid #2 with smaller grid spacing. Finally, a high resolution grid #3 is centered on a region of interest using solution #2 as a boundary condition to compute a fine-grained solution.

2.3 Electrostatic computation of acid-base and redox equilibria

2.3.1 Calculations using a single conformer

In principle, for any titratable compound in a protein one can compute its pK_A (or redox potential) *ab initio* using the thermodynamic cycle depicted in Figure 6. In that approach, the computation is separated into three distinct steps: 1) calculate the proton affinity (PA) in the gas phase; 2) add to the PA the solvation energy difference between deprotonated and protonated state ($\Delta G^{R,S}_{solv,\mu} - \Delta G^{C,S}_{solv,\mu}$); 3) add to the resulting pK_A value in aqueous solution the protein shift as to obtain the final protein pK_A. This step-wise approach has been shown to work well for a heterogeneous group of organic compounds (Busch and Knapp 2004; Busch and Knapp 2005). However, the approach is very sensitive to the level of quantum mechanical theory used to compute the gas phase properties: high accuracy required a large basis set used in the computation of the proton affinities from electronic energies based on density functional theory (DFT). In the case of electron affinities, even DFT and large basis sets were not sufficient to reproduce experimental redox potentials. The authors had to retreat to much more expensive coupled-cluster calculations to obtain good agreement (Busch and Knapp 2005).

Luckily, the situation is much simpler in case of titratable amino acid residues: For all standard residues, such as D, E, C, H, K, R and Y, we know their corresponding pK_A in solution state from measurements of tri- and pentapeptide models with charge-neutral, protected termini (Nozaki, Tanford et al. 1967; Thurlkill, Grimsley et al. 2006). We could use experimental model pK_A values instead of computing the corresponding values ab-initio if we assume their solution-state pK_A to be independent of their side-chain conformation. This is indeed a reasonable assumption in an isotropic continuum such as bulk water. So, basically, we can skip steps 1) and 2) to proceed directly with step 3) in the computation of protein pK_As. As can be seen from Figure 6 the pK_A is essentially determined by the double difference $\Delta\Delta G^{P}_{\mu} = \Delta G^{C,P}_{\mu} - \Delta G^{R,P}_{\mu}$ of the transfer energies of the charged (C) and neutral (R) state:

$$pK^{P}_{A,\mu} = pK^{M}_{A,\mu} + \frac{\Delta\Delta G^{P}_{\mu}}{RT \ln 10} \qquad (14)$$

$pK^{M}_{A,\mu}$ is the pK_A of the solution state model of the residue type corresponding to the titratable residue μ in focus and essentially contains all non-classical contributions to the pK_A. On the other hand $\Delta\Delta G^{P}_{\mu}$ can be calculated using a continuum-electrostatics description of the protein. It is composed of two major terms:

$$\Delta\Delta G^{P}_{\mu} = \Delta\Delta G^{P}_{solv,\mu} + \Delta\Delta G^{P}_{back,\mu},$$

where $\Delta\Delta G^{P}_{solv,\mu}$ represents desolvation and $\Delta\Delta G^{P}_{back,\mu}$ represents the interaction of the titratable residue with the protein matrix.

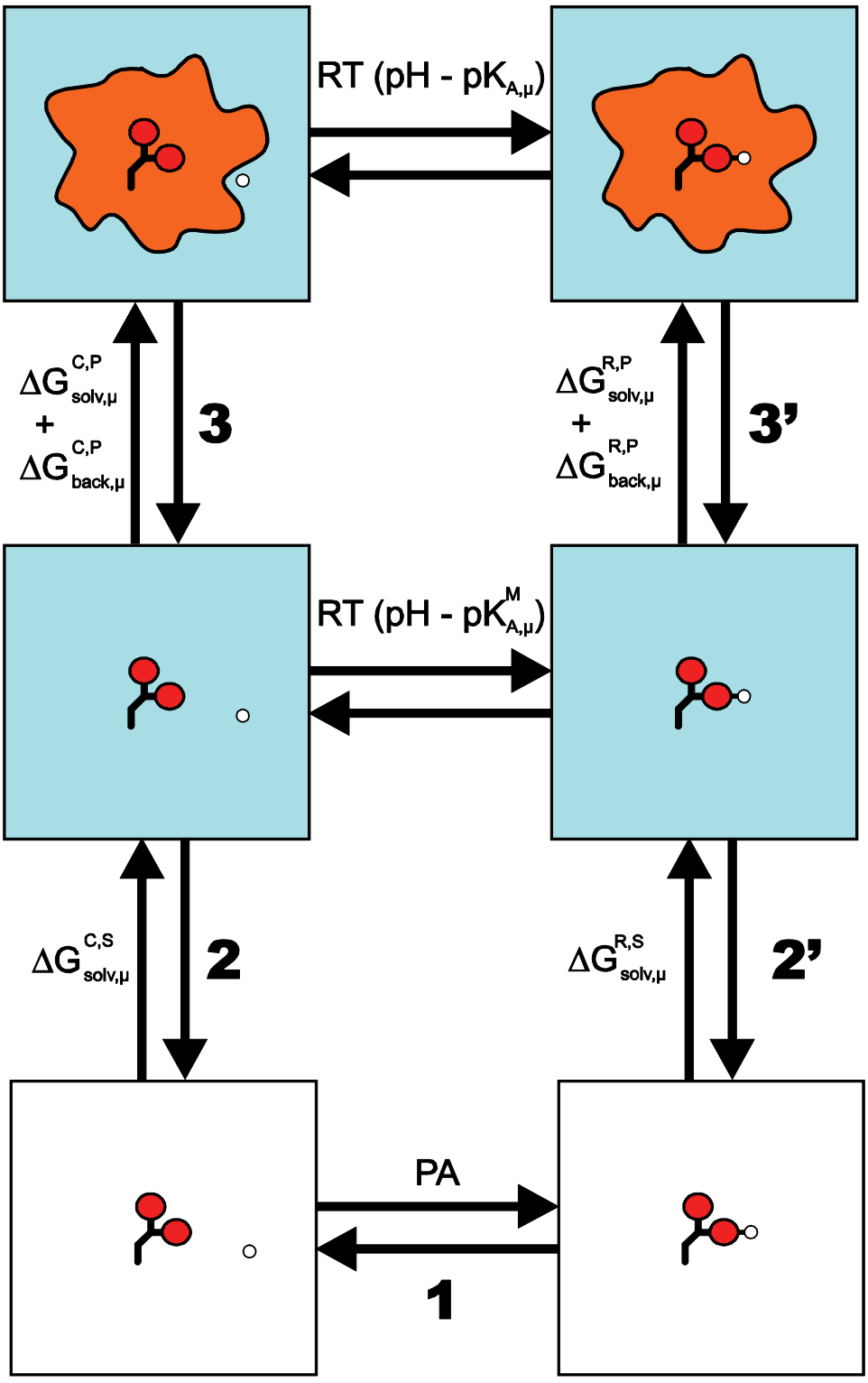

Figure 6 Thermodynamic cycle used to compute protein pK$_A$s from the proton affinity (PA) of a compound µ. While the computation of the PA requires quite a sophisticated level of quantum chemical theory, the protein pK$_A$ can be calculated from classical electrostatics alone. First, the molecule is transferred from gas phase into a solvent (like bulk water) in both protonation states. The difference in the respective solvation energies, $\Delta\mathbf{G}^{\mathbf{C,S}}_{\mathbf{solv},\mu} - \Delta\mathbf{G}^{\mathbf{R,S}}_{\mathbf{solv},\mu}$ ('C' indicates the charged state, 'R' the charge neutral reference state), gives the pK$_A$ shift induced by the solvent. Second, the residue is transferred from aqueous solution into the protein. The resulting desolvation penalties, $\Delta\mathbf{G}^{\mathbf{C,P}}_{\mathbf{solv},\mu}$ and $\Delta\mathbf{G}^{\mathbf{R,P}}_{\mathbf{solv},\mu}$, as well as the interaction with the background charges inside the protein structure, $\Delta\mathbf{G}^{\mathbf{C,P}}_{\mathbf{back},\mu}$ and $\Delta\mathbf{G}^{\mathbf{R,P}}_{\mathbf{back},\mu}$, induce another pK$_A$ shift that, added to the PA and the solvent shift gives the so-called 'intrinsic' pK$_A$ of the residue inside the protein. In practice, however, steps 1) and 2) can be skipped in case of the standard titratable amino acid residues because their pK$_A$ values in aqueous solution are known from experiment.

The solvation energy $\Delta G^{C,P}_{solv,\mu}$ is required to transfer the group μ in its charged state from bulk water into the protein phase because we have to remove the water shell around the molecule and replace it with the solvation shell provided by the protein. In electrostatic theory, the solvation shell is modeled implicitly by the dielectric screening effect describing the polarization of the surrounding dielectric medium and, hence, the formation of induced dipoles around a charge distribution. The potential energy due to the interaction of the solvent with its solute is equivalent to the "self energy"of the solute, i.e. the sum of the products of any point charge inside the solute with the electrostatic potential which it generates itself within the surrounding dielectric medium. Therefore, $\Delta G^{C,P}_{solv,\mu}$ can be computed from the difference in the self energy of the group μ inside the protein and its solution state model (Bashford and Karplus 1990; Ullmann and Knapp 1999):

$$\Delta G^{C,P}_{solv,\mu} = \frac{1}{2}\sum_i Q^{C,M}_{\mu,i}\left[\Phi^{C,P}_{\mu}(\vec{r_i}) - \Phi^{C,M}_{\mu}(\vec{r_i})\right]$$

where $Q^{C,M}_{\mu,i}$ is the partial charge of atom i in the titratable residue μ, $\Phi^{C,P}_{\mu}(\vec{r_i})$ is the electrostatic potential at its position inside the protein, and $\Phi^{C,M}_{\mu}(\vec{r_i})$ is the electrostatic potential in the solution state model. In this work, the partial charges are taken from the CHARMM22 force field where available (MacKerell, Bashford et al. 1998) or calculated ab-initio using the charge fit routines implemented in the program Jaguar (Schrodinger 2009). The factor $\frac{1}{2}$ accounts for the double counting of mutual interactions. Both $\Phi^{C,P}_{\mu}(\vec{r_i})$ and $\Phi^{C,M}_{\mu}(\vec{r_i})$ are calculated from numerical solution of the LPBE (12). Similarly, we compute the solvation energy $\Delta G^{R,P}_{solv,\mu}$ of the neutralized form of residue μ. Following the thermodynamic cycle in Figure 6, we have to subtract $\Delta G^{R,P}_{solv,\mu}$ from $\Delta G^{C,P}_{solv,\mu}$ to obtain $\Delta\Delta G^{P}_{solv,\mu}$ (Bashford and Karplus 1990; Ullmann and Knapp 1999):

$$\begin{aligned}\Delta\Delta G^{P}_{solv,\mu} = &\frac{1}{2}\sum_i Q^{C,M}_{\mu,i}\left[\Phi^{C,P}_{\mu}(\vec{r_i}) - \Phi^{C,M}_{\mu}(\vec{r_i})\right] \\ &-\frac{1}{2}\sum_i Q^{R,M}_{\mu,i}\left[\Phi^{R,P}_{\mu}(\vec{r_i}) - \Phi^{R,M}_{\mu}(\vec{r_i})\right].\end{aligned} \qquad (15)$$

It should be noted that for the computation of the pK_A of even a single titratable group we already have to evaluate four different potential functions and, hence, solve the LPBE four times corresponding to the two charge states of the group times two different dielectric environments. Typically, water is much more polarizable than the protein phase so that $\Delta\Delta G^{P}_{solv,\mu}$ usually adopts positive values. This is why we refer to $\Delta\Delta G^{P}_{solv,\mu}$ by the name of "desolvation penalty".

The protein model also contains permanent dipoles due to the presence of charge neutral, non-titratable residues. Some of them, like S, T, N, Q and the protein backbone, are both potent hydrogen bond donators and acceptors and can quite substantially compensate for desolvation penalty experienced by any titratable group μ transferred into the protein interior. The permanent dipoles are included in the atomic partial charges of the protein taken from the CHARMM22 force field (MacKerell, Bashford et al. 1998) and can be used to compute their contribution to the overall pK_A shift $\Delta\Delta G^{P}_{\mu}$ by the following expression (Bashford and Karplus 1990; Ullmann and Knapp 1999):

$$\Delta\Delta G^{P}_{back,\mu} = \sum_{i=0}^{N_P} q_i^P \left[\Phi_\mu^{C,P}(\vec{r_i}) - \Phi_\mu^{R,P}(\vec{r_i})\right] - \sum_{i=0}^{N_M} q_i^M \left[\Phi_\mu^{C,M}(\vec{r_i}) - \Phi_\mu^{R,M}(\vec{r_i})\right]. \quad (16)$$

The first sum in eq. (16) runs over all background charges q_i^P of the protein and aggregates the potential energy difference between the charged and neutral state of the titratable group μ. The second sum runs over all background charges of the solution state model, i.e. atomic partial charges which do not change between the different protonation states. The electrostatic potentials in eq. (16) are the same as used in eq. (15).

Now we can plug eqs. (15) and (16) into (14) and obtain for the protein pK_A:

$$pK^P_{A,\mu} = pK^M_{A,\mu} + \left(\Delta\Delta G^P_{solv,\mu} + \Delta\Delta G^P_{back,\mu}\right)/RT \ln 10. \quad (17)$$

Using this equation one could in principle compute the pK_A of any titratable group μ inside of a protein directly if it were the only one in the entire protein. This, however, is rarely the case even for small proteins. Usually, different titratable residues would interact with one another in a protein perturbing each other's pK_A value again. Consider for example an ion pair like glutamate and arginine. The glutamate's pK_A will be very low as long as arginine is ionized because the positive charge on arginine stabilizes the deprotonated, ionized state of glutamate even if the desolvation penalty $\Delta\Delta G^P_{solv,\mu}$ is large and $\Delta\Delta G^P_{back,\mu}$ close to zero. Similarly, the arginine's pK_A typically will be large in this situation due to the presence of the negative charge on the ionized glutamate stabilizing its proton. Hence, the precise pK_A values of these residues will depend on each other. In general, determination of any group's pK_A inside a protein requires knowledge of the protonation states of all remaining titratable residues due to mutual electrostatic interactions.

For this reason we first compute the so-called "intrinsic" pK_A ($pK^{intr}_{A,\mu}$) of every protein residue μ according to eq. (17) neglecting the influence of other titratable residues in the protein at this point. Then we calculate the matrix **W** of all pair-wise interactions between the charged states of the residues μ and ν containing the following elements (Bashford and Karplus 1990; Ullmann and Knapp 1999):

$$\Delta\Delta W_{\mu\nu} = \sum_{i=0}^{N_\mu} \left(q^C_{\mu,i} - q^R_{\mu,i}\right)\left[\Phi_\nu^{C,P}(\vec{r_i}) - \Phi_\nu^{R,P}(\vec{r_i})\right] \quad (18)$$

where the sum runs over all partial charges $q^C_{\mu,i}$ and $q^R_{\mu,i}$ of the titratable residue μ. The diagonal elements of **W** are zero by definition as the self-energy is already contained in the intrinsic pK_A. The symbols $\Phi_\nu^{C,P}(\vec{r_i})$ and $\Phi_\nu^{R,P}(\vec{r_i})$ are the electrostatic potentials generated by the charge distributions of residue ν corresponding to its charged and neutralized state, respectively. From its intrinsic pK_A and pair-wise interactions we can calculate the electrostatic energy required to switch residue μ from the reference state x^R_μ into the protonation state x_μ:

$$\Delta G = \left(x_\mu - x^R_\mu\right) \ln(10)\ RT\ \left(pH - pK^{intr}_{A,\mu}\right) + \sum_{\nu\neq\mu} \delta_{\mu\nu}\ \Delta\Delta W_{\mu\nu}.$$

The pH dependent term is equivalent to either the protonation or deprotonation energy as defined by eq. (3) or simply zero depending on the value of $\left(x_\mu - x_\mu^R\right) \in \{-1,0,1\}$. The symbol $x_\mu \in \{0,1\}$ represents the target protonation state where by definition the value of 0 corresponds to the deprotonated state and 1 to the protonated state. The symbol $x_\mu^R \in \{0,1\}$ defines the charge-neutral "reference" protonation state of the titratable group μ. Its value depends on whether the group μ is an acid ($x_\mu^R = 1$) or base ($x_\mu^R = 0$). The symbol

$$\delta_{\mu\nu} = \left|\left(x_\mu - x_\mu^R\right)\left(x_\nu - x_\nu^R\right)\right|$$

returns zero if any of the residues μ and ν are in their reference state because $\Delta\Delta W_{\mu\nu}$ is not defined in this case. Hence, the energy ΔG is zero if the target state is equal to the reference state. ΔG turns out to be the deprotonation energy in case μ is an acid like e.g. aspartic acid and the protonation energy if μ is a basic residue like e.g. arginine.

For a protein with many titratable residues we can now calculate the total electrostatic energy of any protonation state n by adding the individual contributions (Bashford and Karplus 1990; You and Bashford 1995; Ullmann and Knapp 1999):

$$\Delta G^n = \sum_{\mu=1}^{N}\left(x_\mu^n - x_\mu^R\right)\ln(10)\ RT\left(pH - pK_{A,\mu}^{intr}\right) + \sum_{\mu=1}^{N}\sum_{\mu\neq\nu}^{N}\delta_{\mu\nu}\,\Delta\Delta W_{\mu\nu}\,. \qquad (19)$$

The sums run over all titratable groups N of the protein. If all groups are in their charge-neutral reference state x_μ^R the value of ΔG^n becomes zero. If the protein also contains K redox-active groups we can easily generalize this equation to include those as well (Ullmann and Knapp 1999):

$$\begin{aligned}\Delta G^n = &\sum_{\mu=1}^{N}\left(x_\mu^n - x_\mu^R\right)\ln(10)\ RT\left(pH - pK_{A,\mu}^{intr}\right)\\ &-\sum_{\mu=0}^{K}\left(y_\mu^n - y_\mu^R\right)F\left(E - E_\mu^{intr}\right)\\ &+\sum_{\mu=1}^{N+K}\sum_{\mu\neq\nu}^{N+K}\delta_{\mu\nu}\,\Delta\Delta W_{\mu\nu}\end{aligned} \qquad (20)$$

where the middle sum runs over all redox groups and aggregates their corresponding oxidation or reduction energies depending on their reference state y_μ^R which can be either 1 for the oxidized or 0 for the reduced state. The symbol E_μ^{intr} represents the intrinsic redox potential being the hypothetical redox potential if all other titratable groups are in their neutral reference state whether that being protonated, deprotonated, oxidized or reduced. The last term now runs over all pairs of the $N + K$ protonatable and redox-active groups inside the protein and is zero if either μ or ν are in their reference state.

In eqs. (19) and (20) we assume that we can simply add the individual group's contribution to obtain the total potential energy. This requires the underlying electrostatic potentials to be additive which is only guaranteed as long as the linearized approximation of the PBE is valid. For high ionic strengths and charge densities the LPBE description breaks down. This is typically the case if the systems contains multivalent ions like e.g. magnesium cations, phosphate or highly charged macromolecules

like DNA or RNA. Under these conditions one would have to compute the electrostatic potential explicitly for every protonation state of the entire system to accurately calculate the corresponding potential energy.

Finally, we have all the information to compute the protonation/oxidation probilities of all titratable residues in the protein for a given pH and redox potential (Ullmann and Knapp 1999):

$$\langle x_\mu \rangle = \frac{\sum_{n=1}^{2^{N+K}} x_\mu^n e^{-\frac{\Delta G^n}{RT}}}{\sum_{n=1}^{2^{N+K}} e^{-\frac{\Delta G^n}{RT}}}. \tag{21}$$

This is basically an extension of eq. (4) and represents a thermodynamic average of x_μ^n over all 2^{N+K} protein protonation states with $N + K$ being the total number of titratable residues. Population of the individual states follow a Boltzmann-distribution, i.e. high energy states do not significantly contribute to the total average. This is the reason why we can evaluate the sums in eq. (21) more efficiently using the Metropolis-Monte-Carlo (MMC) algorithm (Metropolis, Rosenbluth et al. 1953) which basically works as follows:

1. Randomly choose a titratable group μ.
2. Switch its current protonation state and compute the corresponding energy difference $\Delta\Delta G^n$.
3. If $\Delta\Delta G^n \leq 0$ keep the move and update running average of the protonation state of μ, else keep the move with probability $P = e^{\frac{\Delta\Delta G^n}{RT}}$.
4. Go to step 2.

Typically, after a couple of thousand Monte-Carlo scans the algorithm converges.

2.3.2 Calculations with multiple conformations

So far we have assumed the protein structure remains invariant with changing pH. This, of course, is not appropriate, since the protein environment around ionizable residues is likely to relax following a change of their protonation state. The conformational changes could be limited to a reorganization of the local hydrogen bonding pattern or as large as a global unfolding of the entire tertiary structure. While it is very difficult to predict such (de)protonation triggered conformational reactions, it is actually easy to modify electrostatic theory to include the conformational relaxation of the protein once alternative conformations are known (Ullmann and Knapp 1999):

$$\begin{aligned} G^{n,l} = & \sum_{\mu=1}^{N} \left(x_\mu^n - x_\mu^R\right) \ln(10)\, RT \left(pH - pK_{A,\mu}^{intr,l}\right) \\ & - \sum_{\mu=0}^{K} \left(y_\mu^{nl} - y_\mu^R\right) F\left(E - E_\mu^{intr,l}\right) \\ & + \sum_{\mu=1}^{N+K} \sum_{\mu\neq\nu}^{N+K} \delta_{\mu\nu}\, \Delta\Delta W_{\mu\nu}^l \\ & + G_{conf}^l \end{aligned} \tag{22}$$

The differences between this equation and eq. (20) are only subtle. The total electrostatic energy $G^{n,l}$ is now a function of both the protonation state n and conformation state l. Now, we need to re-compute intrinsic pK_A values of the same residues $pK_{A,\mu}^{intr,l}$ and the matrix $\mathbf{W^l}$ for each conformation l.

This is because structural changes can affect the energetics of all titratable residues even if their coordinates remain the same. Occurring at the surface they modify the dielectric boundary between protein and solvent changing the self-energies defined in eq. (15). Similarly, the interaction with the protein background is modified if non-titratable residues adopt a different conformation. Furthermore, there is an additional correction term G_{conf}^{l} which is necessary because the inclusion of additional conformations resets the zero-energy point. By definition the energy total of eq. (20) becomes zero if all titratable residues are in their charge-neutral reference state. With more than one conformation, however, eq. (22) evaluates to zero only if the protein is residing in an arbitrarily chosen reference conformation state $l = R$ where by definition $G_{conf}^{R} = 0$. In the remaining conformation energies are non-zero due to the interaction of the reference states in a modified protein background. How the values of G_{conf}^{l} are computed shall not be the focus here and will be explained in a later chapter.

The protonation (or oxidation) probabilities are now obtained by the thermodynamic average over both the universe of protein protonation/redox patterns $N + K$ and conformations L (Ullmann and Knapp 1999):

$$\langle x_\mu \rangle = \frac{\sum_{l=1}^{L} \sum_{n=1}^{2^{N+K}} x_\mu^n e^{-\frac{\Delta G^{n,l}}{RT}}}{\sum_{l=1}^{L} \sum_{n=1}^{2^{N+K}} e^{-\frac{\Delta G^{n,l}}{RT}}} \qquad (23)$$

The sums are approximated by a modified MMC procedure. Additional conformation moves randomly switch the protein between the pre-calculated conformations. To enhance the sampling efficiency the “parallel-tempering” method (Earl and Deem 2005) needs to be applied in which several MMC calculations (called replicas) are run in parallel at different temperatures and special tempering moves randomly swap configurations between the replicas (replica exchange).

The approach to treat protein flexibility described above is in principle exact and does not include any additional approximations. However, it suffers from a combinatorial explosion making it very expensive when protein conformational changes are small and localized. Consider a protein with, say, 100 residues whose side chains are flexible. If each residue can adopt only five different rotamers, the implied state space would already contain $5^{100} \approx 10^{70}$ different conformations. Therefore, eq. (22) would require the computation of 10^{70} pairwise interaction matrices $\mathbf{W^l}$ and intrinsic pK_A values $pK_{A,\mu}^{intr,l}$. Luckily, usually only a limited number of conformations, i.e. those with a low energy, would contribute significantly to the average in eq. (23). So a good approximation would be to identify the relevant metastable states from the conformation space and use only those in the computation of protonation probabilities. This is the strategy we adopted for Karlsberg$^+$, the program I have developed for my dissertation and which will be explained in a dedicated chapter.

If the conformational changes of a residue are small (or if the residue is buried) then its effect on the remaining residues, i.e. on their intrinsic pK_A and their pair-wise interactions, due to the change of the dielectric boundary is negligible. In this case one could add all its rotamers to the protein volume and describe the conformational transition just as a change of the atomic partial charges, i.e. the charge corresponding to one conformer is turned off and those of another turned on. This way one does not need to compute the electrostatic energies for every possible combination of rotamers in the protein. Instead, for each rotamer only an additional intrinsic energy (in analogy to the intrinsic

pK_A) and an extra row in $\mathbf{W}$ are computed. The total electrostatic energy of any combination of rotamers can then be calculated from the additive, pairwise contributions in analogy to how the protein protonation pattern is accounted for. In an early adoption (Beroza and Case 1996) of this method – which I like to call the "local conformer" approach in contrast to the "global conformer" method uses by Karlsberg$^+$ – the ionization constants in Lysozyme and Myoglobin were computed by adding to the crystal structural conformation just one extra conformer for every titratable residue. The conformers were created by maximizing the respective residue's solvent accessibility. Inclusion of these "extended" side chain conformers improved the agreement with available experimental pK_A values from Lysozyme on average by 0.4 pK units. The "local conformer" method was later extended to use much more conformers obtained from rotamer libraries or generated by geometry optimization and integrated in an automatic pK_A computation framework called MCCE ("multiconformation continuum electrostatics") (Georgescu, Alexov et al. 2002). The program achieves a comparably high accuracy of about 1 pK unit on average which has been assessed using a large pK_A benchmark data set containing 166 residues. However, the addition of a large number of conformers (up to 10 conformer per side chain of titratable and also non-titratable residues) created the concern that the dielectric boundary might be ill-defined in their calculations (Kieseritzky and Knapp 2007) which might explain the need of rescaling energies due to very strong interactions in order to obtain this level of accuracy (Georgescu, Alexov et al. 2002). Recently, this has been addressed by the MCCE developers in a publication describing the new version of their software called MCCE2 (Huang, Du et al. 2010). Essentially, they replaced the ad-hoc scaling function in the previous version with a scheme in which they estimate the error induced by the artificial boundary: MCCE2 compares the pair wise matrix elements $\Delta\Delta W_{\mu\nu}$ defined similar to eq. (18) computed using the exact dielectric boundary of the crystal structure to the corresponding values obtained after adding all local conformers. The ratio between these values is then used to rescale the pair wise interactions between the conformers. Including this boundary correction MCCE2 achieves an average error of 0.9 pK units, but without a significantly higher value of 1.47 .

2.3.3 The protein dielectric constant

There has been a debate on the appropriate value of the value of the dielectric constant inside the protein volume, although typical values of ε_p range between 2 and 20 (Schutz and Warshel 2001). According to the Clausius-Mossotti relation the dielectric constant ε can be calculated from the molecular polarizability α:

$$\frac{\varepsilon - 1}{\varepsilon + 1} = \frac{N_A \rho_m}{3M} \alpha \qquad (24)$$

where α measures the dipole moment induced in a molecule with molecular weight M by an external electric field, ρ_m is the mass density in the material and N_A Avogadro's constant. For non-polar organic liquids ε typically takes a value of about 2 due to the electrons rearranging in the field. This effect is called electronic polarizability and is, of course, also present in proteins. This is why $\varepsilon_p = 2$ is usually considered a minimum value in protein electrostatics. However, a protein also contains charges and permanent dipoles on its partially mobile amino acid residues which reorient in an externally applied field causing the observed dielectric constant to be higher than the minimum value. The content of charges and dipoles varies between different proteins and even between different regions of the same the protein. So does the value of ε_p, although microscopic molecular dynamics (MD) simulations of proteins suggest the average value to be close to 4 in their hydrophobic core and not larger than 20 in its polar regions (Simonson and Brooks 1996).

It has been argued that the value of ε_p depends not only on physical properties of proteins but more importantly on the nature of the physical model used to describe proteins (Schutz and Warshel 2001). According to this argument protein models can be categorized roughly into microscopic and macroscopic descriptions. Macroscopic models do not explicitly simulate physical effects but incorporate them in the form of phenomenological parameters. Fully microscopic simulations on the other hand do not include the dielectric constant in their equations of motion. Rather, the dielectric constant can be calculated from a time average of the atomic positions in the simulation data. Between fully microscopic and macroscopic theories there are ad-hoc models that fall in between these extremes. Removing some of the degrees of freedom in a molecular simulation typically requires the introduction of a scaling factor or fit parameter compensating for the errors introduced by the approximation. The scaling factor (let's call him ε_p) needs to be larger as more aspects of the protein are ignored by the model – in other words: the more "macroscopic" a theory becomes. The different continuum electrostatic models based on eq. (12) vary in their degree of "macroscopicness": A simple single-conformation model in which the protein structure is treated completely static requires ε_p to be set to a value significantly larger than 4 (8-20) to keep the error in a pK_A computation down. The introduction of different protein conformations allows decreasing the value of the factor making the corresponding model more "microscopic". If the oldest available continuum electrostatic models (Tanford and Kirkwood 1957; Tanford and Roxby 1972) are assessed using a concurrent pK_A benchmark, it turns out that ε_p has to be set to values as high as 80 to obtain average errors that are comparable to modern programs (Schutz and Warshel 2001). The old models (incorrectly) assumed all proteins to have a spherical shape with their ionizable residues equally distributed on its surface.

2.4 Approaches not based on Continuum Electrostatics

2.4.1 All-atom methods

In principle, pK_A values can be computed from fully microscopic MD simulations of the protein and the surrounding solvent if combined with a scheme to improve the sampling efficiency. The first attempt made in this direction was made almost 25 years ago (Warshel, Sussman et al. 1986) when the authors succeeded in computing the free energy of ionization for Aspartate 3 and Glutamate 7 in the pancreas trypsine inhibitor using an "umbrella sampling" method. In umbrella sampling the potential energy surface in MMC or MD simulations is modified in a manner to reduce the barriers between local minima increasing the number of configurations visited (Torrie and Valleau 1977). It is sometimes called "biased" sampling because the algorithm is actually forced to visit states which it would not under normal conditions. In the study, partial charges corresponding to the protonated states of the residue in focus were gradually turned off and in the following replaced with charges representing the deprotonated states. For each step in this charging cycle an MD trajectory was calculated. Integration over the resulting trajectories yielded the free energy of ionization of the residue which had undergone the charging cycle. In constraining the corresponding charges the authors effectively introduced an "infinite" bias in their umbrella sampling which is why their protocol is formally equivalent to the "thermodynamic integration" method of calculating free energies (Kästner and Thiel 2005).

The approach described above neglects the quantum chemical contributions to the pK_A because deprotonation also requires cleavage of a covalent bond between the proton and the titratable residue. Hybrid quantum-classical MD simulations have been introduced to address this issue (Warshel and Levitt 1976; Aaqvist and Warshel 1993; Lin and Truhlar 2007; Senn and Thiel 2009). In these models the potential energy of titratable residues and the proton are calculated *ab initio* based on molecular orbitals (MO) or by hybrid atomic orbitals provided by an empirical valence bond theory (EVB) (Warshel and Levitt 1976; Aaqvist and Warshel 1993). While being very expensive, hybrid simulations allow the study of the kinetics of the underlying proton-transfer process in a rigorous fashion. However, if one is only interested in the pK_A values, i.e. simple free energy differences, the results of classical simulations can be corrected by adding a constant value: the model pK_A introduced in eq. (14) (Warshel, Sussman et al. 1986).

Another limitation of all-atom models described so far is that they require an enormous numerical effort to rigorously incorporate the pair-wise coupling between the protein titratable groups as given in eq. (18). Usually, one assumes a fixed protonation state for the other residues when calculating the ionization energy of a specific group of interest. This is why, all-atom simulations can not provide pH-dependent protonation probabilities and, correspondingly, titration curves. "Constant-pH MD" (CPHMD) simulations have been invented to circumvent this problem: It adds degrees of freedom in the Hamiltonian describing transitions between protonation states of the protein (Lee, Freddie R Salsbury et al. 2004). These "titration coordinates" make it possible to sample both from the ensemble of protein conformations and the ensemble protonation states in the same simulation. CPHMD is actually a special case of λ-dynamics which itself is an umbrella sampling method originally applied to the calculation of binding free energies of ligands (Knight and Charles L Brooks 2009). In CPHMD the completely virtual λ-particles are governed by pH-dependent bias potentials that have been carefully calibrated to reproduce the ionization behavior of isolated model compounds of the major titratable residues appearing in proteins. A variant of CPHMD (Baptista, Teixeira et al. 2002)

does not use continuous titration coordinates but instead works by stopping a classic MD simulation in periodic intervals to perform a continuum electrostatics computation as described in the previous chapter on the current coordinate snapshot. The resulting energies are used to perform a *single* MMC move trying to change the protonation state of the protein after which the MD simulation is continued.

In typical explicit water MD simulations most of the CPU time is spent in propagating the degrees of freedom due to the solvent atoms which usually comprise more than 50% of all particles simulated in the system. Therefore, practical application of constant pH methods requires either massive parallelization or the use of an implicit solvent model such as Generalized-Born (see below) which represents the solvent only in the form of a potential of mean force. Despite of these optimizations the computational demands are still huge compared to continuum electrostatics models so applications were so far limited to much smaller systems. In addition, so far no systematic benchmark has been published proving application of all-atoms models including hybrid-classical MD and CMDPH results in much better accuracy that could justify the effort.

2.4.2 PDLD

Long before numerical methods of solving the LPBE became popular Warshel and co-workers developed a method based on point dipoles called “Protein Dipole Langevin Dipole” (PDLD) (Warshel and Levitt 1976; Warshel and Russell 1984). In this method each protein atom is represented by a point dipole accounting for its electronic polarizability. In this model presence of protein atomic partial charges induce a dipole moment in the surrounding dielectric medium which itself contributes to the polarization of the medium. The solvent is represented by permanent, freely rotatable Langevin dipoles arranged on a lattice around the protein. Their orientation is governed by the local field due to the protein charges. Over the years PDLD has been refined to improve its computational efficiency by scaling of the resulting solvation energies (PDLD/S) and also to include protein structure relaxation (PDLD/S-LRA) using all-atom molecular dynamics simulations (Sham, Chu et al. 1997; Warshel, Sharma et al. 2006). Although yielding similar results, PDLD is not formally equivalent to treatments based on the Poisson equation. Rather, it is more closely related to explicit all-atom descriptions and therefore considered “semi-microscopic” theory (Schutz and Warshel 2001). Although PDLD has been successfully applied to all kinds of biophysical problems in enzymology and protein chemistry (Burykin and Warshel 2003; Burykin and Warshel 2004; Olsson, Sharma et al. 2005; Sharma, Chu et al. 2007; Warshel, Sharma et al. 2007; Pisliakov, Sharma et al. 2008), it has its shortcomings. First, it can not account for the effect of salts diluted in the solvent which elude a point dipole description. Second, while the PDLD method can be used to compute intrinsic pK_A values in a consistent manner similar to equation 16, it turns out to be very expensive if used to evaluate the pair wise electrostatic interactions between the various titratable groups inside the protein to compute the elements of the $\mathbf{W}$ matrix introduced in eq. (18). Therefore, Warshel and his co-workers usually retreat to Coulomb’s law using a high effective dielectric constant or a distance-dependent dielectric constant (Sham, Chu et al. 1997).

2.4.3 Generalized-Born methods

Generalized-Born (GB) theory essentially provides an analytic approximation to the numerical solution of the Poisson equation 10 for any complex molecular geometry. Many implementations of GB theory exist which differ in speed, accuracy and additional features. Many GB methods, but not all of them, e.g. have been extended to include salt effects, therefore, also provide solutions to eq. 12. The basic principles underlying GB theory, however, are always the same. The starting point of

any GB model is the basic Born formula for the solvation energy of a spherical ion with radius a and charge q:

$$\Delta G_{born} = -\frac{q^2}{2a}\left(\frac{1}{\varepsilon_p}-\frac{1}{\varepsilon_w}\right).$$

The Born formula can be formally derived from the analytical solution of eq. (10) in a spherical geometry which, however, is not available for complex, non-regular protein geometries. GB theory starts in considering the protein as a set of individual ions (its atoms) with finite radius (the atomic radii). The total solvation energy due to the transfer of such a polyion from a dielectric medium characterized by ε_p into another characterized by ε_w could be approximated by the some of the individual Born energies plus the differences between the Coloumb energies in the two dielectric media (Bashford and Case 2000):

$$\Delta G_{solv} = -\left(\frac{1}{\varepsilon_p}-\frac{1}{\varepsilon_w}\right)\left(\sum_i \frac{{q_i}^2}{2a_i}+\frac{1}{2}\sum_i\sum_{j\neq i}\frac{q_i q_j}{r_{ij}}\right).$$

Of course, this works only if the distance between the ions were sufficiently large compared to the atomic radii which, in fact, is not true. In GB theory the inter-atomic distances r_{ij} and Born radii a_i are combined into an interpolating function $f_{GB}(r_{ij})$ yielding the fundamental Generalized-Born formula:

$$\Delta G_{GB} = -\frac{1}{2}\left(\frac{1}{\varepsilon_p}-\frac{1}{\varepsilon_w}\right)\sum_i\sum_{j\neq i}\frac{q_i q_j}{f_{GB}(r_{ij})} \qquad (25)$$

with f_{GB} usually given by an expression similar to

$$f_{GB}(r_{ij}) = \sqrt{{r_{ij}}^2 + R_i R_j\, e^{-\frac{{r_{ij}}^2}{4R_i R_j}}}$$

where R_i and R_j are the so-called “effective” Born radii. When the inter-atomic distance r_{ij} becomes large over the Born radii $f_{GB}(r_{ij})$ approaches the value of r_{ij}. For very small distances r_{ij}, however, the value of $f_{GB}(r_{ij})$ approaches the harmonic average $\sqrt{R_i R_j}$ of the effective Born radii of the atoms i and j, respectively. If $R_i = R_j = a_i = a_j$and $q_i = q_j$ we recover the original Born formula (21). The Born radii are determined so that ΔG_{GB} gives maximal agreement with values computed using the exact Poisson equation (10). The Born radii are usually obtained by solving an integral over the protein volume and, therefore, contain the information on the protein shape. The different Born models differ in how exactly the volume integral is solved, i.e. numerically or analytically and how they are calibrated against reference LPBE computations. Common Generalized-Born implementations are GBMV (Lee, Salsbury et al. 2002), GBSW (Im, Lee et al. 2003) and FACTS (Haberthür and Caflisch 2008) and are all available as part of the CHARMM molecular-mechanics package (Brooks, Bruccoleri et al. 1983). The first performs numerical integration of the protein molecular volume and gives results that are the most accurate, i.e. the resulting solvation energies are very close to Poisson electrostatics. GBSW approximates the molecular volume using a simple smoothing function considerably accelerating the integration step. However, GBSW needs a reparametrization of the atomic radii. Finally, FACTS uses a very fast analytical expression to evaluate the effective Born radii. It is currently the fastest available Generalized-Born method, but its

extensive parameterization depends on the use of the CHARMM force field (MacKerell, Bashford et al. 1998) and restricts the available possible values for the dielectric constants to $\varepsilon_w = 80$ and $\varepsilon_p = 1$ or $\varepsilon_p = 2$.

Application of GB in the past has been primarily focused on molecular dynamics simulations to provide a potential of mean force representation of the solvent. Consequently, only a few attempts have been made to use it in pK_A computations (Luo, Head et al. 1998; Onufriev, Bashford et al. 2000; Gordon, Myers et al. 2005). In the earliest attempt known to the author (Luo, Head et al. 1998) pK_A values were computed for small dicarbonic acids as well as the Aspartate dyad in the HIV protease by numerical evaluation of configuration integrals over a finite number of conformations in the relevant protonation states. While the method provided reasonable agreement with experimental values, the use of configuration integrals renders the method very expensive even today because the integrals need to be evaluated for all relevant protonation states of the protein whose number grows exponentially with number of titratable sites as we shown above. A simpler approach was provided by (Onufriev, Bashford et al.) who stayed very close to the original LPBE-based method (Bashford and Karplus 1990) but computed the terms in eqs. (15), (16) and (18) using their own GB method. In their proof of principle the investigators employed only a simplified description of the molecular surface (assuming vdW-volume and applying an empirical correction of the Born radii). An improved version is now available through a web service (Gordon, Myers et al. 2005) providing simple electrostatic pK_A computations. Unfortunately, the authors do not discuss its accuracy and speed compared to the traditional LPBE approach which remains the default method used by their web service.

2.4.4 Empirical methods

Empirical pK_A computations rest on entirely different foundations than traditional methods: Instead of using a more or less accurate physics based description of the protein, they try to correlate certain structural features around a titratable residue with its pK_A by a statistical analysis. Empirical methods were inspired by successful applications of machine-based learning algorithms to other fields of biochemistry such as secondary-structure prediction, trans-membrane helix detection or the automatic identification of potential drug molecules from large libraries. The focus of empirical programs is speed of execution and they typically terminate within seconds even for very large proteins. Current programs are made available through a simple web page which makes them easy to use in contrast to traditional electrostatic or all-atom models requiring a great deal of technical and scientific know-how. This is possibly the most important reason why empirical programs became increasingly popular. The first publication of an empirical method called “PROPKA” (Li, Robertson et al. 2005) has sparked the development of several alternatives (He, Xu et al. 2007; Huang, Du et al. 2010) of which one (He, Xu et al. 2007) we will discuss in greater detail together with PROPKA in a later chapter. The currently most accurate method is “Pred-pKa” (Huang, Du et al. 2010) which is based on a similar model used by PROPKA but applies a more advanced training method based on iterative least-square optimization of two sets of weighting factors specific for the individual amino acids and their physical parameters, correspondingly. The overall mean error generated by “Pred-pKa” is only about 0.6 pH units. However, empirical methods suffer from the limited number of available experimental pK_A values especially for cases in which the influence of the protein is large. This aspect will be discussed in greater detail in another chapter.

3 Results

3.1 Optimizing pK_A computation in proteins with pH adapted conformations

In this publication a new program to compute pK_A's of titratable residues in proteins named "Karlsberg$^+$" is described. It can be used via a simple web interface provided on http://agknapp.chemie.fu-berlin.de/karlsberg.

It is based on an electrostatic continuum description of the protein we extended to account for the effect of structural relaxation of the protein due to neutralization of the titratable residues. Before I give a more detailed explanation of the new approach I would like to address the history that had led to the development of "Karlsberg$^+$".

The accuracy of electrostatic pK_A computations using a single protein conformation

The traditional way pK_A values were computed used the following simplified work flow: Starting with the protein's crystal structure (which, obviously, is the most basic requirement for any method), the coordinates were completed by adding the missing hydrogen atom positions (how exactly will be explained below), calculated the molecular volume based on the structure and solved the LPBE (12) many times to obtain electrostatic energies by eqs. (15), (16) and (18). Finally, these energies are plugged into eq. (21) to compute titration curves for the protein's titratable residues. In an assessment of this single conformational protocol we compared computed pK_A's ($pK_{A,i}^{comp}$) with a benchmark of 185 experimentally measured pK_A values ($pK_{A,i}^{exp}$) and calculated the root mean squared deviation (RMSD) according to

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=0}^{N}\left(pK_{A,i}^{comp} - pK_{A,i}^{exp}\right)^2}. \qquad (26)$$

Disappointingly, a very poor result of RMSD = 2.7 pK units was obtained meaning one could not distinguish between a mild acid (pK_A = 4) and a mild base (pK_A = 7) on average (see Figure 7, upper panel). Even more troubling was the fact that with RMSD = 1.0 pK unit the "null hypothesis", i.e. the approach assuming the computed pK_A's are equal to the corresponding model pK_A's, i.e. $pK_{A,i}^{comp} = pK_{A,i}^{M}$, was on average much more reliable than electrostatic computations. In other words: Doing no calculations at all gave better pK_A predictions on average. In a careful analysis of the reasons for its failure one of the most important issues turned out to be the protein structure used in electrostatic computations.

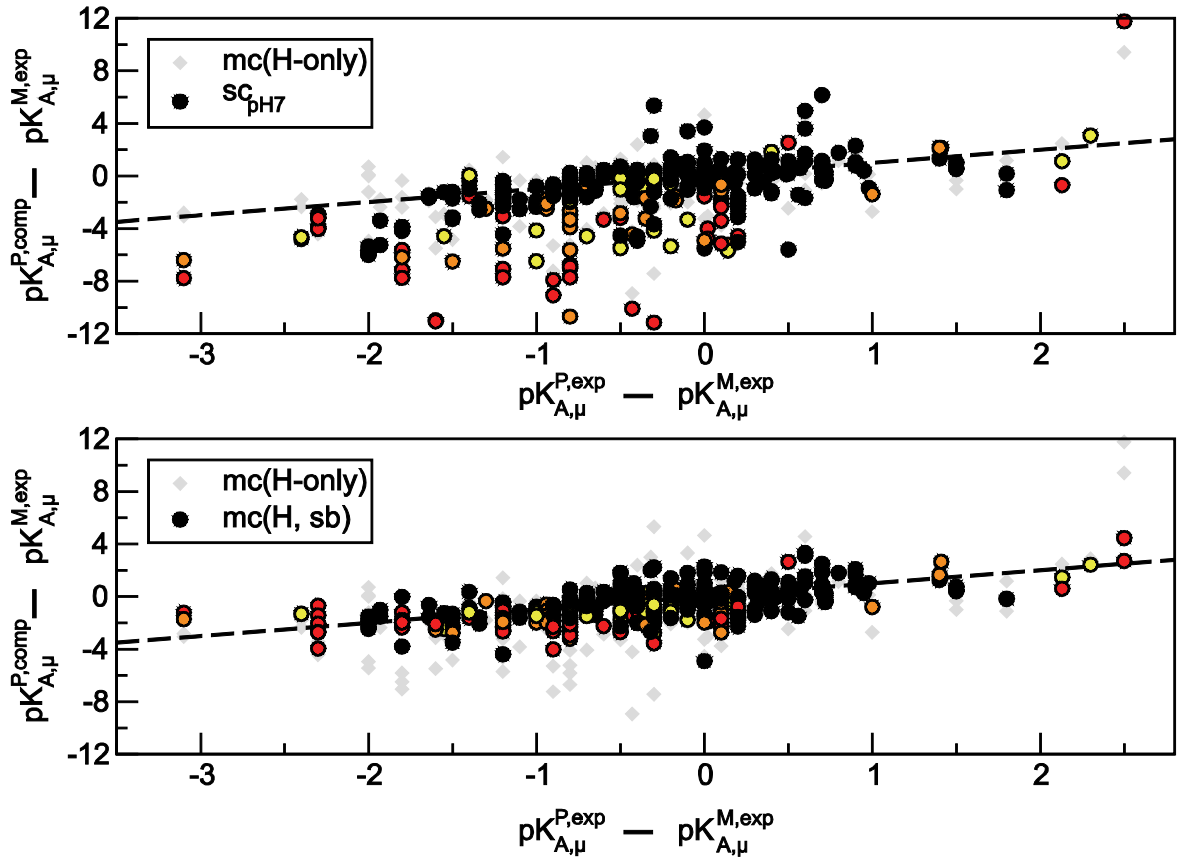

Figure 7 Computed pK_A shifts plotted against experimental pK_A shifts. A pK_A shift is the difference between the pK_A value inside the protein ($\mathbf{pK^P_{A,\mu}}$) and the corresponding value of the same titratable group μ in aqueous solution ($\mathbf{pK^M_{A,\mu}}$). For titratable groups on which the protein exerts only a weak net influence have $\mathbf{pK^P_{A,\mu} - pK^M_{A,\mu} \approx 0}$ and the corresponding points on the correlation plot will lie close to (0,0).

This is the case for the majority of experimental pK_A values available (observe the point density in the plots is highest in the center). Colored points are due to titratable groups located in salt bridges: The color tone is proportional to the number of side chain contacts in the corresponding ion pair and, hence, measures the strength of the salt bridge (yellow – few contacts, red – large number of contacts).

Upper panel: Correlating the results of computations using a single, self-consistent conformation obtained at pH = 7 (black points, 'sc_{pH7}') as well as computations using eight conformations obtained at different pH values in the range of -8 to 20 differing only in the hydrogen positions (grey diamonds, 'mc(H-only)'). Observe that the largest, absolute pK_A shifts were calculated for groups involved in strong or very strong salt bridges (orange and red points), whereas the experimental shifts do not show this dependency. The RMSD calculated according to equation (26) was 2.7 in case of results computed from a single conformation ('sc_{pH7}') and 1.9 with multiple hydrogen conformers ('mc(H-only)').

Lower panel: Results of computations in which salt bridges were allowed to relax (five conformations at pH = -8 and 20, and a single conformation at pH 7). Note how the colored points move closer to the diagonal. The resulting RMSD according to equation (26) was 1.1 pK units.

Problems using crystal structures

Problem: Missing hydrogens

Crystal structures, with the exception of a few structures with sub-atomic resolution, are usually missing the hydrogen atoms even if every heavy atom position is defined. They define the hydrogen bonding network around titratable groups determining the polarity of the local protein environment. The corresponding electrostatic interactions can be quite strong which is why the correct placement of hydrogen bonds can have a tremendous effect on the pK_A. While modeling the hydrogen atom positions is essential, it is also tricky, because the coordinates depend on the protonation pattern and, hence, indirectly on the pK_A. For example, the hydroxyl group of serine might donate a hydrogen bond to aspartate because the partially positively charged hydrogen atom is attracted by the ionized oxygens in the carboxyl group. In aspartic acid, however, the carboxyl group is protonated and could in turn donate a hydrogen bond to serine now functioning as a hydrogen bond acceptor. Hence, we run into a hen-egg problem, since we can not construct hydrogen atom positions without knowledge of the protonation pattern which, in turn, we can not predict without hydrogens.

Solution: Self-consistent geometry optimization

Traditionally, this problem was solved by means of a simple two-step optimization procedure. First, hydrogen atoms were constructed assuming titratable residues to adopt its standard protonation state at pH 7, i.e. bases and acids ionized with the exception of tyrosine and cysteine, followed by the first geometry optimization. In our group we employed the molecular mechanics package CHARMM (Brooks, Bruccoleri et al. 1983) in combination with the all-hydrogen force field CHARMM22 (MacKerell, Bashford et al. 1998) for this task. The 'HBUILD' command was used to generate initial hydrogen atom coordinates and the 'MINI' command to perform a constrained geometry optimization in which non-hydrogen atom positions were fixed. Next, electrostatic energy calculations were performed on the resulting protein structure to determine the actual protein protonation pattern at pH 7. Finally, the hydrogen positions were optimized again with the modified protonation pattern and the pK_A's computed based on the new protein structure. The approach can be generalized by repeating the two steps, electrostatic energy computations and geometry optimization, iteratively until the pattern does not change anymore. The resulting conformation can be called "pH adapted" because it is

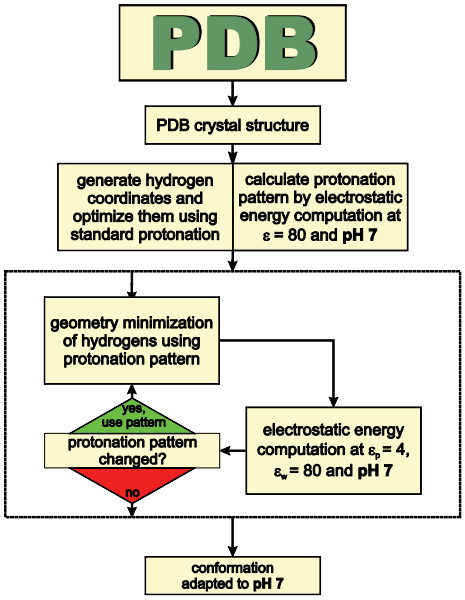

Figure 8 Flowchart describing self-consistent geometry optimization of hydrogen positions at pH 7. Starting from the crystal structure hydrogen atoms are added assuming standard protonation. Using the resulting structure, the protonation pattern is initialized by electrostatic energy computations in a homogeneous dielectric volume with $\boldsymbol{\varepsilon = \varepsilon_w = \varepsilon_p = 80}$. This step is new in Karlsberg[+] and was introduced to avoid a bias towards standard protonation. In the following geometry optimization and subsequent pK_A computations with $\boldsymbol{\varepsilon_p = 4}$ are repeated until the calculated protonation pattern converges. The end result is a protein structure whose hydrogen bonding pattern is consistent with the most likely protonation state obtained at pH 7. Of course, the procedure can be generalized to also include heavy atom positions in the geometry optimization phase.

consistent with most likely protonation pattern at the corresponding pH. In analogy to quantum chemical geometry optimizations this has been called "self-consistent" geometry optimization (Rabenstein, Ullmann et al. 1998). However, just using a similar, but slightly modified protocol (Figure 8) did not improve the RMSD value.

Problem: The pH in crystallization conditions causes a structural bias in ion pairs

Self-consistency alone, however, does not solve a more fundamental problem: There is a strong structural bias inherent in protein crystal structures. Most proteins have been crystallized only at a single pH usually at around a value of 7. At pH 7 most proteins contain quite a lot of ion pairs, i.e. pairs of acidic and basic residues that form strong hydrogen bonds with each other. In an attempt to determine the pK_A of any member of these salt-bridges experimentally the pH is varied from acidic to basic conditions tracking its ionization state. During the experiment it is very likely the local structure of these ion pairs changes upon neutralization of the residue in focus (Figure 9). It is important to note that in titration experiments, no assumption is made on the local structure of these ion pairs. In contrast, using only a single, self-consistent protein conformation like in traditional pK_A computations implies the protein structure remains unchanged over the whole pH range scanned. By investigating the benchmark set I found that about one third of the data involved residues (mostly aspartic and glutamic acids) in ion pairs and that in the cases where the computations generated the largest errors this share rose to almost 100% (see Figure 7, upper panel).

In rare cases crystallographers succeeded in crystallizing the same protein at different pH. One example is myoglobin, a heme protein involved in oxygen storage, for which there are in total five structures obtained at pH 4 (two structures), 5 (two) and 6 (one). In a previous study aimed at computing the pK_A values of the histidines in myoglobin, it was demonstrated that the agreement with experimental values improved significantly (from RMSD = 2.9 to RMSD = 0.7 pK units considering only myoglobin pK_A's) if one allowed the protein to adopt the different conformations observed crystallographically (Rabenstein and Knapp 2001). Those calculations were based on eq. (22) in which the corresponding conformation energies (G^{l}_{conf}) were fitted to reproduce crystallographic occupancies and to obey the constraint that the probability of finding the protein in any conformation within the considered pH range 3-7 is unity.

Solution: Multiple pH adapted protein conformations

Now, in an early incarnation of Karlsberg$^+$, I generated for every structure in the benchmark set a small set of seven hydrogen conformations by means of a modified selfconsistent geometry optimization procedure in which electrostatic energies were calculated at different pH values. The idea was to allow hydrogen bonding patterns to relax in response to the different ionization states occupied by the titratable residues. Since in a true prediction one does not know their pK_A values I simply scanned the whole pH spectrum between strongly acidic and basic conditions. With each conformation we increased the pH in steps of three starting from a low value of pH = -8 for the first one so that, in effect, the optimizations were done with respect to the most likely protein protonation pattern at pH values of -8, 1, 4, 7, 10, 13 and 20, respectively. The resulting structures were called "pH adapted conformations" (PACs). At low pH acidic groups like Asp and Glu are neutralizing and therefore trigger a rearrangement of the surrounding hydrogen bonds. On the other end of the spectrum, basic groups like lysine and arginine will neutralize at high pH values. The protonation pattern at pH values around 7, finally, will be close to the standard protonation pattern and, hence, the corresponding conformations close to the traditional selfconsistent conformation. Since experimental conformations for which crystallographic occupancies exist were not included, I

had to compute the conformation energy G^{l}_{conf} from scratch. In principle, G^{l}_{conf} can be calculated according to (Ullmann and Knapp 1999):

$$G^{l}_{conf} = \left(U^{l}_{ff} - U^{R}_{ff}\right) + \left(\Delta G^{l}_{solv} - \Delta G^{R}_{solv}\right). \quad (27)$$

In eq. (27) U^{l}_{ff} and U^{R}_{ff} represent the internal energies required to switch the protein from conformation l into the reference conformation R, respectively. ΔG^{l}_{solv} and ΔG^{R}_{solv} are the solvation energies required to transfer the protein from a homogeneous medium characterized by $\varepsilon = \varepsilon_p = 4$ into aqueous solution. U^{l}_{ff} and U^{R}_{ff} could be computed using the CHARMM energy function that includes terms for bonds, angles, dihedrals, Coulomb and van-der-Waals terms. Here, only the electrostatic component was computed analytically with infinite cut-off. ΔG^{l}_{solv} and ΔG^{R}_{solv} were computed using APBS. The reason why G^{l}_{conf} can not be computed in a single step is related to the numerical problem called "grid artifact". To evaluate the electrostatic energy of an inhomogeneous continuum the LPBE has to be solved numerically on a grid. The resulting energy includes the interaction of the grid points with each other bearing no physical meaning and being solely an artifact of the discretization procedure. By separating the solute-solute $(U^{l}_{ff} - U^{R}_{ff})$ and solvent-solvent $(\Delta G^{l}_{solv} - \Delta G^{R}_{solv})$ contributions to G^{l}_{conf} this problem is avoided because the grid energy cancels in the differences ΔG^{l}_{solv} and ΔG^{R}_{solv}.

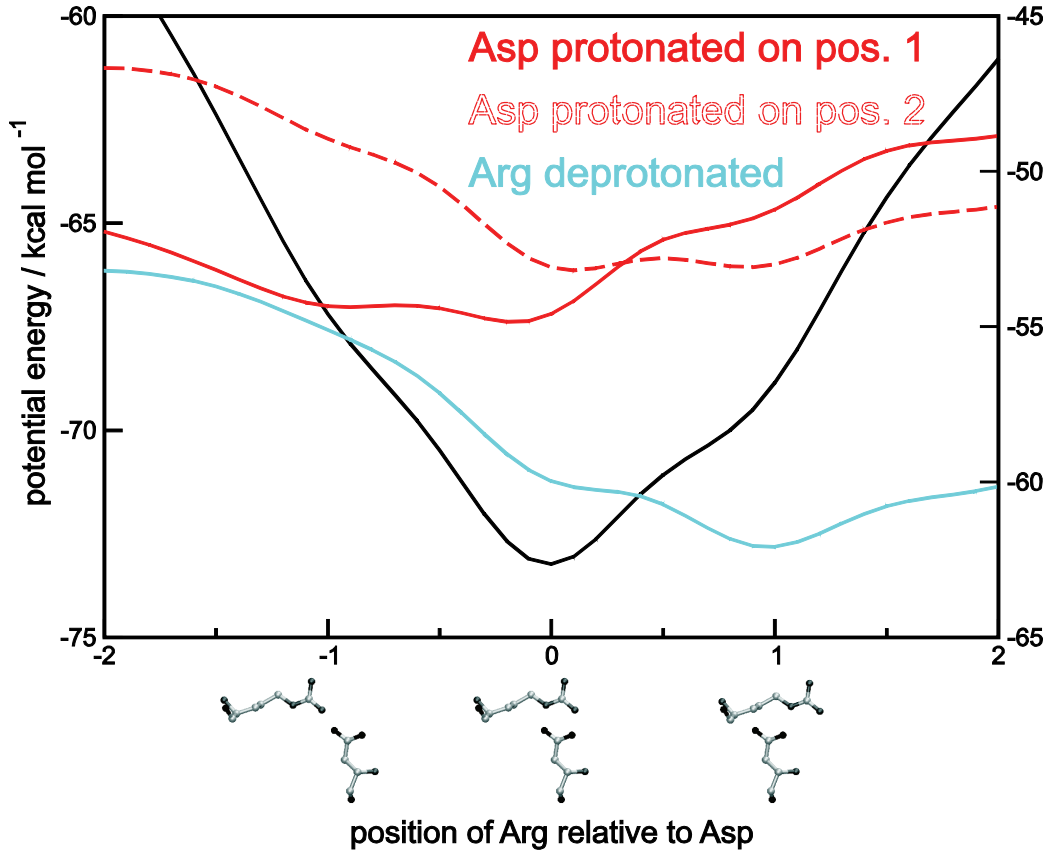

Figure 9 Electrostatic component of the potential energy of different configurations of a model arginine-aspartate salt bridge placed in a homogeneous dielectric continuum with $\varepsilon = \varepsilon_p = 4$. The black curve gives the potential energy in the case of both residues being charged. Its minimum lies at 'zero' when aspartate forms to hydrogen bonds with arginine. If arginine is neutral the minimum is shifted to +1 where it only donates a single hydrogen bond (blue curve). If aspartate is protonated and, hence, neutralized, a second minimum appears in either position -1 or +1 depending on the position of the proton (red curves). Note that the red and blue curves were drawn using the second y-axis on the right side for clarity.

The results were encouraging: The RMSD decreased significantly from about 2.7 using only a single conformation to 1.9 pK units using the seven hydrogen conformations but was still poor compared to the null hypothesis (RMSD = 1.0 pK units). When I allowed not only the hydrogen bonds to relax, but also the side chains of titratable groups present in salt-bridges upon neutralization of either residue, I obtained RMSD = 1.1 pK units being only slightly worse than the null hypothesis. If only residues were considered which in experiments exhibited strong pK_A shifts, i.e. $\left|\mathrm{pK}_{\mathrm{A,i}}^{\mathrm{M}} - \mathrm{pK}_{\mathrm{A,i}}^{\mathrm{exp}}\right| \geq 1.8$, the RMSD remained essentially constant beating the null hypothesis giving RMSD = 2.4 pK units for this subset. The remarkable accuracy of Karlsberg$^+$ also becomes obvious if one compares the position of points in the upper and lower panel of Figure 7: All colored points marking residues in salt-bridges that were previously far off the diagonal in the correlation diagram moved much closer to it.

The Karlsberg$^+$ procedure

The final version of Karlsberg$^+$ generates not only hydrogen conformations of the protein structure, but also tries to find alternative side chain positions for the salt-bridges. This is done by randomization of the corresponding side-chain positions (Asp, Glu, Arg, Lys and Tyr): The algorithm first detects salt-bridges automatically by means of a distance cutoff (max. 4 Å between hydrogen bond donor and acceptor atoms). Buried salt-bridges not located on the protein surface are discarded at this stage because structural changes in these places tend to destabilize the resulting conformations in favor of the crystal structure. Then Karlsberg$^+$ randomizes their side chain dihedral angles one ion pair after the other. Following each randomization step the corresponding atom positions are geometry optimized locally (50 steps done using the “steepest descent” minimizer and 500 steps using the “Adapted-Base Newton-Raphson” algorithm). Next, all ion-pairs are geometry optimized together (500 steps of “Adapted-Base Newton-Raphson”). Randomization and energy minimization is repeated 30 times generating 30 random global conformations of the protein in total. Finally, the protein conformation with the lowest CHARMM energy is used in an electrostatic energy computation to determine the most likely protonation pattern. All geometry optimizations are performed applying a weak harmonic constraint (force constant 0.1 kcal mol^{-1} $Å^{-2}$) on the starting coordinates to ensure that the resulting conformations dot not diverge too much from the original crystal structure. The whole procedure is repeated until self-consistency of coordinates and protonation pattern is achieved (Figure 10). Using different random seeds at least five PACs are generated at pH -8, where all acids are charge neutral, and pH 20, where all bases are charge neutral. Karlsberg$^+$ generates also an additional PAC at pH 7 with its heavy atom positions being identical to the crystal structure but geometry optimizing the hydrogen positions self-consistently.

In order to obtain pK_A values from the generated PACs, Karlsberg$^+$ calculates for reach of them G^{l}_{conf} as well as a list of intrinsic pK_A's ($pK^{intr,l}_{A,\mu}$) according to eq. (17) and a matrix of pair-wise interactions $\mathbf{W^l}$ according to eq. (18). Finally, Karlsberg$^+$ simulates a pH titration by means MMC sampling according to eq. (23). The MMC algorithm selects PACs having the lowest total electrostatic energy at any given pH and, therefore, allows the protein structure to relax in response to a change of the protein protonation pattern.

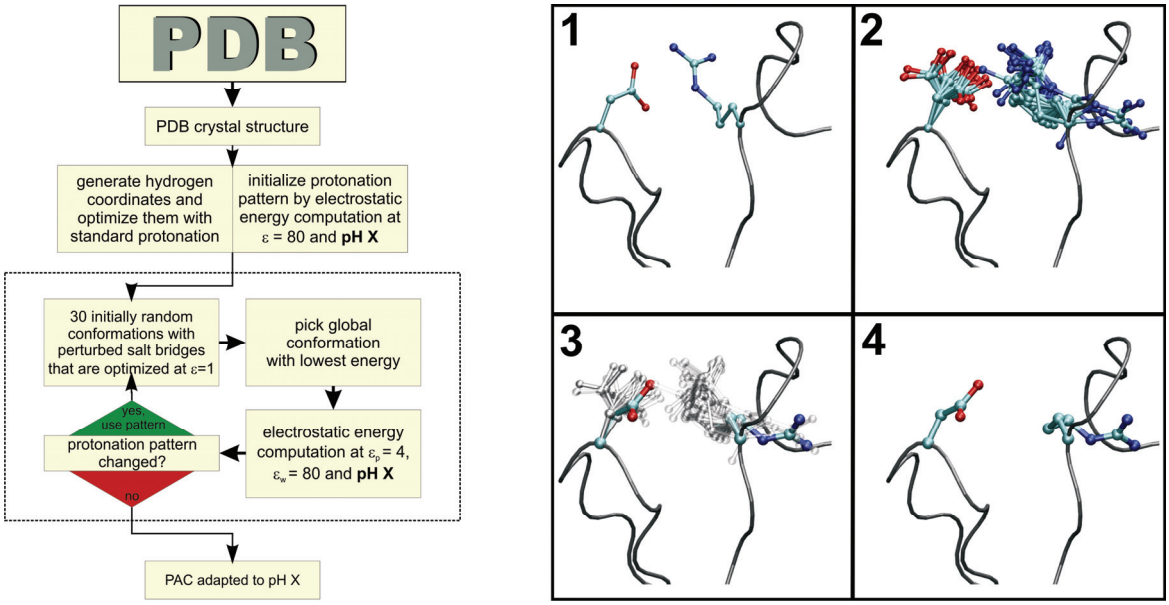

Figure 10 Self-consistent optimization procedure as implemented by Karlsberg$^+$ V1.

Left: Flowchart of the fully automated iterative procedure which bears resemblance with Figure 8 but includes a random-search of alternative side-chain conformers of ion pairs.

Right: The random search procedure in detail: 1) identification of an ion pair; 2) individual randomization of the corresponding side-chain positions (the backbone is left untouched) followed by geometry optimization with harmonic constraint, which is done for every ion pair in the structure and then repeated 30 times; 3) picking of lowest energy configuration among the generated 30 protein conformations; 4) electrostatic energy computation using the minimum structure.

3.2 Improved pK_A prediction: combining empirical and semi-microscopic methods

Looking for a pragmatic approach to improve the accuracy of pK_A computations even further, I developed a simple consensus protocol involving both electrostatic computations with Karlsberg$^+$ and predictions based on two empirical approaches. In the Karlsberg$^+$ publication we were comparing its performance with the empirical program called PROPKA (Li, Robertson et al. 2005) carrying out computations using an empirical scoring function. With RMSD = 0.9 pK units we found PROPKA's accuracy to be slightly better on average. After publication of Karlsberg$^+$ we learned of a second empirical program we refer to under the name of "PKAcal" (He, Xu et al. 2007) that exhibited a similar accuracy of RMSD = 0.9. While PROPKA's scoring function consists of terms bearing physical meaning like the number of protein backbone hydrogen bonds involving the titratable residue or the degree of solvent exposure, in case of PKAcal the terms have no meaning other than statistical significance and are optimized by a machine-learning algorithm.

In our comparisons we noticed that the biggest problem of empirical programs are not the algorithms themselves but lack of information to learn from due to the limited quantity and, more importantly, diversity of available experimental pK_A values: Most of the data consists of pK_A values that are identical or very close to the model pK_A of the corresponding residue type (eq. (14)) meaning that the protein has practically no influence on its value. In the benchmark set used in this paper e.g. only 24% of the pK_A's exhibit a shift of at least 1 pK unit compared to the model pK_A, i.e. more than $^3/_4$ of the data consists of values that are practically indistinguishable from the model pK_A judging based on the RMSD ≈ 1 pK produced by all three programs considered here. This is why the correlation plots produced by both PROPKA and PKAcal resemble of straight lines with a slope of zero (Figure 11). Correspondingly, the RMSD produced by the empirical programs increased significantly on distilled subsets of the benchmark containing only strongly shifted pK_A's. In comparison the accuracy provided by Karlsberg$^+$ is stable and better for these difficult targets.

Inspired by a large-scale benchmark (Davies, Toseland et al. 2006) that compared many different programs, I sought to find a consensus approach combining the pK_A predictions of PROPKA, PKAcal and Karlsberg$^+$ that provides a higher accuracy than the individual programs alone. I was testing two approaches: i) the "difference" method and ii) multi-linear regression. In the first approach the pK_A in question was computed using all three programs. If the difference ΔpK_A^{diff} between the results of PROPKA and PKAcal was larger than a constant c, the consensus method would return the value computed by Karlsberg$^+$ and otherwise the arithmetic mean of the two empirical values:

$$pK_A^{comp} = \begin{cases} \dfrac{pK_A^{PKAcal} + pK_A^{PROPKA}}{2} & \text{if } \left|\Delta pK_A^{diff}\right| \leq c \\ pK_A^{KBPLUS} & \text{if } \left|\Delta pK_A^{diff}\right| > c \end{cases}. \qquad (28)$$

The value of c was determined by minimzing the RMSD as defined by eq. (26). The lowest RMSD obtained was 0.70 pK units with $c = 1.2$.

I adopted the second method directly from (Davies, Toseland et al. 2006) where it was applied on PROPKA together with several other electrostatics based programs. The predicted pK_A returned by the multi-linear regression approach is simply a weighted average of the computed pK_A's of all three programs:

$$pK_A^{comp} = \alpha\, pK_A^{PROPKA} + \beta\, pK_A^{PKAcal} + \gamma\, pK_A^{KBPLUS} \qquad (29)$$

where the weights α, β, γ like c were determined by least-square fit against the benchmark set to minimize the RMSD. I obtained RMSD = 0.64 pK units using $\alpha = 40\%, \beta = 40\%, \gamma = 20\%$. Not by coincidence, the weight of 20% assigned to the Karlsberg$^+$ is very close to the concentration of "interesting" pK_A values in the benchmark mentioned above. I like to mention that 0.64 pK units was the smallest RMSD reported for any protein pK_A prediction method for two years after its publication until this record was broken by "Pred-pKa" (Huang, Du et al. 2010), a brand new empirical program using the scoring function of PROPKA but training it systematically using an advanced machine-learning scheme. The authors reported an RMSD of 0.60 pK_A units obtained on a benchmark set that was even significantly larger than the one used by us.

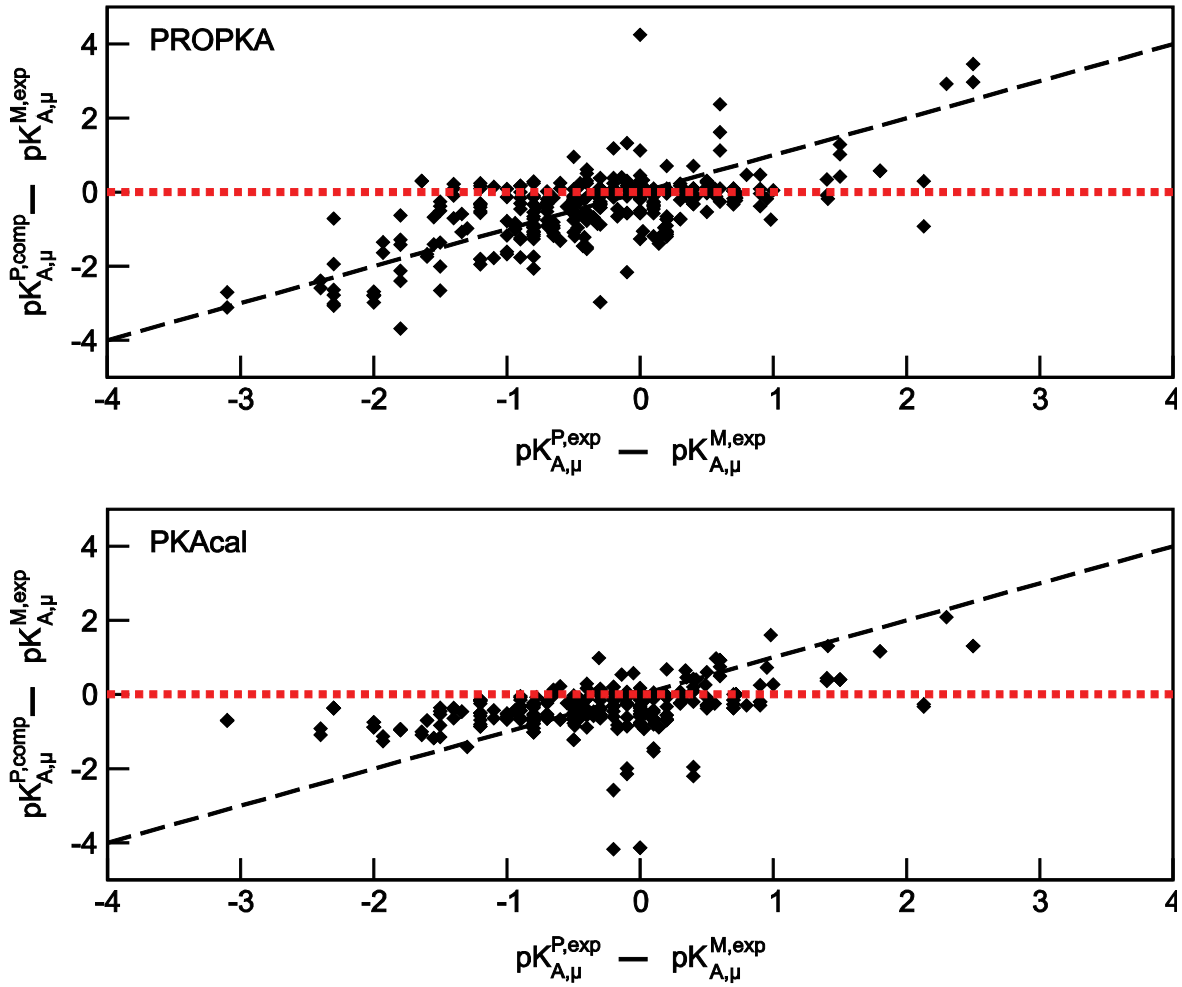

Figure 11 Computed pK_A shifts as calculated by two empirical programs, namely PROPKA (upper panel) and PKAcal (lower panel), plotted against experimental pK_A shifts. The correlation plots shown above are characteristic for empirical pK_A calculation programs having a tendency to predict small, i.e. almost zero, protein induced pK_A shifts. The bias is caused by the fact that most of the available experimental pK_A values, that are used to train these programs, imply zero pK_A shifts. For comparison the "predictions" made by the null hypothesis, i.e. assuming the protein pK_A shift is always zero, are given by the red dotted line.

3.3 Charge transport in the ClC-type proton-chloride anti-porter from *Escherichia coli*

The ClC super-family of chloride channels has been established in 1990 with the publication of the primary sequence of its first member protein (Jentsch, Steinmeyer et al. 1990) and since then recognized as a very large and important family of chloride channels with its members being ubiquitously expressed in almost any cell of eukaryotic organisms (including *Homo sapiens*) and prokaryotes as well (Miller 2006). In 2002, the first crystal structures of two bacterial homologues became available (Dutzler, Campbell et al. 2002), among them the ClC homologue from *Escherichia coli* (EClC). Back then and in a later crystallographic study (Dutzler, Campbell et al. 2003), the structural architecture was thought to be representative for all known ClC-type chloride channels. However, shortly thereafter it was realized that EClC is actually a new type of secondary active transporter (Accardi and Miller 2004). Accordingly, EClC pumps protons across the membrane driven by a chloride gradient or the other way around. Interestingly, anionic chlorides move opposite to the cationic protons in a fixed 2:1 stoichiometry suggesting a defined microscopic transport mechanism at work inside of EClC.

Experimental facts

While the pathway the chlorides take through EClC was firmly established by the crystallographic identification of three discrete chloride binding sites that are located at the periplasmic surface ($Cl^{(1)}$ "outsides" the bacterial cell), in the center of the protein ($Cl^{(2)}$) and at the intracellular surface ($Cl^{(3)}$) (Dutzler, Campbell et al. 2003), the available crystal structures do not provide enough information to deduce the corresponding proton transfer pathway (PT pathway). Limited information has been inferred from various mutants of EClC: Two conserved glutamates, E148 and E203 (Figure 12), are essential for PT and converting them into glutamines or other non-titratable residues completely abolishes active proton transport but retains passive chloride transport (Accardi and Miller 2004; Accardi, Walden et al. 2005; Lim and Miller 2009). E203 is considered to be the proton entry site close to the intracellular lumen and E148, correspondingly, the proton exit site. E148

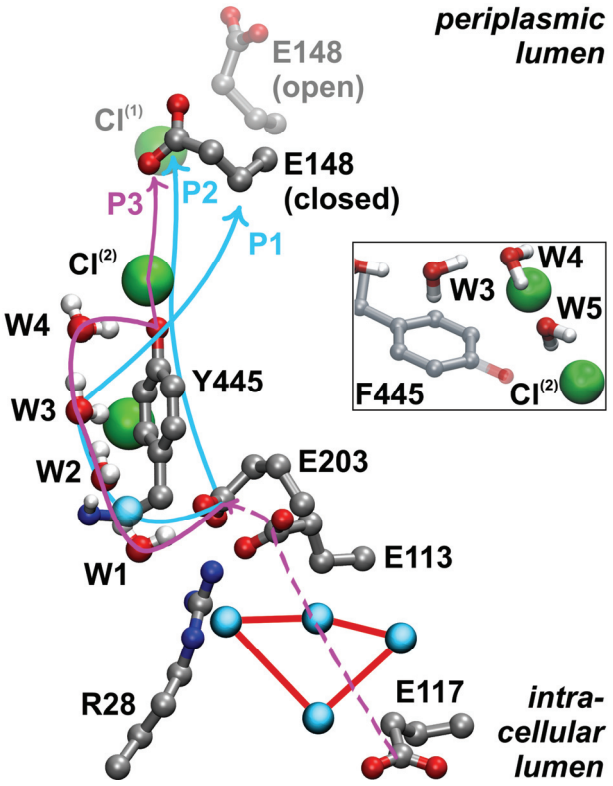

Figure 12 Possible proton transfer pathways in EClC and important titratable groups. The pathways labeled 'P1' and 'P2' were identified using a novel pore searching algorithm. In addition, 'P2' has been investigated in an MD simulation study of EClC where up to seven buried waters (not shown) were modeled in the hydrophobic space between E203 and E148. I am proposing the protons to follow path 'P3' (magenta color), which was derived from 'P1' by addition of a fourth water molecule previously not recognized. Known crystal waters are depicted by blue balls. Interestingly, there are four crystal waters in an internal cavity located not far from the putative proton entry site E203 (blue balls connected by a red line). It is speculated that proton uptake happens at E117 and not E203 which might only relay intracellular protons.

Inlay: Locations of the buried waters W3-5 in the Y445F mutant of EClC where the hydroxyl group of tyrosine is missing and replaced by a fifth water W5 according to my modeling.

can switch between a "closed" conformation binding to the extracellular chloride binding site and an "open" conformation exposed to the periplasmic space (Accardi, Walden et al. 2005; Lim and Miller 2009). However, it remains a mystery how protons pass the large hydrophobic gap of about 15 Å separating E203 from E148. Large distance tunneling as observed in case of electrons (Page, Moser et al. 1999) is impossible for protons with their much larger mass. No titratable residues reside in that area which could support PT except for a conserved tyrosine (Y445) whose acidic oxygen is still 12 Å away from E148.

Previous theoretical works

A pore-searching algorithm on EClC led to the identification of two potential PT pathways, named 'P1' and 'P2' (Kuang, Mahankali et al. 2007). The former was found to be large enough to hold three single-filed buried water molecules (W1-3) including the crystal water W2 that forms a hydrogen bond with the backbone of Y445 in the 1OTS crystal structure of WT EClC. The protons could possibly be conducted along this "water wire", however, the PT pathway 'P1' does not end directly on the proton exit site E148 (Figure 12). The second pathway 'P2' on the other hand ends on E148 but is too small to contain buried waters. Correspondingly, no crystal waters were observed in this region. Nevertheless, up to seven buried waters were modeled into 'P2' in a recent molecular dynamics (MD) study (Wang and Voth 2009). With these additional buried waters the investigators successfully simulated the transfer of an excess proton from E203 downstream onto E148 in quantum-classical MD simulations. They predict E203 to act as a proton shuttle switching its conformation upon protonation to form a hydrogen bond with the buried waters in 'P2' and establish the PT pathway. However, in the crystal structure of the EClC mutant E203Q the side chain of Q203 corresponding essentially to a protonated E203 is nearly indistinguishable from the WT coordinates (Accardi, Walden et al. 2005).

This study

For this thesis I investigated 'P1' more closely. In preparations I tried to add as much water molecules into its pore as possible and was able to stabilize a linear chain of four buried waters (W1-4) in the WT structure of EClC (Figure 12). The fourth water (W4) was previously unpredicted and located close to the hydroxyl group of Y445 to which it donates a hydrogen bond. Because Y445 is also hydrogen bonded to the central chloride ($Cl^{(2)}$) and the latter only 4 Å away from E148, this alternative water wire would establish a new potential PT pathway 'P3' connecting E203 all the way downstream with E148. An important implication of the new pathway is, if correct, that chloride must be, at least transiently, protonated.

I explored this hypothesis by characterizing the energetics of proton transfer along the putative pathway by means of electrostatic energy computations. Instead of computing the pK_A values of the involved titratable groups (E203, W1-4, $Cl^{(2)}$, E148) by globally varying the pH as I did in the benchmark calculations using Karlsberg$^+$, the individual protonation energies were probed at a fixed pH = 4.5 (the pH optimum of EClC) so as to study the system under conditions as close as possible to the physiological setting. This way I obtained the non-equilibrium pK_A (or "action pK_A") of $Cl^{(2)}$ when for example the neighboring chlorides ($Cl^{(1)}$, $Cl^{(3)}$) remain deprotonated and do not titrate at the same time as $Cl^{(2)}$ as it will happen under equilibrium conditions varying the external solvent pH. However, as in case of previous Karlsberg$^+$ calculations the relaxation of the protein structure was accounted for in a limited fashion by selfconsistent geometry optimization of the EClC structure. In this study, geometry optimization was limited to hydrogens and the buried waters W1-4 (salt-bridges located on the protein surface were not included in the optimization to save computational resources). In fact,

every titratable group was associated with two protein conformations: one obtained by fixing the corresponding residue in its protonated state and a second where the same residue was deprotonated (and, therefore, was the same for any group).

Action pK_As were recomputed with different chloride loading states varying the amount of chloride anions inside the transporter between zero and three chlorides per subunit. As a result, pK_As for the central chloride $Cl^{(2)}$ ranging between -14 and -7 were obtained using $\mathrm{pK}_{\mathrm{A,HCl}}^{\mathrm{M}} = -6$, the pK_A of hydrochloride in aqueous solution (Robinson and Bates 1971). The higher values around -7 corresponded to the crystallographically resolved loading states with two and three chlorides inside EClC. Although the chloride pK_As were the lowest of all titratable groups in the proposed PT pathway, they were not dramatically lower compared to the pK_A values of the nearby buried waters W3 and W4 (ranging between -10.5 and -3.8) because the protein does a poor job to stabilize oxonium ions. Furthermore, a protonation energy of 70 kJ/mol (corresponding to pK_A = -7) is compatible with the rather slow kinetics of EClC (1000 protons s^{-1}) assuming the production of hydrochloride to be rate limiting. To investigate the hypothesis in detail a reduced quantum system was set up consisting of the waters W2, W3, W4, Y445, $Cl^{(2)}$ and its ligands as well as E148. With an excess proton located on W4, indeed, a structure was obtained in which a proton resided on the chloride after geometry optimization in vacuum (Figure 13). Since the potential energy decreased strictly monotonically, the result also suggest protons can transfer activationless from W4 onto $Cl^{(2)}$.

A similar linear water chain was successfully modeled into the mutants Y445F and E203H of EClC explaining their almost WT-like activity. In case of Y445F the water wire consists of five buried waters with W5 replacing the hydroxyl group of Y445. The fact that histidine can replace the putative proton entry group E203 was at first surprising but could be explained by the very acidic pK_A calculated for

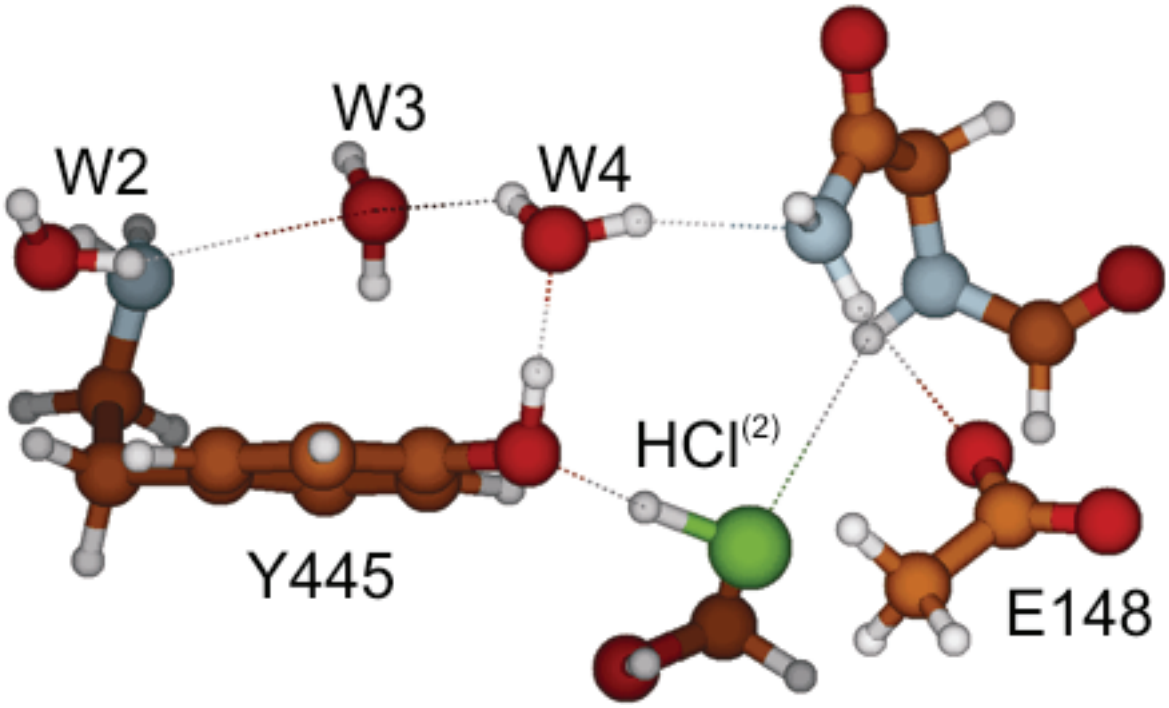

Figure 13 Final geometry obtained by placing an excess proton on W4 and performing an ab initio optimization of a reduced model system of the proposed proton transfer pathway in EClC. The excess proton ends up on the hydroxyl group of Y445 and is involved in a hydrogen bond with W4. The proton originally resident on Y445, however, moved onto the central chloride $Cl^{(2)}$. The hydrogen chloride species donates a hydrogen bond to Y445 (and not to E148).

H203 that is qualitatively the same as for E203 (-0.5 versus -0.9). A closer inspection of the protein environment surrounding E203 also revealed that it might not be the entry group at all. First of all, E203 is not directly solvent accessible at least not in the conformation present in the WT crystal structure. Moreover, an unusual, protonated glutamate E113 is donating a hydrogen bond to E203 and is partially solvated by an internal cavity characterized by a negative electrostatic potential. This residue, together with E203, is conserved among prokaryotic ClC-type antiporters. Therefore, it is tempting to speculate that a proton is conducted from E117 located at the intracellular protein surface via this cavity onto E113 when it deprotonates and injects a proton into 'P3'.

EClC exchanges a proton on the intracellular surface with two extracellular chloride ions either being in proton or chloride pumping mode. The reasons for the strict 2:1 stoichiometry are unknown but must be intrinsic to the underlying reaction mechanism. Although my results do not deliver a stringent explanation why two chlorides are needed, they allowed the construction of a plausible working model of the transporter cycle. Judging from the electrostatic energies of about 100 proton and chloride loading states of EClC I identified a dozen or so of states which, when adopted by EClC in a specific order, indeed lead to the observed stoichiometry (Figure 14). The proposed reaction cycle has several interesting properties: (i) The proton conducting chains are doubly protonated during a specific phase of the transport cycle where one proton resides on W4 and a second on E148, which is left-over from the previous cycle. (ii) EClC evolves through states with one, two and three chlorides and, as a consequence, all three chloride binding sites are involved in the exchange cycle although it has been postulated before that the $Cl^{(3)}$ site is not essential (Miller and Nguitragool 2009). (iii) To obtain the 2:1 stoichiometry I had to assume that the proton once located on $Cl^{(2)}$ is, although highly energetic, kinetically trapped, which would leave enough time for a second chloride to enter the intracellular lumen. However, this assumption is plausible as the proton most probably arrives on $Cl^{(2)}$ forming a strong hydrogen bond with the hydroxyl group of Y445 and not facing the E148 carboxylate. This was confirmed by quantum chemical geometry optimization of a model system with excess proton (Figure 13).

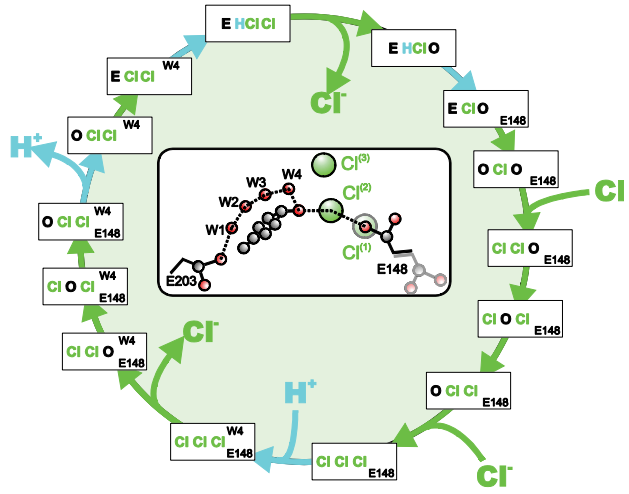

Figure 14 Proposed transport cycle operative in EClC. Green arrows indicate charge transport events involving exclusively chloride anions. Blue arrows indicate proton transfer steps. With one turn of this reaction cycle, two chlorides are moved inside the bacterial cell and one proton outsides which is consistent with the 2:1 stoichiometry observed in experiments of EClC.

Outlook

This study went beyond of just proposing a potential proton transport pathway in EClC: To our knowledge it was the first time a serious attempt was made to elucidate the transport mechanism of EClC. However, parts of the proposed reaction cycle are based on simplifications and speculation due to a complete lack of experimental information regarding the chloride or proton transport mechanism operative in EClC. New experimental insides might shake our model in its foundation.

Indeed, on Oct, 29th 2010, during the review phase of the manuscript, a crystal structure of a eukaryotic ClC-type transporter isolated from the thermophilic alga *Cyanidioschyzon merolae* (CmClC) was published in Science (Feng, Campbell et al. 2010). Its most important feature is that it might represent a new possible intermediate in the exchange mechanism. Using a notation introduced in the manuscript, where the symbols 'Cl', 'E' and '0' written from left to right refer to the periplasmic, central and cytoplasmic chloride binding site in the EClC channel occupied by either chloride, E148 or nothing, respectively, the new EClC conformation can be represented by the string 'Cl E Cl'. In previous structures, the proton exit site E148 adopted two possible conformations: Either occupying the periplasmic chloride binding site ('E Cl Cl') or giving way to an external chloride in the protonated state ('Cl Cl Cl; E148'). In CmClC, the homologous glutamate E210 is occupying the central chloride binding site and forming a hydrogen bond with the hydroxyl group of Y445, while the periplasmic and internal chloride sites are saturated with chloride (Figure 15).

If such a conformation also exists in EClC, it means chloride might not need to be protonated so that protons can reach the exit site E148. In preliminary computations using the new structure (where I have not attempted to model a water wire) I obtained pK_A = +0.7 for E210 under non-equilibrium conditions. This is significantly higher than for chloride in the same position (at least -7, value calculated in EClC, see above) and, hence, protonation of E210 requires considerably less energy. Although the result seems to suggest the activation barrier of the PT process to be considerably lower in this alternative scenario, where glutamate instead of the central chloride $Cl^{(2)}$ is involved in PT, the reader should be reminded that, according to the calculations done with EClC, protonation of the waters W3 or W4 would require almost as much energy as protonation of chloride: The corresponding oxonium pK_As were calculated to be about -6 in the EClC loading state 'Cl Cl Cl; E148' in which, according to my model of the exchange cycle, proton uptake on the intracellular side is expected to take place.

Another remarkable feature of the CmClC variant is the fact that the putative proton entry site E203 being conserved in both prokaryotic and eukaryotic homologues previously recognized (including the human forms hClC-3, hClC-4 and hClC-5) is replaced with threonine (T269) which is commonly thought of a non-titratable residue due to the comparably weak acidity of its hydroxyl group (pK_A of ethanol is about 16). It was speculated T269 might deprotonate transiently under non-equilibrium conditions (Feng, Campbell et al. 2010). On first glance, it looks like deprotonated T269 could be stabilized by the resulting ion pair together with lysine K171 in CmClC, and, hence, relay protons similar to chloride. However, T269 and K171 do not form a hydrogen bond with each other in the CmClC structure (Figure 15). Thus, according to my preliminary computations the T269 pK_A is as high as 29 rendering deprotonation at physiological pH extremely unlikely as this would require about 120 kJ/mol at pH 7. Also, substituting glutamate in position 203 by threonine or serine, completely abolished transport activity in EClC (Lim and Miller 2009) implying that even if threonine would be involved in PT in CmClC it is not active in the EClC transporter.

With all these dramatic differences to EClC it is tempting to speculate that protons follow a unique pathway in CmClC that differs from the one in EClC. However, until crystal waters are resolved in CmClC or buried waters modeled inside the structure, the picture of the PT pathway in CmClC remains incomplete making it very difficult to draw conclusions on its function.

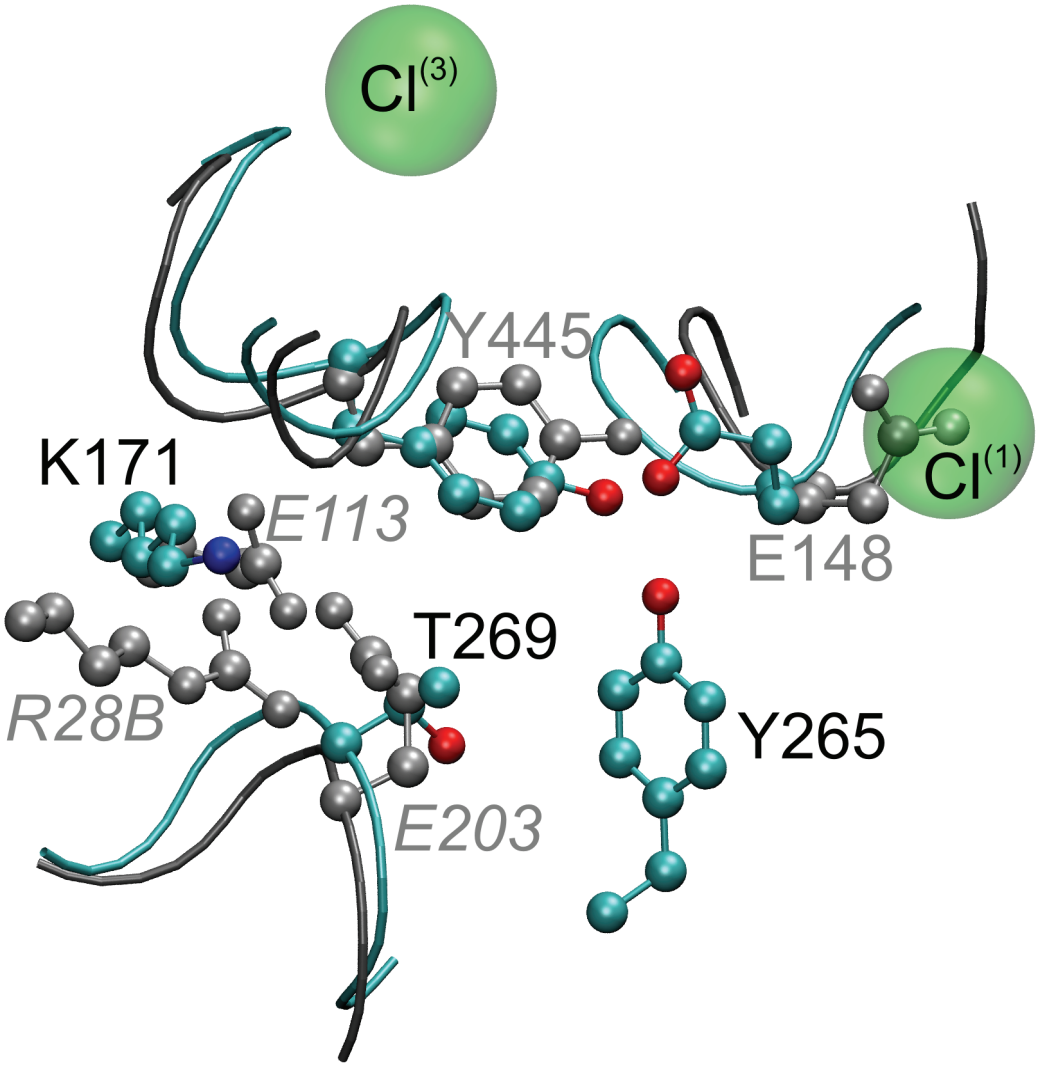

Figure 15 The crystal structure of EClC (PDB code: 1OTS, in grey) superimposed on the brand new crystal structure of CmClC (PDB code: 3ORG, in color). Two key residues are conserved (E148 and Y445), while the putative proton entry site E203 is lost in the eukaryotic transporter. E203 is replaced by threonine T269 and the authors of the crystallographic study speculate it might be capable of ionization. Judged from preliminary pK_A computations where I found pK_A = 29, however, this seems rather unlikely. Interestingly, the protonated glutamate E113 which I suspect of relaying protons from the intracellular lumen is replaced by K171. A homologous lysine is also found at this position in other known eukaryotic transporters (as, for example, in the human systems hClC-4, hClC-5 and hClC-7).

4 Conclusion and outlook

For this thesis I developed a new method to improve electrostatic pK_A computations, namely Karlsberg$^+$. The new program extends the established electrostatic description of the protein by accounting for conformational flexibility of a protein in an explicit manner. Allowing the protein structure to relax in response to (de)protonation events at its titratable residues Karlsberg$^+$ calculates pK_A values with considerable higher accuracy than traditional electrostatic approaches as tested using a large benchmark of experimentally known protein pK_A values. The mean error produced by Karlsberg$^+$ is in the order of 1 pK unit, while the traditional, single conformational approach gives an error of about 3 pK units using the same structures and the same dielectric constant. By combining Karlsberg$^+$ with empirical pK_A prediction tools in a new consensus method, I obtained an average errors as low as 0.64 pK units – the best result by any method until recently. Since electrostatic pK_A computation tools are notoriously difficult to use, the Karlsberg$^+$ web service (http://agknapp.chemie.fu-berlin.de/karlsberg) was established rendering electrostatic pK_A computations with Karlsberg$^+$ as simple as uploading a crystal structure to a web form. However, Karlsberg$^+$ can also be used offline and extended for more complex tasks involving non-protein groups. I have demonstrated this by application of Karlsberg$^+$ in a detailed study of the EClC proton-chloride transporter. The program was used to optimize the modeled, buried water wire inside the large membrane protein and study the PT reactions at the waters. In EClC conformational changes at a conserved glutamic residue play a central role in the transport mechanism. Karlsberg$^+$ proved an invaluable tool to generate the exploding number of combinations of conformations, chloride and proton occupancies of this system.

The encouraging results do not mean, however, there is no room for improvement left. In 2009 I participated in the "pK_A cooperative" in which participating programs had to calculate about a hundred pK_A values (targets), some of them measured and published, but most of them not. The results were devastating for all programs but, especially, for Karlsberg$^+$. Almost all targets were mutant residues inside Staphylococcal Nuclease (SNase), a DNA restricting enzyme. Interestingly, the mutants were not made from wild-type SNase but from a re-designed, hyperstable variant of SNAse. Some of these mutants were specifically designed to probe the dielectric constant inside the protein (García-Moreno, Dwyer et al. 1997; Karp, Gittis et al. 2007). Correspondingly, the charged residues whose pK_A values were to be calculated, namely aspartate, glutamate or lysine, are located in strongly hydrophobic environments. The pK_A values obtained with Karlsberg$^+$ were usually > 12 (aspartate, glutamate) or < 0 (lysine) suggesting that the corresponding titratable residues must be neutral under physiological conditions. While the latter prediction was consistent with experiments, our calculated pK_A values were far away from the published experimental values. For example, in the mutant V66D Karlsberg$^+$ calculated a pK_A of 16 which is exceptionally high for an acidic residue like aspartate. The experimental pK_A of 9 is still very high, but 7 pK units below the predicted value. Why did Karlsberg$^+$ overestimate the pK_A shift so drastically?

The result implies that the polarizability of the protein is apparently higher in the environment of these mutant side chains than what is predicted by classical continuum electrostatics (Isom, Castañeda et al. 2010). The value of the protein dielectric constant used by Karlsberg$^+$ is $\varepsilon = \varepsilon_p = 4$ which is at the lower end of the spectrum of values used by different programs. Correspondingly, programs using higher values around 10 achieved better agreement with experiments. However, the apparent dielectric constant necessary to reproduce some experimentally measured pK_A values goes

up to a value of 38, and larger structural changes triggered by ionization of the corresponding residues could be ruled out in most cases (Isom, Castañeda et al. 2010). But what effect is it then that so extraordinarily stabilizes charges in the hydrophobic core of protein? Interestingly, buried waters are observed in the local environment of the mutant titratable groups in some crystal structures which are not observed in the wild-type structure. It is tempting to speculate that water penetrates the protein caused by the insertion of charged species. Usually, crystal waters are removed by Karlsberg$^+$ because internal waters are thought to be represented by internal cavities with a dielectric constant of 80 as detected by calculating the protein's molecular volume (not to be confused with the solvent-accessible surface area, SASA). However, as Figure 16A shows, with the usual parametric setup (probe radius of 1.4 Å) Karlsberg$^+$ fails to detect narrow cavities filled by single-filed waters. The reason for this failure is probably the smoothing procedure applied to remove small high dielectric interstitia at the protein surface by inflating the atomic radii (Figure 16B) – a technique which is applied by all popular finite-difference Poisson-Boltzmann solvers. With a modified protocol this artifact can be avoided resulting in considerably better SNase pK_A values in calculations with a new Karlsberg$^+$ version (Meyer and Knapp 2010).

Another way out would be to treat single-filed waters in an explicit manner, i.e. keep buried crystal waters in Karlsberg$^+$ calculations. By treating these waters flexible, Karlsberg$^+$ can account for the relaxation of the hydrogen bonding networks involving these waters and titratable residues. By including the crystal waters shown in Figure 16A the correct pK_A of aspartate in the V66D mutant of SNase was reproduced. Unfortunately, many mutant structures lack crystal waters near the mutant side chain. This must not mean, however, that there are none: As the example of the EClC proton-chloride transporter shows, crystallographers might simply fail to resolve them. Alternatively, water may enter the protein transiently triggered by ionization of the titratable, mutant residue. With Karlsberg$^+$ one could prepare protein conformations with and without water at the group of interest and study the pH dependence of water penetration.

Last but not least I would like to mention that Karlsberg$^+$ can also help to improve the accuracy of redox computations: The concept of pH adapted conformations (PACs) can be easily generalized to redox adapted conformation (RACs), i.e. protein conformations that depend on the redox state of redox active groups inside the protein. In a recent study, Karlsberg$^+$ was used to calculate the redox potential of the iron-sulfur complex in different variants of rubredoxin with a deviation of only 15 mV to the experimental values. The high accuracy allowed the interpretation of the differences in redox potentials between various rubredoxin mutants and variants from two different species in terms of the formation or removal of a small number of hydrogen bonds. Quite frequently, proton and redox transfer processes are intricately coupled like in the cytochrome c oxidase or photosynthetic reaction centers. So, hopefully, there are going to be many more exciting applications of Karlsberg$^+$.

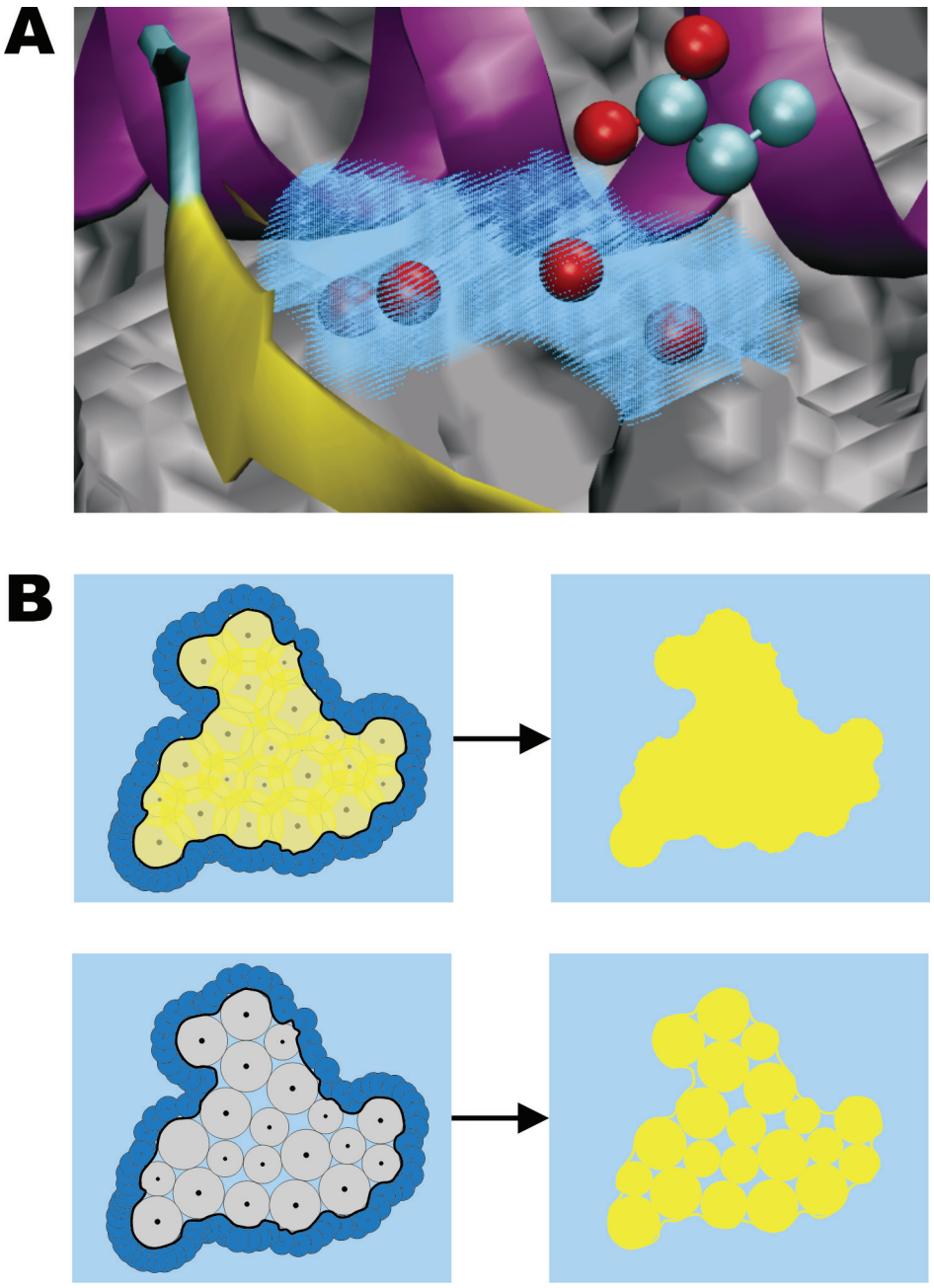

Figure 16 Molecular volume routines implemented in popular Poisson-Boltzmann solvers fail to detect internal cavities with pores so narrow that they accommodate only single-filed waters.

A: View inside the molecular volume of the V66D mutant of SNase. Four crystal waters are observed in close proximity of the Aspartate D66 spanning through an arc-shaped pore. The grey surface in the background represents the interface between protein and solvent. At both feet of the "water arc" the interface starts protruding into the molecular volume gradually narrowing but stops right at the positions of the two outer crystal waters. The blue shaded volume is obtained using a modified molecular volume routine which fully reproduces the water arc (Meyer and Knapp 2010).

B: Molecular volume generation with (upper row) and without (lower row) inflation of atomic radii. The molecular volume is usually obtained by rolling a sphere with a value equal to the average molecular radius of the solvent molecule (in case of water this is 1.4 Å) over the solvent-accessible surface area (SASA, black contour line) after inflating atomic radii by the value of the probe radius. This is done to avoid the creation of small interstitial (lower row, second picture) which, although not solvent-accessible, are assigned the solvent dielectric constant. Interstitia are artifacts caused by atoms being represented as hard spheres with a sharp boundary in Poisson-Boltzmann electrostatics.

English summary

Karlsberg$^+$ is a new method developed to improve the accuracy of electrostatic pK_A and redox computations of titratable groups in proteins. The program describes proteins using Poisson-Boltzmann electrostatics like others before, but Karlsberg$^+$ also allows protein structures to relax in response to changes of ionization states of individual groups. For that purpose, the program computes pK_A values in proteins by electrostatic energy computations using a small number of optimized protein conformations derived from crystal structures. In these protein conformations hydrogen positions and geometries of salt bridges on the protein surface were determined by "self-consistent" geometry optimization considering the most likely protonation pattern at least at three different pHs (acidic, neutral and basic). The special treatment of salt bridges at protein surfaces by Karlsberg$^+$ is most relevant, since they likely open at low and high pH, and justified by the results of large benchmark computations. Without conformational flexibility enhanced electrostatics the root mean square deviation (RMSD) between experimental and computed took a value of 2.7 pK units. In contrast, the RMSD obtained with Karlsberg$^+$ was only 1.1 pK units using the same force field parameters (value of the protein dielectric constant, atomic radii and partial charges). Comparable benchmark computations were performed with two different empirical programs that are not based on continuum electrostatics, namely, PROPKA and PKAcal. PROPKA deploys an empirical, but physically motivated energy function with about 50 adjusted parameters. PKAcal uses a meaningless statistical scoring function containing 91 free parameters. PROPKA reproduced the pK_A values with highest overall accuracy. However, differentiating the data set into weakly and strongly shifted experimental pK_A values, PROPKA's accuracy is found to be better if the pK_A values are weakly shifted but on equal footing with that of Karlsberg$^+$ for more strongly shifted values. PKAcal reproduces strongly shifted pK_A values very badly but weakly shifted values slightly better than PROPKA. Different consensus approaches were tested combining predictions from all three methods to find a general procedure for more accurate pK_A predictions. An RMSD value of 0.64 was obtained by multilinear regression of the independent predictions of all three methods being better than for any of the individual programs alone. The latest methodological advances were applied in a study of the ClC-type proton-chloride anti-porter from *Escherichia coli*, the first transporter identified in this superfamily of chloride channels. Pathways mechanism of proton and chloride translocation and their coupling are up to now unclear. Four stable buried waters were modeled into both subunits of wild-type (WT) EClC structure. Together they form a "water wire" connecting the assumed proton entry site on the intracellular and proton exit site on the periplasmic surface. For proton transfer to work it was necessary to assume the transient production of hydrochloride in the central chloride binding site of EClC. Results of electrostatic energy computations using Karlsberg$^+$ and additional quantum chemical calculations suggested that protonation of chloride is energetically feasible. Based on a characterization of more than a hundred chloride occupancies and protonation states, a working model of the EClC transport cycle was constructed. Accordingly, EClC evolves through states involving up to two excess protons and between one and three chlorides which was required to fulfill the experimentally observed 2:1 stoichiometry. The proposed mechanism of coupled chloride-proton transport in EClC is consistent with available experimental data and allows making predictions on the importance of specific amino acids, which may be probed by mutation experiments.

Zusammenfassung auf Deutsch

Karlsberg$^+$ ist ein neues Verfahren, um die Genauigkeit elektrostatischer pK_A- und Redoxpotenzialberechnungen von titrierbaren Gruppen in Proteinen zu steigern. Ebenso wie andere Programme basiert Karlsberg$^+$ auf dem Poisson-Boltzmann-Formalismus. Es erlaubt aber darüberhinaus abhängig vom Ionisierungszustand der Residuen die Struktur des Proteins zu ändern. Dazu werden pK_A-Werte mit mehreren Proteinkonformationen, die von Kristallstrukturen abgeleitet werden, gleichzeitig berechnet. Diese Konformationen unterscheiden sich in der Lage und Geometrie von Wasserstoffatomen sowie Salzbrücken an der Proteinoberfläche und werden durch „selbstkonsistente“ Geometrieoptimierung bei mindestens drei verschiedenen pH-Werten (sauer, neutral, basisch) erzeugt, wobei das jeweils wahrscheinlichste Protonierungsmuster berücksichtigt wird. Die besondere Behandlung von Salzbrücken an Proteinoberflächen ist dabei sehr bedeutsam, weil solche Ionenpaare bei extremen pH-Werten typischerweise nicht stabil sind. Das wird durch die Ergebnisse umfangreicher Benchmark-Rechnungen mit Karlsberg$^+$ gestützt: Für die Ergebnisse aus Rechnungen mit statischen Proteinstrukturen wurde eine mittlere Standardabweichung von 2.7 pK-Einheiten zu den experimentellen Datenpunkten erhalten. Zum Vergleich wurde ein Wert von 1.1 pK-Einheiten mit Karlsberg$^+$ erhalten, bei der gleichen Wahl der variablen Parameter (Dielektrizitätskonstante im Protein, atomare Radien und Partialladungen). Vergleichbare Benchmark-Rechnungen wurden mit zwei empirischen Programmen, PROPKA und PKAcal, durchgeführt, die pK_A-Werte nicht aus elektrostatischen Energien berechnen. PROPKA verwendet eine empirische, physikalisch motivierte Energiefunktion mit 52 anpassbaren Parametern. PKAcal benutzt eine einzig auf Statistik beruhende Scoring-Funktion, die über 91 freie Parameter enthält. PROPKA erreichte die höchste Genauigkeit aller Programme. Differenzierte man die experimentellen Daten zwischen stark und schwach verschobenen pK_A-Werten, so zeigte sich, dass PROPKA schwach verschobene pK_A-Werte besser reproduzierte als stark verschobene, bei denen es mit Karlsberg$^+$ gleichauf lag. PKAcal reproduzierte starke Verschiebung sehr schlecht, war aber leicht basser als PROPKA bei schwachen Verschiebungen, d.h. einfachen pK_A-Werten. Verschiedene Konsensstrategien wurden ausprobiert, um ein allgemeines Verfahren zu entwickeln, das zu einer höheren Genauigkeit führt. Die beste Variante, eine multilineare Regression der drei unabhängigen Vorhersagen, erzielte eine Standardabweichung von nur 0.64, die besser war als für jedes der drei Programme alleine. Die methodologischen Fortschritte, die in den Vorarbeiten gemacht wurden, galt es dann anzuwenden in einer Untersuchung über die Funktionsweise eines neuartigen und ungewöhnlichen Proton-Chlorid-Antiporters aus der Familie der ClC-Chloridkanäle. Der erste Transporter dieser Art (EClC) wurde in *Escherichia coli* vor ein paar Jahren identifiziert. Bisher ist nur wenig über seine Funktionsweise bekannt, weder über die Transportwege der Protonen und Chloriden noch über die Energetik der Transportprozesse, geschweige denn etwas über den mikroskopischen Reaktionsmechanismus. In dieser Arbeit ist es gelungen, vier interne Wassermoleküle im Innern beider Proteinketten des Transporters zu stabilisieren. Die Wassermoleküle bilden zusammen einen „Wasserdraht“, d.h. ein protonenleitende Verbindung, die den möglichen Protoneneingang auf der intrazellulären mit dem Ausgang auf der periplasmatischen Oberfläche verbinden. Damit der Protonentransport unter diesen Umständen auch funktioniert, muss jedoch angenommen werden, dass durch Protonierung eines Chlorids in der zentralen Bindungsstelle des EClCs Chlorwasserstoff zumindest transient erzeugt wird. Die Ergebnisse elektrostatischer Energieberechnungen mittels Karlsberg$^+$ und zusätzliche quantenchemische Modellrechnungen legen nahe, dass die Bildung von Chlorwasserstoff in EClC tatsächlich möglich ist. Auf der Basis der über hundert charakterisierten Chlorid- und Wasserstoffbesetzungszustände von

EClC wurde eine Arbeitshypothese über der Transportmechanismus aufgestellt. Danach durchläuft EClC verschiedene Zustände mit ein bis zwei Protonen und mit ein bis drei Chloriden im Transportweg. Letzteres ist notwendig, damit der Mechanismus die beobachtete 2:1 Stöchiometrie erfüllen kann. Der vorgeschlagene Mechanismus des gekoppelten Protonen-Chlorid-Transports in EClC passt zu den bekannten experimentellen Daten und ermöglichte es, einige Vorhersagen über die funktionelle Relevanz bisher unbeachteter Aminosäurereste zu machen, die sich durch Punktmutationsexperimente leicht verifizieren lassen.

References

Aaqvist, J. and A. Warshel (1993). "Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches." Chemical Reviews **93**(7): 2523-2544.

Accardi, A. and C. Miller (2004). "Secondary active transport mediated by a prokaryotic homologue of ClC Cl- channels." Nature **427**(6977): 803-7.

Accardi, A., M. Walden, et al. (2005). "Seperate Ion Pathways in a Cl-/H+ Exchanger." The Journal of General Physiology **126**(6): 563-570.

Baker, N. A., D. Sept, et al. (2001). "Electrostatics of nanosystems: Application to microtubules and the ribosome." PNAS **98**(18): 10037-10041.

Baptista, A. M., V. H. Teixeira, et al. (2002). "Constant-pH molecular dynamics using stochastic titration." Journal of Chemical Physics **117**(9): 4184-4200.

Bashford, D. and D. A. Case (2000). "Generalized Born Models of Macromolecular Solvation Effects." Annual Review of Physical Chemistry **51**: 129-152.

Bashford, D. and M. Karplus (1990). "pKa's of Ionizable Groups in Proteins: Atomic Detail from a Continuum Electrostatic Model." Biochemistry **29**: 10219-10225.

Benkovic, S. J. and S. Hammes-Schiffer (2003). "A Perspective on Enzyme Catalysis." Science **301**(5637): 1196 -1202.

Beroza, P. and D. A. Case (1996). "Including Side Chain Flexibility in Continuum Electrostatic Calculations of Protein Titration." Journal of Physical Chemistry **100**: 20156-20163.

Boschitsch, A. H., M. O. Fenley, et al. (2002). "Fast Boundary Element Method for the Linear Poissonâˆ'Boltzmann Equation." The Journal of Physical Chemistry B **106**(10): 2741-2754.

Brooks, B. R., R. E. Bruccoleri, et al. (1983). "CHARMM: A Program for Macromolecular Energy Minimization and Dynamics Calculations." J. Comp. Chem. **4**: 187-217.

Burykin, A. and A. Warshel (2003). "What Really Prevents Proton Transport through Aquaporin? Charge Self-Energy versus Proton Wire Proposals." Biophysical Journal **85**(6): 3696-3706.

Burykin, A. and A. Warshel (2004). "On the origin of the electrostatic barrier for proton transport in aquaporin." FEBS Letters **570**(1): 41-46.

Busch, M. S. a. and E.-W. Knapp (2004). "Accurate pKa determination for a heterogeneous group of organic molecules." Chemphyschem **5**(10): 1513-22.

Busch, M. S. a. and E.-W. Knapp (2005). "One-Electron Reduction Potential for Oxygen- and Sulfur-Centered Organic Radicals in Protic and Aprotic Solvents." Journal of the American Chemical Society **127**(45): 15730-15737.

Davies, M. N., C. P. Toseland, et al. (2006). "Benchmarking pKa prediction." BMC Biochemistry **7**(18): 1471-2091.

Debye, P. and E. Hückel (1923). "Zur Theorie der Elektrolyte I." Physik Z. **24**: 185-206.

Debye, P. and E. Hückel (1923). "Zur Theorie der Elektrolyte II." Physik Z. **24**: 305-324.

Dutzler, R., E. B. Campbell, et al. (2002). "X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity." Nature **415**(6869): 287-94.

Dutzler, R., E. B. Campbell, et al. (2003). "Gating the Selectivity Filter in CLC Chloride Channels." Science **300**: 108-112.

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

Feng, L., E. B. Campbell, et al. (2010). "Structure of a Eukaryotic CLC Transporter Defines an Intermediate State in the Transport Cycle." Science **330**: 635-641.

García-Moreno, B. E., J. J. Dwyer, et al. (1997). "Experimental measurement of the effective dielectric in the hydrophobic core of a protein." Biophysical Chemistry **64**(1-3): 211-224.

Georgescu, R. E., E. G. Alexov, et al. (2002). "Combining Conformational Flexibility and Continuum Electrostatics for Calculation pKas in Proteins." Biophysical Journal **83**: 1731-1748.

Gordon, J. C., J. B. Myers, et al. (2005). "H++: a server for estimating pKas and adding missing hydrogens to macromolecules " Nucleic Acids Research **33**(Web Server Issue): W368-W371.

Haberthür, U. and A. Caflisch (2008). "FACTS: Fast analytical continuum treatment of solvation." Journal of Computational Chemistry **29**(5): 701-715.

He, Y., J. Xu, et al. (2007). "A statistical approach to the prediction of pK(a) values in proteins." Proteins **69**(1): 75-82.

Huang, R.-B., Q.-S. Du, et al. (2010). "A fast and accurate method for predicting pK_A of residues in proteins." Protein Engineering, Design & Selection **23**(1): 35-42.

Im, W., M. S. Lee, et al. (2003). "Generalized born model with a simple smoothing function." Journal of Computational Chemistry **24**(14): 1691-1702.

Ishikita, H., A. Galstyan, et al. (2007). "Redox potential of the non-heme iron complex in bacterial photosynthetic reaction center." Biochimica et Biophysica Acta (BBA) - Bioenergetics **1767**(11): 1300-1309.

Ishikita, H. and E.-W. Knapp (2003). "Redox Potential of Quinones in Both Electron Transfer Branches of Photosystem I." Journal of Biological Chemistry **278**(52): 52002-52011.

Ishikita, H. and E.-W. Knapp (2004). "Variation of Ser-L223 Hydrogen Bonding with the QB Redox State in Reaction Centers from Rhodobacter sphaeroides." Journal of the American Chemical Society **126**(25): 8059-8064.

Ishikita, H. and E.-W. Knapp (2005). "Control of Quinone Redox Potentials in Photosystem II:â€‰ Electron Transfer and Photoprotection." Journal of the American Chemical Society **127**(42): 14714-14720.

Ishikita, H. and E.-W. Knapp (2005). "Redox Potentials of Chlorophylls and Î²-Carotene in the Antenna Complexes of Photosystem II." Journal of the American Chemical Society **127**(6): 1963-1968.

Ishikita, H., B. Loll, et al. (2005). "Redox Potentials of Chlorophylls in the Photosystem II Reaction Centerâ€ " Biochemistry **44**(10): 4118-4124.

Isom, D. G., C. A. Castañeda, et al. (2010). "Charges in the hydrophobic interior of proteins." Proceedings of the National Acadamy of Sciences of the United States of America **107**(37): 16096-16100.

Jentsch, T. J., K. Steinmeyer, et al. (1990). "Primary structure of Torpedo marmorata chloride channel isolated by expression cloning in Xenopus oocytes." Nature **348**: 510-514.

Jiang, L., E. A. Althoff, et al. (2008). "De Novo Computational Design of Retro-Aldol Enzymes." Science **319**(5868): 1387-1391.

Kaplan, J. and W. F. DeGrado (2004). "De novo design of catalytic proteins." Proceedings of the National Acadamy of Sciences of the United States of America **101**(32): 11566-11570.

Karp, D. A., A. G. Gittis, et al. (2007). "High Apparent Dielectric Constant Inside a Protein Reflects Structural Reorganization Coupled to the Ionization of an Internal Asp." Biophysical Journal **92**(6): 2041-2053.

Kästner, J. and W. Thiel (2005). "Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration"." Journal of Chemical Physics 2005 Oct 8;123(14):144104. **123**(14).

Kieseritzky, G. and E. W. Knapp (2007). "Optimizing pKA computation in proteins with pH adapted conformations." Proteins: Structure, Function, and Bioinformatics.

Knight, J. L. and I. Charles L Brooks (2009). "λ-Dynamics Free Energy Simulation Methods." Journal of Computational Chemistry **30**: 1692-1700.

Kuang, Z., U. Mahankali, et al. (2007). "Proton pathways and H+/Cl- stoichiometry in bacterial chloride transporters." Proteins **68**(1): 26-33.

Kuramitsu, S., K. Ikeda, et al. (1977). "Effects of Ionic Strength and Temperature on the Ionization of the Catalytic Groups, Asp 52 and Glu 35, in Hen Lysozyme." Journal of Biochemistry **82**: 585-597.

Lee, M. S., J. Freddie R Salsbury, et al. (2004). "Constant-pH Molecular Dynamics Using Continuous Titration Coordinates." Proteins: Structure, Function, and Bioinformatics **56**: 738-752.

Lee, M. S., F. R. Salsbury, et al. (2002). "Novel generalized Born methods." Journal of Chemical Physics: 10606-10614.

Li, H., A. D. Robertson, et al. (2005). "Very Fast Empirical Prediction and Rationalization of Protein pKa Values." Proteins: Structure, Function, and Bioinformatics **61**: 704-721.

Lim, H. H. and C. Miller (2009). "Intracellular proton-transfer mutants in a CLC Cl-/H+ exchanger." J Gen Physiol **133**(2): 131-8.

Lin, H. and D. Truhlar (2007). "QM/MM: what have we learned, where are we, and where do we go from here?" Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) **117**(2): 185-199.

Luo, R., M. S. Head, et al. (1998). "pK_A Shifts in Small Molecules and HIV Protease: Electrostatics and Conformation." Journal of the American Chemical Society **120**(24): 6138-6146.

MacKerell, A. D., Jr., D. Bashford, et al. (1998). "All-hydrogen Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins using the CHARMM22 Force Field." J. Phys. Chem B **102**: 3586-3616.

Metropolis, N., A. W. Rosenbluth, et al. (1953). "Equation of State Calculations by Fast Computing Machines." Journal of Chemical Physics **21**: 1087-1092.

Meyer, T. and E.-W. Knapp (2010).

Miller, C. (2006). "ClC chloride channels viewed through a transporter lens." Nature **440**: 484-489.

Miller, C. and W. Nguitragool (2009). "A provisional transport mechanism for a chloride channel-type Cl^-/H^+ exchanger." Philosophical Transactions of the Royal Society B **364**: 175-180.

Nozaki, Y., C. Tanford, et al. (1967). [84] Examination of titration behavior. Methods in Enzymology, Academic Press. **Volume 11:** 715-734.

Olsson, M. H. M., P. K. Sharma, et al. (2005). "Simulating redox coupled proton transfer in cytochrome c oxidase: Looking for the proton bottleneck." FEBS letters **579**(10): 2026-2034.

Onufriev, A., D. Bashford, et al. (2000). "Modification of the Generalized-Born Model Suitable for Macromolecules." Journal of Physical Chemistry B **104**(15): 3712-3720.

Page, C. C., C. C. Moser, et al. (1999). "Natural engineering principles of electron tunnelling in biological oxidation-reduction." Nature **402**(6757): 47-52.

Perutz, M. (1978). "Electrostatic Effects in Proteins." Science **201**: 1187-1191.

Pisliakov, A. V., P. K. Sharma, et al. (2008). "Electrostatic basis for the unidirectionality of the primary proton transfer in cytochrome c oxidase." Proceedings of the National Academy of Sciences **105**(22): 7726-7731.

Post, C. B. and M. Karplus (1986). "Does Lysozyme Follow the Lysozyme Pathway? An Alternative Based on Dynamic, Structural, and Stereoelectronic Consideration." Journal of the American Chemical Society **108**: 1317-1319.

Rabenstein, B. (2000). Monte-Carlo-Methoden zur Simulation der Faltung und Titration von Proteinen. Fachbereich Biologie, Chemie, Pharmazie. Berlin, Freie Universität.

Rabenstein, B. and E. W. Knapp (2001). "Calculated pH-Dependent Population and Protonation of Carbon-Monoxy-Myoglobin Conformers." Biophysical Journal **80**(3): 1141-1150.

Rabenstein, B., G. M. Ullmann, et al. (1998). "Calculation of protonation patterns in proteins with structural relaxation and molecular ensembles - application to the photosynthetic reaction center." European Biophysics Journal **27**: 626-637.

Robinson, R. A. and R. G. Bates (1971). "Dissociation Constant of Hydrochloric Acid from Partial Vapor Pressures over Hydrogen Chloride-Lithium Chloride Solutions." Analytical Chemistry **43**(7): 969-970.

Rocaa, M., B. Messera, et al. (2007). "Electrostatic contributions to protein stability and folding energy." FEBS letters **581**(10): 2065-2071.

Röthlisberger, D., O. Khersonsky, et al. (2008). "Kemp elimination catalysts by computational enzyme design." Nature **453**(8 May 2008): 190-195.

Schrodinger, L. (2009). Jaguar. New York, NY.

Schutz, C. N. and A. Warshel (2001). "What Are the Dielectric "Constants" of Proteins and How To Validate Electrostatic Models?" Proteins: Structure, Function, and Genetics **44**: 400-417.

Senn, H. M. and W. Thiel (2009). "QM/MM Methods for Biomolecular Systems." Angewandte Chemie International Edition **48**(7): 1198-1229.

Sham, Y. Y., Z. T. Chu, et al. (1997). "Consistent Calculations of pKa's of Ionizable Residues in Proteins: Semi-microscopic and Microscopic Approaches." Journal of Physical Chemistry B **101**(22): 4458–4472.

Sharma, P. K., Z. T. Chu, et al. (2007). "A new paradigm for electrostatic catalysis of radical reactions in vitamin B12 enzymes." Proceedings of the National Academy of Sciences **104**(23): 9661-9666.

Sharp, K. A. and B. Honig (1990). "Electrostatic interactions in macromolecules: theory and applications." Annual Review Biophysics and Biophysical Chemistry **19**: 301-332.

Simonson, T. and C. L. Brooks (1996). "Charge Screening and the Dielectric Constant of Proteins:â€‰ Insights from Molecular Dynamics." Journal of the American Chemical Society **118**(35): 8452-8458.

Tanford, C. and J. G. Kirkwood (1957). "Theory of Protein Titration Curves." Journal of the American Chemical Society **79**(20): 5333-5339.

Tanford, C. and R. Roxby (1972). "Interpretation of Protein Titration Curves. Application to Lysozyme." Biochemistry **11**(11): 2192-2198.

Thurlkill, R. L., G. R. Grimsley, et al. (2006). "pK values of the ionizable groups of proteins." Protein Science **15**(5): 1214-1218.

Torrie, G. M. and J. P. Valleau (1977). "Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling." Journal of Computational Physics **23**: 187-199.

Tynan-Connolly, B. M. and J. E. Nielsson (2007). "Redesigning protein pK_A values." Protein Science **16**: 239-249.

Ullmann, G. M. and E. W. Knapp (1999). "Electrostatic models for computing protonation and redox equilibria in proteins." European Biophysics Journal **28**: 533-551.

Wang, D. and G. A. Voth (2009). "Proton transport pathway in the ClC Cl-/H+ antiporter." Biophys J **97**(1): 121-31.

Warshel, A. (1981). "Electrostatic basis of structure-function correlation in proteins." Accounts of Chemical Research **14**: 284-290.

Warshel, A. and M. Levitt (1976). "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme." Journal of Molecular Biology **103**(2): 227-249.

Warshel, A. and S. T. Russell (1984). "Calculations of electrostatic interactions in biological systems and in solutions." Quarterly reviews of biophysics **17**(3): 283-422.

Warshel, A., P. K. Sharma, et al. (2007). "Electrostatic Contributions to Binding of Transition State Analogues Can Be Very Different from the Corresponding Contributions to Catalysis:Phenolates Binding to the Oxyanion Hole of Ketosteroid Isomerase." Biochemistry **46**(6): 1466-1476.

Warshel, A., P. K. Sharma, et al. (2006). "Modeling electrostatic effects in proteins." Biochimica et Biophysica Acta **1764**: 1647-1676.

Warshel, A., F. Sussman, et al. (1986). "Free Energy of Charges in Solvated Proteins: Microscopic Calculations Using a Reversible Charging Process." Biochemistry **25**(26): 8368-8372.

You, T. J. and D. Bashford (1995). "Conformation and Hydrogen Ion Titration of Proteins: A Continuum Electrostatic Model with Conformational Flexibility." Biophysical Journal **69**: 1721-1733.