Benchmarking a fast proton titration scheme in implicit solvent for biomolecular simulations

Fernando Luís Barroso da Silva*,†,‡ and Donal MacKernan‡,¶

Departamento de Física e Química, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. do café, s/no. – Universidade de São Paulo, BR-14040-903 Ribeirão Preto – SP, BRAZIL, Fax: +55 (16) 3315 48 80; Tel: +55 (16) 3315 42 22; E-mail: fernando@fcfrp.usp.br, UCD School of Physics & UCD Institute for Discovery, University College Dublin, Belfield, Dublin 4, Ireland., and CECAM-IRL, University College Dublin, Belfield, Dublin 4, Ireland.

E-mail: fernando@fcfrp.usp.br

Abstract

pH is a key parameter for technological and biological processes, intimately related to biomolecular charge. As such, it controls biomolecular conformation and inter molecular interactions, for example, protein/RNA stability and folding, enzyme activity, regulation through conformational switches,protein-polyelectrolyte association, and protein-RNA interactions. pH also plays an important role in technological systems in food, brewing, pharma, bioseparations and biomaterials in general. Predicting

*To whom correspondence should be addressed

† Departamento de Física e Química, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. do café, s/no. – Universidade de São Paulo, BR-14040-903 Ribeirão Preto – SP, BRAZIL, Fax: +55 (16) 3315 48 80; Tel: +55 (16) 3315 42 22; E-mail: fernando@fcfrp.usp.br

‡ UCD School of Physics & UCD Institute for Discovery, University College Dublin, Belfield, Dublin 4, Ireland.

¶ CECAM-IRL, University College Dublin, Belfield, Dublin 4, Ireland.

the structure of large proteins and complexes remains a great challenge, experimentally, industrially, and theoretically, despite the variety of numerical schemes available ranging from Poisson-Boltzmann approaches to explicit solvent based methods. In this work we benchmark a fast proton titration scheme against experiment and several theoretical methods on the following set of representative proteins: [HP36, BBL, HEWL (triclinic and orthorhombic), RNase, SNASE (V66K/WT, V66K/PHS, V66K/Δ+PHS, L38D/Δ+PHS, L38E/Δ+PHS, L38K/Δ+PHS), ALAC and OMTKY3] routinely used in similar tests due to the diversity of their structural features. Our scheme is rooted in the classical Tanford-Kirkwood model of impenetrable spheres, where salt is treated at the Debye-Hückel level. Treating salt implicitly dramatically reduces the computation time, thereby circumventing sampling difficulties faced by other numerical schemes. In comparison with experimental measurements, our calculated pK_a values have the average, maximum absolute and root-mean-square deviations of $[0.4 - 0.9]$, $[1.0 - 5.2]$ and $[0.5 - 1.2]$ pH units, respectively. These values are within the ranges commonly observed in theoretical models. They are also in the large majority of the cases studied here more accurate than the NULL model. For BBL, ALAC and OMTKY3, the predicted pK_a are closer to experimental results than other analyzed theoretical data. Despite the intrinsic approximations of the fast titration scheme, its robustness and ability to properly describe the main system physics is confirmed.

KEYWORDS: protein titration, Monte Carlo Simulations, Tanford and Kirkwood model, protein electrostatics.

Introduction

Constant-pH simulations are becoming routine for biological systems.[1–6] As a measure of the concentration of hydrogen ions in the solution, pH indicates the availability of pro-

tons to go from solution to titratable sites on a biomolecule (when there are H^+ available in the solution, i.e. at lower pH, the acid regime) or from a biomolecule to the solution (when the solution is lacking H^+, i.e. at higher pH, the basic regime). Thus, pH is intimately related to biomolecular charge. As such, it controls both intra and inter-protein interactions. Classical examples of processes controlled by pH include protein/RNA stability and folding,[7–11] enzyme activity,[12,13] regulation through conformational switches,[14] protein-polyelectrolyte association,[15,16] protein-protein complexation[17–19] and protein-RNA interactions.[20,21] pH also plays a key rule in technological systems in food, brewing, pharma, bioseparations and biomaterials in general.[22–26]

Ideally, a quantum mechanical treatment should be used when studying pH effects, since it involves the formation and breakage of chemical bonds. However, due to the very large number of titratable sites, different protein conformations and all other charged species in an electrolyte solution, the corresponding CPU time costs are typically prohibitive. Instead, the common procedure to estimate pH-effects in molecular simulations is to assign atomistic partial charges to titratable sites as a function of pH at the beginning of the calculations (i.e. the user makes a choice between neutral or protonated amino acids as a function of pH during the simulation initial setup) and keeps them unchanged during the run. Poisson-Boltzmann (PB) solvers and empirical methods are frequently used for this purpose.[28–31] The underlying approximation is that dynamical changes of the local protein environment (exposure of the titratable side chains to water, interactions with other proteins, other titratable groups, other charged species, salt and free counter-ions in the solution) during the molecular dynamics simulation do not affect the protonation process and vice-versa. The strong protonation-conformation coupling is largely neglected. One consequence of the fact that charges of these ionization sites are fixed through the course of the simulation is that important physical mechanisms cannot be modelled, such as the attractive mesoscopic forces between macromolecules in solution arising from fluctuations in proton charge predicted by the Kirkwood-Shumaker (KS) theory.[32] Proton charge fluctuation is at the origin of the

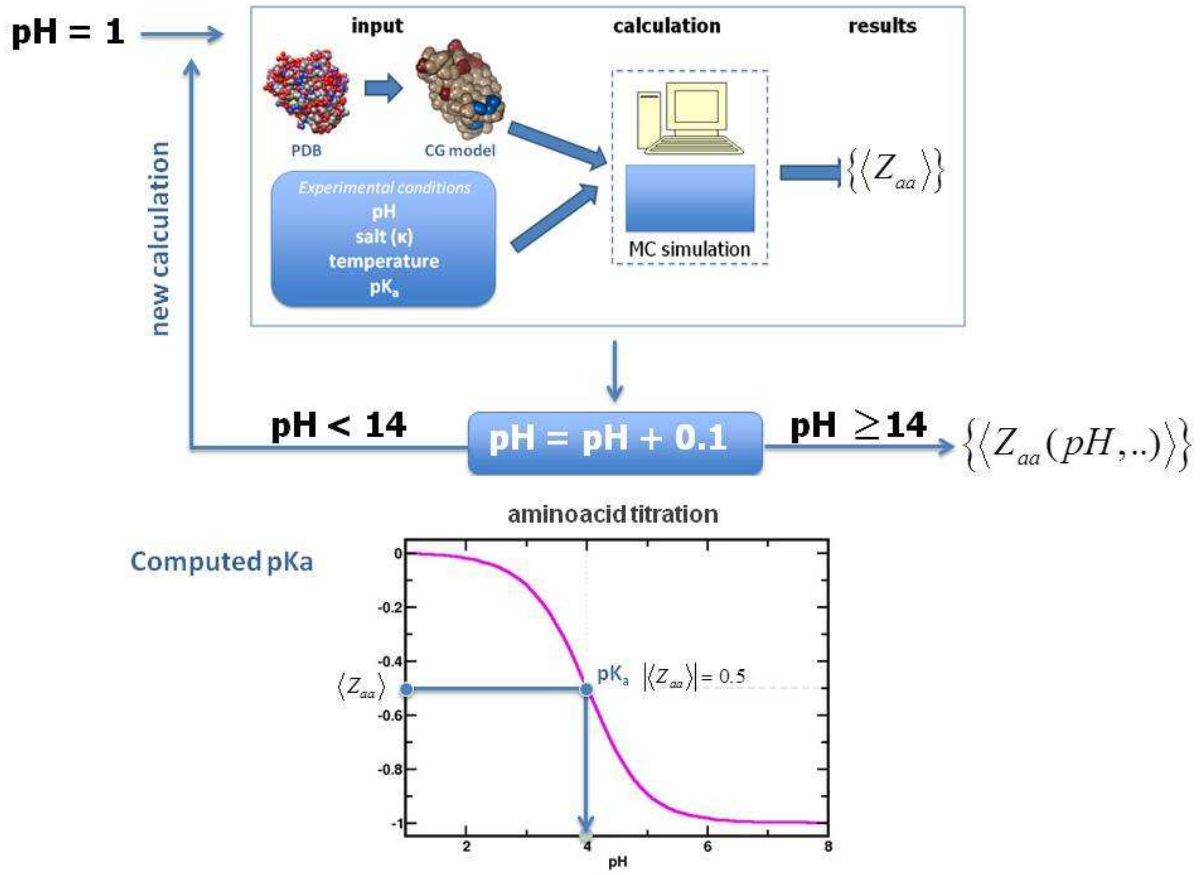


Figure 1: Scheme for theoretical titration plot and the prediction of pK_as. Input parameters are the experimental conditions (temperature, pH and salt concentration as described by the κ_c parameter). The intrinsic pK_0 values of the amino acid model compounds are taken from experimental data given by Nozaki & Tanford.[27] The calculation is repeated from pH 1 to pH 14 using small intervals (in this work we used 0.1) at a given ionic strength. See the text for more details.

so called "charge regulation mechanism", and is essential to explain protein complexation particularly at low salt and at pH regimes closer to pI (the isoelectric point).[15,18,33,34]

Attempts to model the effect of pH on biomolecular structure have been made through the combination of molecular dynamics (MD) with protonation numerical schemes, a good example being the work done by Baptista and collaborators,[1] combining both titration and conformational sampling. Their titration scheme is based on a continuum modeling of the solvent and a mean-field description of the electrolyte solution given by the linear PB equation. Other recent variants include full flexibility of the titrated molecule and the replacement

of the implicit solvent model by full atomistic representations.[2,3,6,35–37] The advantage of the latter, in principle, is a more detailed understanding of proton translocation between the protein titratable sites and the solvent. While the technical details are different for each of these methods, on practice they share is that most if not all the force field parameters obtained at a given set of experimental pH and salt conditions are used for all other explored pH regimes. Some caution should also be taken when combining different water like-models used in a mixed way to describe solvent, hydronium, hydroxide and salt properties. For instance, the introduction of titratable water as done in replica-exchange (REX) constant-pH MD (CpHMD) simulations[37] can produce artifacts in the solvent structure and dynamics which can affect protein conformation, the diffusion of mobile charged species (added salt and counter-ions) and all their interplay. The parametrization of a good dissociative water potentials is *per se* a complex and independent research field.[38] Typically much more computationally expensive simulations are also necessary for this class of methods (10ns as reported by Shen and collaborators,[37] or 40 ns for a simple dipeptide as quantified by Chen & Roux[4]), and can result in limited conformational sampling, and insufficient modelling of charge fluctuations. The slow convergence of detailed molecular solvent explicit models may explain why more empirical methods such as PROPKA (at negligible CPU cost) seem to obtain similar outcomes for pK_a predictions. In fact, PROPKA results are in general even more precise and faster than popular PB solvers.[39]

It is clearly necessary to develop fast titration schemes that do not exhibit slow convergence problems and can be applied on macromolecular systems with *multiple* titrating objects each containing several ionizable sites. Simulations at atomistic resolution of large complexes are virtually impossible today using even massive computational resources. However, it is also of fundamental importance to be able to correctly describe the main physical aspects of such systems given their biological and industrial importance. The need to perform constant-pH (CpH) simulations for protein-protein systems at several solution pH and salt regimes motivated us to propose an alternative and faster proton titration scheme for

this large scale scenario.[40] Based on a coarse-grained (CG) description of the system and rooted in the classical Tanford-Kirkwood model,[41,42] we proposed a proton titration scheme for proteins that could be used to dramatically reduce the computation time by making the calculation independent of the ionic strength cost and boosting sampling. This model in combination with the Monte Carlo method has been intensively applied in several biomolecular systems (e.g.[19,25,43,44]), and we already showed how this method can be successfully used to study protein-protein interactions.[19,25,40] The method was also recently extended to RNA titration where similar or even better outcomes were observed at much lower computational costs in comparison with other theoretical approaches including all atom CpHMD simulations).[45]

Although this fast proton titration scheme (FPTS) for proteins was not originally designed to improve pK_a predictions, recently, we observed[25] that our predicted pK_a values are surprisingly within the range of values given by different theoretical models.[37,46] For instance, comparing our pK_a results with the theoretical pK_a calculations published by Wallace & Schen[35] for the N-terminal domain of the Major Ampullate Spidroin 1, the average and maximum absolute deviations were, respectively, 0.4 and 1.2 pH units for the wild-type protein.[25] This was indeed a good indication that this simple coarse-grained model is able to capture the main physical features with computationally cheaper simulations, an important step forward in the application of CpH methods in larger systems, as is typically the situation for industrial and nanotech applications.

The present work investigates further the accuracy of FPTS for a larger variety of protein systems, including comparison with experiment and other theoretical outcomes obtained by more sophisticated and CPU demanding simulations. In particular, we focus the comparison on the latest proposed methods for CpHMD simulations in explicit solvent using the pH-REX method,[35,37] another of its variants, the pH-titration MD (pHtMD),[5] and the hybrid nonequilibrium Molecular Dynamics–Monte Carlo (neMDMC) simulation method.[4] Proteins were selected primarily to allow the present FPTS calculations to be compared with results

obtained by the methods cited above. Another criterion was the possibility to explore the accuracy of the different pH methods on proteins having a wide variety of titratable sites: proteins rich with exposed surface titratable residues accessible to solvents; deeply buried ones; and more flexible chains. Such comparisons may provide additional physical insights to improve the accuracy of biomolecular electrostatics methods in general.

This paper is organized as follows. We first review some fundamental basic details of the FPTS, which is then followed by benchmarking with the other theoretical methods. We conclude with a summary of the pK_a results for all studied protein systems, analyzed both quantitatively and qualitatively.

Theoretical background

The fast proton titration scheme used here follows a physical chemistry formalism rooted in the Tanford-Kirkwood (TK) model.[41,42] This classical dielectric continuum model assumes that the protein may be modeled as a hard-sphere inscribing charged sites (the titratable groups) placed at specific locations and immersed in a medium with high dielectric permitivitty. The salt ions and other charged ligands are not explicitly included in the model. TK appealed to the construction of an effective interaction, eliminating explicit reference to the mobile particles and describing them by the Debye-Hückel approach. Therefore, the model assumed a mean-field approximation, neglecting ion-ion correlation effects. However, the success of the TK model may be seen by the number of investigations where it has been invoked to study the interactions between charged ligands and proteins, membranes and other macromolecules (e.g.[47–50]).

Let us start reviewing a few key theoretical concepts. Consider the dissociation of a weak acid (HA),

$$HA \rightleftharpoons H^+ + A^-$$

with the corresponding thermodynamic equilibrium constant *at a given experimental condition*

$$K_a = \frac{a_{H^+} a_{A^-}}{a_{HA}} \tag{1}$$

where the a's are the activities. For an ideal system, a stoichiometric equilibrium constant (K_s) is often used

$$K_s = \frac{c_{H^+} c_{A^-}}{c_{HA}} \tag{2}$$

where the c's stand for concentrations. Deviations from the ideal behavior due to molecular interactions are effectively taken into account by the activity coefficients (γ)

$$a_i = \gamma_i c_i \tag{3}$$

It follows that

$$K_a = K_\gamma K_s = \frac{\gamma_{H^+}\gamma_{A^-}}{\gamma_{HA}} \times \frac{c_{H^+} c_{A^-}}{c_{HA}} \tag{4}$$

which can be re-written as

$$-logK_a = -log\big(\gamma_{H^+} c_{H^+}\big) - log\Big(\frac{\gamma_{A^-}}{\gamma_{HA}}\Big) - log\Big(\frac{c_{A^-}}{c_{HA}}\Big)$$

By definition, $pK_a = -logK_a$ and $pH = -log(a_{H^+}) = -log(\gamma_{H^+} c_{H^+})$ which yields to

$$pK_a = pH - log\Big(\frac{\gamma_{A^-}}{\gamma_{HA}}\Big) - log\Big(\frac{c_{A^-}}{c_{HA}}\Big) \tag{5}$$

or,

$$-ln\Big(\frac{c_{A^-}}{c_{HA}}\Big) = -ln\Big(\frac{\gamma_{HA}}{\gamma_{A^-}}\Big) - \Big(pH - pK_a\Big) ln10 \tag{6}$$

The term $-ln\big(\frac{c_{A^-}}{c_{HA}}\big)$ can be identified as the free energy between the protonated and deprotonated states ($\beta\Delta A_{HA\rightarrow A^-}$), where $\beta = 1/K_BT$, K_B ($= 1.3807 \times 10^{-23}\, J.mol^{-1}.K^{-1}$) is the Boltzmann constant and T is the temperature (in Kelvin).

From this physical chemical approach, an effective potential for the protonation/deprotonation process can be written as

$$w = \Delta E - \big(pH - pK_a\big) ln10 \tag{7}$$

where ΔE should describe all the molecular interactions that produce deviations from ideal behavior (e.g. interaction with other charged amino acids, counter-ions, added salt, etc.). The second term accounts for the free energy change of the (de)protonation process for a single amino acid, not affected by the presence of the rest of the protein, nor by the any other mobile charged (added salt and counter-ions). pH becomes a simple (input) parameter in this phenomenological approach.

The term ΔE can be obtained from the TK model. Accordingly to TK,[41,42] the electrostatic free energy (G^{el}) for a protein containing N_p ionizable sites with valency z_i immersed in an electrolyte solution in the absence of a dielectric inhomogeneity is given by[50]

$$G^{el} = \frac{e^2}{8\pi\epsilon_0} \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} z_i z_j (A_{ij} - C_{ij}) \quad , \tag{8}$$

where e is the elementary charge ($e = 1.602 \times 10^{-19} C$) and ϵ_0 is the the vacuum permittivity ($\epsilon_0 = 8.854 \times 10^{-12}\, C^2/Nm^2$).

The direct Coulombic interaction between the charges of the protein is accounted for by A_{ij}:

$$A_{ij} = \frac{1}{\epsilon_s r_{ij}} \quad ,$$

where ϵ_s and r_{ij} are, respectively, the solvent dielectric constant and the spatial separation distance between charges i and j. This term is independent of the salt concentration. The

effect of mobile counter-ions and added salt is described by the C_{ij} term. A critical discussion of these terms can be seen in ref.[42]

The C_{ij} term can be estimated by the excess chemical potential (μ^{ex}) for a free charged hard sphere ion with valency z and radius R_s using the Debye-Hückel (DH) theory of strong electrolytes,

$$\mu_F^{ex} = -\frac{\kappa z^2 e^2}{8\pi\epsilon_0\epsilon_s k_B T(1+2\kappa R_s)} \quad , \tag{9}$$

where κ is the inverse DH screening length which is proportional to the square root of the salt concentration. For a bulk number density of specie k equals to $n_{0,k}$,[51]

$$\kappa = \left[\frac{e^2}{\epsilon_0\,\epsilon_s\,k_B T}\sum_1^N n_{0,k}\,(z_k)^2\right]^{\frac{1}{2}} \tag{10}$$

where N is the total number of mobile charge species in the system.

The concept of κ as a scaling parameter that measures how effective is the Coulomb shielding is, has been revisited by different authors in the colloidal literature.[52–55] For protein electrostatics, it was observed that a simple modification of the definition of κ to include the counter-ions concentration (κ_c) in the summation of Eqn. (10), as suggested by Beresford-Smith and coworkers,[52] better describes the system.[50] Therefore, we followed this modified definition of κ to κ_c on the FPTS. As a consequence, κ_c now depends on the protein protonation state.

Based on these arguments, it was proposed that the titration process given by Eq. 7 can be modeled as[40]

$$w_{TK} = \frac{e^2}{4\pi\epsilon_0\,\epsilon_s}\Big[\sum_{i>j}^{N_p}\frac{z_i z_j}{r_{ij}} - \frac{Z_p^2\kappa_c}{2(1+\kappa_c b)}\Big] + \lambda(pH-pKa)ln10 \tag{11}$$

where $Z_p = \sum_i^{N_p} z_i$, λ equals either -1 (deprotonation) or $+1$ (protonation) and b is assumed to equal the radius of a sphere that inscribes the protein (R_s).

It is worth mentioning possible weaknesses of this scheme: a) assuming a spherical shape for the biomolecule when calculating its excess chemical potential to describe the salt effect (see Eqn. 9) might introduce artifacts for more elongated biomolecules although for both RNA molecules and lactoferrin (an elongated milk protein) studied earlier, only small deviations between computed and experimental quantities have been observed.[40,45] However, a well-known side effect is that pI predictions will be unaffected by the ionic strength since anisotropic-salt interactions are neglected.

b) The intrinsic pK_a values are taken from experimental data for "isolated" amino acids at a given experimental condition, and we only account for the *difference* in free energy between the residue in the protein and this reference state for which the pK_a was originally obtained. Doing so, all the interactions (solvation effects, dispersion, polarization, etc.) implicitly included in the pK_a measurement are assumed to be the same in both environments. Of course, this implies that sites deeply buried will probably not be accurately described (see the discussion below). Yet, as most of the titratable groups of proteins are close to the surface, this is expected to have only a minor effect on the majority of biomolecules. We will refer to this equilibrium constant of the amino acid model compounds as pK_0. From hereon, pK_a will be used here for the equilibrium constants when the titratable residue is at a particular protein conformation and salt solution.

c) The mesoscopic description of the amino acids as single beads decreases possible differences between rotamers. Titratable sites end up at similar positions when the amino acids are reduced to a spherical object. Although for protein-protein interactions this approximation will have virtually no effect when calculating free energy of interactions, due to the long range nature of electrostatic interactions, this has an impact on the precise description of hydrogen bonds. Specifically for pK_a studies, a natural direction to improve the results could be to combine a less coarse grained description of the amino acids with the formalism for multiple-site titration proposed by Beroza and co-authors[56] together with a proton isomerism titration scheme.[57]

Choice of pK_0 and buried amino acids

As mentioned above, intrinsic pK_a values are taken from experimental data for "isolated"amino acids, and this choice has a strong effect on deeply buried residues. Actually, this is an intrinsic feature of all numerical schemes that invoke a phenomenological approach as does the FPTS to describe pH effects. From a physical chemistry perspective, it is trivial to mathematically demonstrate that buried amino acids will always be a difficult case in this framework:

Consider a dissociation chemical reaction,

$$\mathrm{H\ A}_{prot} \rightarrow \mathrm{H}^{+} + \mathrm{A}^{-}_{prot} \qquad \Delta G_1$$

where the subscript *prot* is used to indicate that the amino acid A is located in the protein interior, and ΔG_1 is the free energy associated with the process at this given experimental condition. From basic physical chemistry, it follows that K_1 ($\Delta G_1 = -RTlogK_1$, where $R = 8.314 J/molK$ is the gas constant) is dependent of the experimental conditions (temperature, pressure, ionic strength, solvation, concentration of other species, etc.). Changing any of such conditions will clearly affect K as often measured by ΔpKs ($pK = -logK$).

If the reaction takes place in pure water instead, i.e., the amino acid A is not part of the protein structure, the equivalent reaction is:

$$\mathrm{H\ A}_{w} \rightarrow \mathrm{H}^{+} + \mathrm{A}^{-}_{w} \qquad \Delta G_2$$

where A_w indicates that A is in pure water, and ΔG_2 is the corresponding associated free energy. Note that the proton binding is just the opposite reaction for both situations.

Essentially, ΔG_1 and ΔG_2 are measuring the acid-base behaviour of this amino acid in different environments. Consequentially, $\Delta G_1 \neq \Delta G_2$. This is because the binding of a proton to A^{-}_{prot} changes the interaction of A^{-}_{prot} with the surrounding species, which are different in comparison to the pure water case. As a consequence, one can write that:

$$\Delta G_1 = \Delta G_2 + \Delta G_c$$

where ΔG_c reflects the free energy involved in electrostatic and non-electrostatic interactions changes. It is convenient to make a partition between electrostatic (ele) and other contributions (other), $\Delta G_c = \Delta G_c^{(ele)} + \Delta G_c^{(other)}$. When assuming that pK_a is the intrinsic pKvalue taken from experimental data for the isolated amino acid at a given experimental condition (as done by us), strictly speaking, we are setting $\Delta G_c^{(other)} = 0$. An approximation is introduced at this point since this pK_a value is rigorously valid *only* at this particular special situation. In a different environment, say deeply buried in a protein structure, solvation, for instance, will be very much different than the previous case (when the molecule was isolated), $pK_1 = pK_2 + pK_c$. As a result, $pK_1^{(other)} \neq pK_2^{(other)}$. During the model derivation, on Eqn. 1, a specific experimental condition is taken for grant (could be either 1 or 2, neglecting the term $pK_c^{(other)}$). Only computing the electrostatic terms means that all other contributions are assumed to be either the same in both conditions or their difference is quite small [$\Delta pK^{(other)} \approx 0$]. This is far from reality in the buried amino acids situation ($pK_1^{(other)} \neq pK_2^{(other)}$). Conversely, for residues closer to the surface, $\Delta G_c^{(other)}$, and this assumption is much less of an approximation ($pK_1^{(other)} \approx pK_2^{(other)}$).

Convergence properties of the FPTS

Despite its apparent limitations, the outcomes obtained by the FPTS for other macromolecular systems studied before are similar to or even better than many other theoretical approaches (see also the tables presented below here).[25,45] This is achieved at much lower computational cost. For this reason, the FPTS does offer an optimal compromise between accuracy and efficiency. It was already demonstrated that charges obtained by the FPTS converge to their equilibrium values in less than 10^5 MC steps for systems with strong interactions between titrating sites.[45] This is achieved in very short CPU time, orders of magnitude smaller even than the typical time observed for PB approaches. For instance,

Wang and co-authors quoted 9185.2s for the energy runtime in a single AMD Opteron 2356 processor (8 cores and 2.3 GHz) for the large protein 6-phosphogluconate dehydrogenase (PDB id 2zyg).[58] Using the FPTS, the CPU time for 10^5 MC steps decreases to 96 s in a personal notebook (Intel i7-3630QM and 2.40 GHz - running ubuntu 12.04). One should also see that the FPTS has a further advantage in comparison with the PB: contrary to the PB methods that only provide an averaged value for the titratable sites' charges at a given pH, instantaneous charges are directly obtained in a FPTS calculation. Also, due to the MC protonation/deprotonation process, these charges can fluctuate as a function of the solution pH. This is of fundamental importance when exploring the charge regulation mechanism that can be so relevant for protein complexation.[15,32,34,59]

Model and Methodology

Titration scheme

The fast proton titration scheme (FPTS) for proteins was described in detail in ref.[40] Here, we briefly review it including specific comments for the pKa calculation.

The protein is described at a mesoscopic level where amino acids are represented by charged van der Waals particles of radii (R_i) and valences z_i. For the sake of simplicity, internal degrees of freedom are neglected (i.e. bond lengths, angles and dihedral angles are kept fixed). Values for R_i were taken from ref.[43]

Glutamic acid (GLU), aspartic acid (ASP), tyrosine (TYR), cysteine (CYS) not involved in SS bridges, lysine (LYS), histidine (HIS), argine (ARG) and the C (CTR) and N (NTR) terminals have titratable groups. We employed in this work pK_0 values given by Nozaki & Tanford[27] for them. Their protonation states were allowed to change according to the solution pH, salt conditions and other charged amino acids (valences vary between -1 and 0 or 0 and $+1$, for acid and basic amino acids, respectively). All the other residues are assigned with a constant zero charge.

For the protonation/deprotonation process, Eq. 11 is converted into the following MC protocol:[40]

1. Titratable sites on the protein are placed at their experimentally determined positions and kept fixed. The protein crystal and/or NMR structure coordinates as given by RCSB Protein Data Bank (PDB)[60] define their positions.
2. A titratable site is chosen by random.
3. If protonated, try to move the proton *charge* to the bulk solution (deprotonation process). If deprotonated, try to move the proton *charge* from the bulk solution to the site (protonation process). κ_c should be updated to reflect the corresponding changes in the number of counterions in the electrolyte solution (see Eq. 10). In an analogous manner, Z_p is also updated for the appropriate protein net charge after the protonation/deprotonation process. 4. This trial charge movement is accepted with probability

$$min\Big(1, e^{(-\beta \Delta w_{TK})}\Big)$$

5. The process is repeated from step 2 until the convergence is reached. Several MC steps are necessary due to the interplay of multiple titratable sites present in the protein structure.

In the context of the present work, the main outcomes of each MC run are the average charge of each amino acid ($< q_{aa} >=< z_{aa} > e$) and the average total protein charge ($< Q_{total} >=< Z_p > e$). Other physical chemical properties such as the average total dipole moment ($< \mu_{total} >$) and the average protein charge fluctuation parameter [also known as protein charge capacitance ($< C_{total} >$)] can also be directly obtained.[40] A scheme of the simulation protocol can be seen in Fig. 1.

For most of the applications, a proper description of the amino acids charges and their fluctuations as a function of solution pH is needed. However, for benchmarking theoretical models, it is often used pK_a values which are not directly obtained from the MC run. From a physical chemical description, the pK_a value for each individual titratable amino acid at

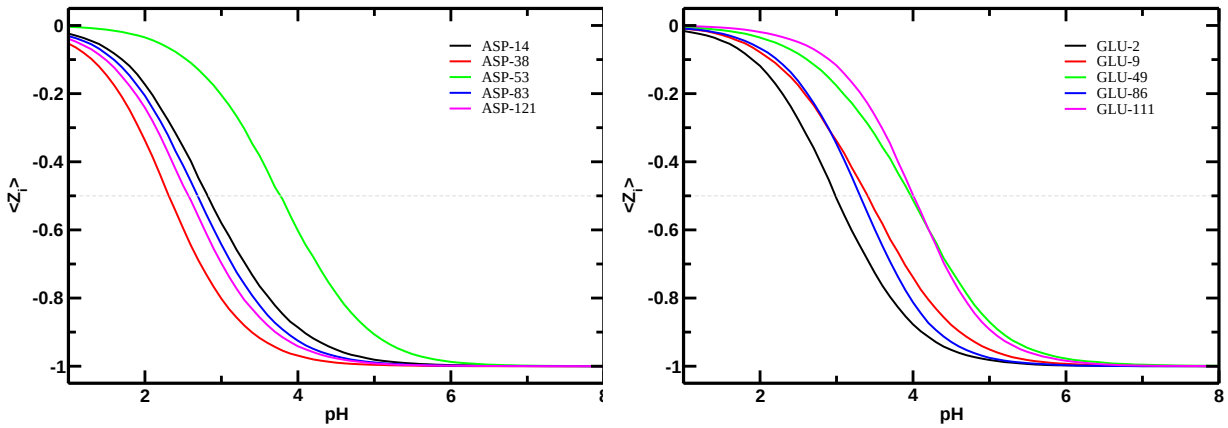


Figure 2: Computed titration plots of the acid amino acid residues ASP (left panel) and GLU (right panel) of ribonuclease A at 60mM salt concentration. The dashed gray lines indicate the half of the protonated states which is used to predicted the theoretical pK_a. Data are taken from the titration simulations with the FPTS. The intrinsic pK_0 values of the amino acid model compounds are 4.0 and 4.4, respectively, for ASP and GLU.[27]

a particular micro-environment (specified by the neighbor charges and salt conditions) is defined as the pH where this residue is half-protonated (e.g. $< z_i(pH) >= -0.5$, for GLU). This is measured from a titration plot. Therefore, it is necessary to generate a theoretical titration plot for each of the ionizable amino acids, to determine in what pH condition the titratable site is 50% occupied by the proton. This requires a series of simulations varying solution pH from 1 to 14 using small intervals (in this work we used 0.1) to produce the theoretical titration plot. From this series of simulations at different pH conditions and given salt concentration, one obtains a set of average charge numbers for all titratable residues ($\{< z_i(pH) >\}$). Observing in this set of charge $\times$ pH data when each residue is half-protonated leads to the prediction of its pK_a value, as it is done in wet laboratorial experiments – see Fig. 1. It is worth mentioning that this association constant is also particularly useful to quantify how specific micro-environments affects the proton binding reaction. It can be used to describe several physical events (e.g. protein stability, macromolecular assembly, binding of ligands, conformational changes, added salt effects, etc.) and their dependency on the environment.[8] Nevertheless, the knowledge of the pK_a does not directly give the charge fluctuation ($C_{total} =< Z_P^2 > - < Z_P >^2$).[40]

Proteins

Protein structures were all obtained from the RCSB Protein Data Bank (PDB).[60] For the sake of comparison with other theoretical works, missing residues in their studies were not included here. All PDB files were edited before the calculations. Water molecules and hetero atoms were removed. For atoms that had two records included in the PDB file due to their different occupancies, the higher occupancy case was selected, and the others deleted. Different sets of proteins were studied here. In the first set, the thermostable actin binding 36-residue subdomain of chicken villin headpiece – HP36 (PDB id 1VII), the 45-residue binding domain of 2-oxoglutarate dehydrogenase multi-enzyme complex – BBL (PDB id 1W4H) and the 129-residue hen egg white triclinic lysozyme – HEWL (PDB id 2LZT) were used for the comparison with both the GB and the REX-CpHMD method.[37] In another set, simulations were performed with the 124-residue ribonuclease A – RNase (PDB id 7RSA), the 135-residue staphylococcal nuclease – SNASE (PDB ids 3D6C, 2RKS and 2SNM), and the 122-residue α-lactalbumin – ALAC (PDB id 1F6S). Calcium ions were removed from the ALAC structure. This set of proteins is particularly interesting to explore the effect of the locations of the titratable sites in the protein conformation. In RNase, the comparison is done for surface residues. SNASE has buried amino acids whose pK_as should be more challenging for the present fast titration scheme to predict (particularly given the use of the reference pK_0). This system was also studied by the pHtMD, and gives us the opportunity to compare the calculated pK_as with another recent developed method.[5] ALAC is a good example of an intrinsically flexible protein chain. For the comparison with the hybrid neMDMC,[4] the 56-residue turkey ovumucoid third domain – OMTKY3 (PDB id 1OMU) and the 129-residue hen egg-white orthorhombic lysozyme – HEWL2 (PDB id 1AKI) were utilized. All low-energy NMR solution structures for OMTKY3 as given by PDB were used to define its amino acids positions. Each structure was submitted to an independent simulation run. Results were averaged with a uniform weight.

This selected set of proteins also offered the possibility to explore a variety of site-site interactions. The total number of ionizable sites in each protein was quite diverse, ranging from 12 to 55: (a) HP36 – 12, (b) BBL – 17, (c) HEWL – 40, (d) RNase – 36, (e) SNase – 54 (for PDB id 3D6C), 54 (for PDB id 2RKS) and 55 (for PDB id 2SNM), (f) ALAC – 49, (g) OMTKY3 – 16, and (h) HEWL2 – 32.

Monte Carlo simulations

We performed standard Metropolis Monte Carlo (MC) simulations[61,62] using the titration scheme above described for all protein sets. The aqueous solution dielectric constant and temperature were fixed at $\epsilon = 78.7$ and $T = 298$K, respectively. Solution pH was varied from 1 to 14. Each system was simulated in a specific salt concentration as given by the correspondent experimental data: (a) HP36, 150 mM, (b) BBL, 200 mM, (c) HEWL, 50mM, (d) RNase, 60mM, (e) SNASE, 100mM, (f) ALAC, 150mM, (g) OMTKY3, 10mM, (h) HEWL2, 100mM. The number of MC steps for production runs was at least 10^6 steps after equilibration (10^5 MC steps). Calculations were performed with the Faunus biomolecular simulation package,[63] where the FPTS was already implemented.[40]

For the sake of comparisons, additional calculations with PropKa (version 3.1) and default parameters[64] were also carried out. Together with the predicted pK_a values, PropKa provided indicators on how buried amino acids are in the folded protein conformation. Available published data obtained by the Generalized-Born (GB) method is included too. The quality of the outcomes from the FPTS and other theoretical methods is finally scrutinized by means of a comparison with the so-called "NULL model", where site-site interactions are altogether neglected.[65,66] In this model, any ionizable site titrates as if no other titratable site was present in the system, *i.e.* one assumes that the pK_a of a given amino acid at any experimental condition is identical to its model compounds given zero pK_a shifts ($\Delta pK_a = pK_a - pK_0 = 0$). In the current study, we used for these calculations the pK_0's given by Nozaki & Tanford[27] as done in the MC runs. The accuracy will be discussed in

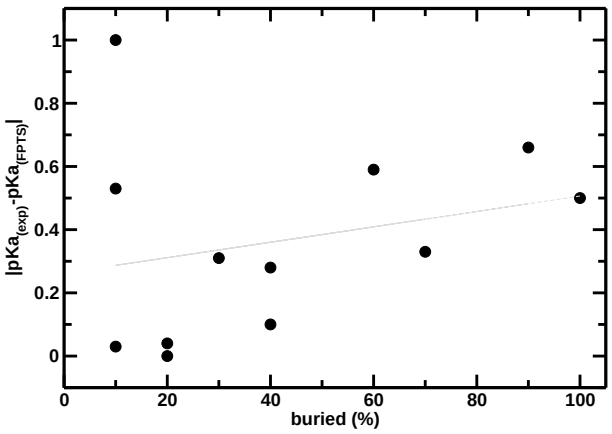


Figure 3: The effect of burial of amino acid residues on computed pK_a for ribonuclease A. Experimental and computed pK_as by the FPTS in absolute numbers are compared with the percentage of buried charges. This percentage was predicted with the PropKa software. Experimental and FPTS are the same ones given in Table 2. The linear regression line has a coefficient equal to 0.0024362. The correlation coefficient is 0.25.

terms of the maximum absolute deviation (MAX), the averaged absolute deviation (AAD), the root-mean-square deviation (RMSD) and the linear correlation coefficient (r) between the experimental and computed pK_as. Bold numbers in the tables indicate the cases where experimental and theoretical data have pKa shifts in opposite directions (e.g. $pK_{a,exp} - pK_0 > 0$ and $pK_{a,theoretical} - pK_0 < 0$, or the contrary) and/or no shift is observed with respect to the isolated behavior (e.g. $pK_{a,exp} - pK_0 > 0$ and $pK_{a,theoretical} - pK_0 = 0$). This is used to indicate if the model is able to correctly predict the protonation state. For practical use, the general trends are the results that matter most, that is, if the model suggests protonation for an amino acid that is actually found protonated in a given structure and vice versa.

Results and discussion

pK_a calculation for HP36, BBL and HEWL

Predicting precise pK_a values of proteins is a lively research field.[67] Our comparisons here are focused on the validation process of the fast protonation scheme aimed for including proton equilibria in multiscale simulations with many ionizable macromolecules. The FPTS was

Table 1: Calculated and experimental pK_a values of HP36, BBL and HEWL proteins. Salt concentrations were 150 mM for for HP36, 200 mM for BBL, and 50mM for HEWL. [a] Data taken from ref.[37] [b] Sampling was based on 1 ns. [c] The 10ns trajectory was broken in two halves (0-5 and 5-10ns). MAX, AAD, RMSD and r mean the maximum absolute, the average absolute and the root-mean-square (RMS) deviations and the linear correlation coefficient between the experimental and computed pK_as, respectively, in this table and the next ones.

Residue	Experiment[a]	GB [a]	All-atom REX-CpHMD [a,c]			PropKa	FPTS	NULL
		0-1[b]	0-5	5-10	0-10			
HP36								
Asp44	3.10(1)	3.2(1)	2.0	3.0	2.6(5)	3.78	3.7	4.0
Glu45	3.95(1)	3.5(1)	4.3	**4.5**	**4.4(1)**	**4.57**	**4.5**	4.4
Asp46	3.45(12)	3.5(1)	2.4	3.7	3.1(6)	**4.08**	3.8	4.0
Glu72	4.37(3)	3.5(1)	**4.4**	**4.4**	**4.4(0)**	**4.43**	3.5	4.4
	MAX	0.9	1.1	0.6	0.5	0.7	0.9	0.9
	AAD	0.4	0.6	0.2	0.3	0.5	0.6	0.5
	RMSD	0.5	0.8	0.3	0.4	0.6	0.6	0.6
BBL								
Asp129	3.88(2)	3.2(0)	2.2	3.2	2.7(5)	3.68	3.7	4.0
Glu141	4.46(4)	**4.3(0)**	**4.0**	**4.4**	**4.2(2)**	4.51	**3.8**	4.4
His142	6.47(4)	7.1(0)	**5.9**	**5.8**	**5.8(0)**	6.37	6.5	6.3
Asp145	3.65(4)	2.8(2)	3.0	3.1	3.1(0)	3.76	3.7	4.0
Glu161	3.72(5)	3.6(3)	4.2	3.9	4.0(2)	**4.59**	4.1	4.4
Asp162	3.18(4)	3.4(3)	2.9	3.5	3.2(3)	2.32	3.2	4.0
Glu164	4.50(3)	4.5(1)	5.7	4.6	5.2(6)	4.54	**3.8**	4.4
His166	5.39(2)	5.4(1)	4.4	4.4	4.4(0)	5.78	6.0	6.3
	MAX	0.9	1.7	1.0	1.2	0.9	0.7	0.9
	AAD	0.3	0.8	0.4	0.6	0.3	0.3	0.4
	RMSD	0.5	0.9	0.5	0.7	0.5	0.4	0.5
HEWL								
Glu7	2.6(2)	2.6(1)	3.6	3.4	3.5(1)	3.98	3.3	4.4
His15	5.5(2)	5.3(5)	5.1	5.1	5.1(0)	**6.71**	5.6	6.3
Asp18	2.8(3)	2.9(0)	2.5	3.3	2.9(4)	3.41	2.8	4.0
Glu35	6.1(4)	**4.4(2)**	8.5	8.7	8.6(1)	6.51	**3.5**	4.4
Asp48	1.4(2)	2.8(2)	0.1	1.1	0.6(6)	1.81	3.4	4.0
Asp52	3.6(3)	**4.6(0)**	**5.4**	**5.6**	**5.5(1)**	3.83	3.3	4.0
Asp66	1.2(2)	1.2(4)	0.6	0.8	0.3(7)	1.85	3.0	4.0
Asp87	2.2(1)	2.0(1)	0.8	2.1	1.5(7)	3.27	3.2	4.0
Asp101	4.5(1)	**3.3(3)**	6.1	5.7	5.9(2)	**3.92**	**2.9**	4.0
Asp119	3.5(3)	2.5(1)	3.0	3.3	3.2(1)	3.58	3.2	4.0
	MAX	1.7	2.4	2.6	2.5	1.4	2.6	2.8
	AAD	0.5	1.3	0.9	1.0	0.7	1.0	1.4
	RMSD	0.9	1.4	1.2	1.2	0.8	1.4	1.6
For these three proteins								
	MAX	1.7	2.4	2.6	2.5	1.4	2.6	2.8
	AAD	0.5	1.0	0.6	0.7	0.5	0.7	0.9
	RMSD	0.7	1.2	0.9	0.9	0.6	1.0	1.2
	r	0.86	0.87	0.86	0.87	0.92	0.67	0.67

compared with several other recent theoretical calculations (GB, REX-CpHMD, PropKa), the NULL model and available experimental data. Proteins are grouped here as a function of the available data found in the references. In table 1, we present the results for the comparison with published GB and REX-CpHMD data. PropKa predictions and the NULL model data are also included in this table. For these three proteins, PropKa is able to reproduce the experimental results with the smallest deviations followed by the GB method. Both the maximum absolute and root-mean-square (RMS) deviations are in line with previous observations.[39] The FPTS achieved similar performance in terms of the RMS deviations with these two methods while the observed maximum absolute deviation is compatible with the REX-CpHMD results. Each method shows relatively better agreement with the experimental data for a specific protein system. This will be illustrated on the following comparisons too. The smallest MAX, AAD and RMSD values for BBL was achieved with the FPTS. All theoretical methods based on molecular simulations have difficulties to reproduce well the HEWL-Glu35 experimental data due to the manner in which the pH is taken into account. This particular amino acid (Glu35) is deeply buried in the protein structure (66% according to PropKa) where the common approximation assumed for Eqn. 6 is to adopt pK_0 to describe the reference pK_a. The more deeply buried is the residue, the less reliable is this approximation. The reference (fully solvated environment) and the final configuration (poorly solvent exposed environment) states clearly have different microenvironments.

Following ref.,[2] the linear correlation coefficients between the experimental and computed pK_as were calculated and are given in the tables. The r results are strongly dependent on the protein system, as can be seen comparing data reported in all tables. For the sake of improving statistics, the three proteins (HP36, BBL and HEWL) are grouped together. Analyzing them (22 experimental points) as a single set, the empirical predictor PropKa gives the best r value (0.92). Both the CPHMD and the GB methods show $r \approx 0.86$ while both the FPTS and the NULL model give 0.67. A common argument for the discrepancy observed for rigid models like FPTS given is the lack of conformational changes that should accompany

Table 2: Calculated and experimental pK_a values of ribonuclease A. Salt concentrations was 60 mM. na = data not available. [a] Data taken from ref.[68] For deviations, only residues with available experimental data were used.

Residue	Experiment[a]	All-atom REX-CpHMD [a]	propKa	FPTS	NULL
Asp14	1.8	3.4	0.3	2.8	4.0
Asp38	2.1	3.0	2.7	2.3	4.0
Asp53	3.7	**4.0**	3.4	3.8	4.0
Asp83	3.3	3.2	3.9	2.7	4.0
Asp121	3.0	2.7	0.0	2.6	4.0
Glu2	2.6	3.6	0.6	3.0	4.4
Glu9	na	3.8	2.2	3.4	4.4
Glu49	4.3	3.3	2.1	4.0	4.4
Glu86	4.0	4.5	1.1	3.3	4.4
Glu111	na	3.4	2.4	4.0	4.4
His12	6.0	5.8	3.8	5.9	6.3
His48	6.1	4.9	5.4	**7.0**	6.3
His105	6.5	**6.4**	6.9	**6.3**	6.3
His119	6.5	**5.6**	7.6	6.4	6.3
	MAX	1.6	3.0	1.0	2.2
	AAD	0.7	1.5	0.4	1.1
	RMSD	0.8	1.7	0.5	1.1
	r	0.88	0.83	0.95	0.93

the changes in ionization states.[10] In fact, partially introducing the conformational effect via different configurations extracted from a MD trajectory generated with fixed charges improves the r values for RNA systems.[45]

It is well known that is a challenge for theoretical models is to beat the NULL model predictions,[69] making such comparison a real critical test in benchmark studies. For HP36, results obtained by the FPTS and GB are equivalent to the NULL model. The REX-CpHMD performs better in this respect than the other methods if all the trajectory (0-10ns) is used. For the initial 5ns, the REX-CpHMD demonstrates difficulties to come closer to the NULL model accuracy. Conversely, for the BBL system, the expensive REX-CpHMD has the worst accuracy when compared with NULL model, while the FPTS data indicate higher precision. All descriptors (MAX, AAD and RMSD) are smaller for the FPTS. An intermediate situation is observed for the HEWL system. Analyzing all the three proteins together, one can observe that the FPTS is able to predict pK_a's with a higher accuracy than the NULL model.

In terms of the correlation between the protonation states observed in the calculations and in the experiments, GB revealed the best agreement, followed by both the FPTS and REX-CpHMD methods. PropKa fails to predict the correct direction of the pK_a's trends in six cases (see the bold numbers in the table). Although good results were observed for propKa in terms of MAX, AAD and RMSD, it predicts three pK_a's in the opposite direction for HP36. For instance, the pK_a for Glu45 is shifted from 4.4 to 3.95 ($\Delta pK = +0.45$) in the experiments while propKa predicts a shift in the opposite direction 4.57 ($\Delta pK = -0.17$). For these three proteins, from the 22 possible values for comparison, propKa has a mistake in 6 of them (72.7% of success). Conversely, for the same set of data, 17 have the correct pK_a shifts signal predicted by the FPTS (77.3% of success).

pK_a calculation for ribonuclease A

Ribonuclease A was investigated before by the all-atom REX-CpHMD method.[68] This is an example of a protein system where MD based methods are in closer agreement with

experimental data then PropKa. This enzyme is interesting for benchmark studies because it has several surface amino acids with strongly shifted pK_as. Accordingly to Wallace & Shen,[68] it is an excellent system for testing the theoretical methods particularly in respect with the description of local electrostatic interactions. Three amino acids (GLU-2, ASP-14 and ASP-38) are known to exhibit the largest experimental pK_a shifts with respect to the ideal case (i.e. $pK_a - pK_0$ is larger than what is observed for other titratable groups) being critical to be reproduced by theoretical methods. This can be observed in the typical titration plots for the acid amino acids shown in Figure 2. Such graphics measure the degree of protonation, and are equivalent to the unprotonated fractions plots commonly reported in the literature.[68] Particularly, ASP-38 is highlighted in the literature due to its strong interaction with LYS-1, LYS-41 and ARG-10. All these residues GLU-2, ASP-14 and ASP-38 behavior are well reproduced by the FPTS as seen in table 2. Experimental pK_as are 2.6, 1.8 and 2.1, respectively, for GLU-2, ASP-14 and ASP-38. Computed pK_as for the same amino acids are overestimated, 3.6(-1.0), 3.4 (-1.6) and 3.0(-0.9), by the all-atom REX-CpHMD method, and, 3.0(-0.4), 2.8 (-1.0) and 2.3 (-0.2), by the FPTS. The differences between the experimental and theoretical results, given between the parenthesis, indicate that the FPTS is able to capture the experimental shifts of the critical ASP-38 (0% buried). The outcome is also good for GLU-2, another superficial amino acid (10%). All methods (including PropKa) have more difficulty to simulate ASP-14 (50% buried). In general, there is a *tendency* for the worst pK_a predictions to be for buried amino acids. This can be seen in Figure 3 where the difference between experimental and computed pK_a by the FPTS is compared with the the percentage of buried charges in the protein structure. Of course, other effects due to the assumed approximations in the model may affect the results too.

For this particular system, FPTS gives the smallest MAX (1.0), AAD (0.4) and RMSD (0.5) than any other method – see table 2. The linear correlation coefficient (0.95), based on the 12 experimental available points reported in this table, confirms the ability of the method to quantitatively describe the experimental data. All these descriptors are also much

Table 3: Calculated and experimental pK_a values of buried amino acids of staphylococcal nuclease. Salt concentrations was 100 mM. $\Delta_{REX-CpHMD}$, Δ_{FPTS} and Δ_{NULL} are defined as the difference between the theoretical model (REX-CpHMD, FPTS and NULL) and the experimental data.
[a] Data taken from ref.[68] [b] Original PDB file (3D6C) was modified with the mutation L38D. [c] Original PDB file (2SNM) was modified with the three mutations, P117G, H124L, and S128A. [d] Same as [c] with two additional modifications: substitutions of G50F and V51N and deletion of residues 44–49.

PDB id	Protein	Residue	Experiment[a]	All-atom REX-CpHMD [a]	$\Delta_{REX-CpHMD}$	FPTS	Δ_{FPTS}	NULL	Δ_{NULL}
3D6C	L38D/Δ+PHS [b]	Asp38	7.2	6.6	-0.6	**3.2**	-4.0	4.0	-3.2
3D6C	L38E/Δ+PHS	Glu38	7.0	6.9	-0.1	**3.9**	-3.1	4.4	-2.6
2RKS	L38K/Δ+PHS	Lys38	10.4	9.3	-1.1	**11.4**	1.0	10.4	0.0
2SNM	V66K/WT	Lys66	6.4	7.5	1.1	**11.0**	4.6	10.4	+4.0
2SNM	V66K/PHS[c]	Lys66	6.35	6.9	0.55	**11.0**	4.6	10.4	+4.1
2SNM	V66K/Δ+PHS[d]	Lys66	5.8	7.0	1.2	**11.0**	5.2	10.4	+4.6

smaller than the values of 2.2, 1.1 and 1.1 given by the NULL model, respectively, for MAX, AAD and RMSD. PropKa fails in this respect for this particular system, while it predicts the proper pK_a shift trend. The FPTS results for HIS48 and HIS105 are in opposite directions in relation to the experimentally observed shifts (-0.2 x $+0.7$ for HIS48 and $+0.2$ x 0.0 for HIS105). The all-atom REX-CpHMD method has difficulties with three residues (ASP53, HIS105 and HIS119).

pK_a calculation for SNASE

The general protocol observed in the literature to investigate the reliability of a given method is its application to proteins of different structural characteristics.[46,67,68] Methods for predicting pK_a values in biomolecules based on the differences between the the residue in the protein and the reference state for which pK_0 is originally obtained have a strong tendency to fail to compute pK_a for buried amino acids. This was already observed in Figure 3, and is primarily due to the desolvation energy that is not account in this approximation.

Staphylococcal nuclease contains buried groups, and as such is often used in benchmark studies. Deeply buried groups used before to test the sensibility of the all-atom REX-CpHMD method were employed here. Results are seen in table 3 for some mutants. The differences between the theoretical model (REX-CpHMD and FPTS) and the experimental data are in the same direction although the limitation of the FPTS for these amino acids

is evident. However, FPTS performs poorly for these specific amino acids. The accuracy of the FPTS is worst than the NULL model for these specific amino acids. Based on this data, the FPTS can only qualitatively predict these experimental shifts. The errors (up to 5 pH units) are the highest ones observed in all pK_a calculations done with the FPTS. Similar high MAX numbers (5.2 pK_a units) were reported before for the PropKa method.[46] Nevertheless, the FPTS can still be utilized to provide insight into the physical mechanisms of biological processes (e.g. the understanding of the functions of enzymes with buried charged amino acids), and describe the charge fluctuations entailed in the KS complexation mechanism.

A more complete investigation of the accuracy of the FPTS for SNASE is done in Table 4 where the available experimental data for other residues are used to benchmark some theoretical pK_a predictors. This analysis includes the data from the NULL model and the pH-titration MD (pHtMD) scheme, another variation of the CpHMD methods based on the performance of a consecutive series of MD simulations with small pH changes.[5] As can be seen, the results obtained by the FPTS are comparable with the other methods in terms of the usual observed deviations. As a matter of fact, the outcomes are slightly better than the PropKa data for both MAX (4.3 x 3.4), AAD (0.9 x 0.7) and RMSD (1.3 x 1.0). In this case, the r values revealed a good correlation for the computed pK_as by the FPTS. Considering only the same residues employed to analyse the pHtMD method ($r = 0.85$), the best correlation coefficient is observed for the FPTS data ($r = 0.88$). r also improves from 0.33 to 0.53 for PropKa. In fact, excluding ASP-19, ASP-21 and ASP-40 from the deviation analysis decreases the FPTS MAX, AAD and RMSD to 1.2, 0.5 and 0.5, respectively. These three amino acids are deeply buried in the protein interior and gives the highest deviations [ASP-19 is 76% (1.4), ASP-21 is 98% (3.4), and ASP-40 is 67% (0.6)]. Once more, this finding corroborates the previous trend for buried charges.

In terms of RMSD and AAD, the accuracy of the FPTS for this system (RMSD=1.0 and AAD=0.7) is slightly better than the NULL model (RMSD=1.1 and AAD=0.9) and at least equivalent to the REX-CpHMD pK_a data (RMSD=1.0 and AAD=0.8). The pHtMD method

Table 4: Calculated and experimental pK_a values of staphylococcal nuclease (SNase Δ+PHS). Salt concentrations was 100 mM. Calculations were done with the PDB structure 3BDC. For deviations and the linear correlation coefficients, only residues with available experimental data were used (ASP-77 and ASP-83 were not included due to their uncertainties). The number of experimental points utilized for the r calculations are given between parenthesis. na = data not available. [a] Data taken from ref.[5] [b] This analysis was carried out only with the smaller number of experimental points (14) used for the r calculation for the pHtMD method (ASP-19, ASP-21 and ASP-40 were removed from the correlation calculation).

Residue	Experiment[a]	pHtMD[a]	REX-CpHMD[a]	PROPKA[a]	FPTS	NULL
Asp19	2.21	na	4.1	**4.22**	3.61	4.0
Asp21	6.54	na	na	**2.29**	**3.17**	4.0
Asp40	3.87	na	3.1	**4.04**	3.3	4.0
Asp77	< 2.2	2.64	3.6	2.25	2.61	4.0
Asp83	< 2.2	na	na	2.72	2.55	4.0
Asp95	2.16	3.23	3.6	2.69	2.67	4.0
Glu10	2.82	3.83	4.4	3.7	3.46	4.4
Glu43	4.32	4.03	na	**4.96**	**4.69**	4.4
Glu52	3.93	3.77	4.3	3.87	4.3	4.4
Glu57	3.49	3.74	4.3	**4.41**	4.06	4.4
Glu67	3.76	3.84	4.39	3.61	3.33	4.4
Glu73	3.31	3.84	4.2	**4.51**	3.31	4.4
Glu75	3.26	4.16	4.0	3.65	3.44	4.4
Glu101	3.81	3.69	3.5	**5.25**	3.54	4.4
Glu122	3.89	3.27	3.8	3.83	3.51	4.4
Glu129	3.75	3.76	4.28	**4.48**	2.95	4.4
Glu135	3.76	3.58	4.2	3.27	3.22	4.4
His8	6.50	**6.08**	na	**6.29**	6.37	6.3
His121	5.25	5.89	na	**6.43**	**6.49**	6.3
	MAX	1.1	1.9	4.3	3.4	2.5
	AAD	0.4	0.8	0.9	0.7	0.9
	RMSD	0.6	1.0	1.3	1.0	1.1
	r	0.85 (14)	-0.03 (13)	0.33 (17)	0.59 (17)	0.58 (17)
				0.53 (14)[b]	0.88 (14)[b]	0.87 (14)[b]

Table 5: Calculated and experimental pK_a values of α-lactalbumin. Salt concentrations was 150 mM. [a] Data taken from ref.[68] For deviations and the linear correlation coefficients, only residues with available experimental data were used (GLU-1, ASP-78, ASP-82 and ASP-88 were not included). The number of experimental points utilized for the r calculations is 15.

Residue	Experiment[a]	All-atom REX-CpHMD [a]		PROPKA	FPTS		NULL
		25^{o}C	43^{o}C		25^{o}C	43^{o}C	
Glu1	na	3.8	3.8	4.6	4.9	4.9	4.4
Glu7	4.9	4.8	4.6	**3.6**	**4.2**	**4.1**	4.4
Glu11	4.7	4.6	4.6	4.8	**4.3**	**4.3**	4.4
Asp14	3.5	2.9	2.9	3.8	3.2	3.2	4.0
Glu25	4.9	**3.9**	**4.0**	**2.3**	**3.9**	**3.9**	4.4
Asp37	4.2	**3.7**	**3.8**	**2.3**	4.1	4.1	4.0
Asp46	3.8	2.7	2.6	2.5	3.8	3.8	4.0
Glu49	4.0	**5.0**	**4.9**	**4.9**	**4.4**	**4.4**	4.4
Asp63	4.5	4.4	4.3	**1.8**	**3.7**	**3.7**	4.0
Asp64	4.1	**3.0**	**2.9**	**3.8**	4.1	4.1	4.0
Asp78	na	2.3	2.7	2.9	3.6	3.6	4.0
Asp82	na	4.6	4.7	3.5	4.4	4.5	4.0
Asp83	4.5	**3.3**	**3.1**	**4.0**	4.5	4.6	4.0
Asp84	4.1	**2.5**	**2.4**	**2.4**	4.5	4.5	4.0
Asp87	4.4	**3.7**	**3.8**	**2.0**	4.1	4.1	4.0
Asp88	na	5.9	5.9	9.1	4.2	4.2	4.0
Asp97	3.5	2.6	2.5	3.4	3.1	3.1	4.0
Glu113	4.1	3.9	3.7	**4.5**	3.8	3.8	4.4
Asp116	3.5	2.6	2.5	3.8	3.7	3.7	4.0
	MAX	1.6	1.7	2.7	1.0	1.0	0.5
	AAD	0.7	0.8	1.1	0.4	0.4	0.4
	RMSD	0.9	0.9	1.4	0.5	0.5	0.4
	r	0.68	0.68	-0.16	0.58	0.58	0.53

shows higher accuracy for this specific system, and propKa the worst one. PropKa has also predicted opposite pK_a shifts directions for several residues. Asp21 and Glu43 are the only two cases where the FPTS gives the opposite experimental behavior for the pK_a shifts. His8 is the critical case for calculated amino acids with the pHtMD method. Data is not available for four cases including Asp19, Asp21 and Asp40 where deviations were observed for the PropKa. It is unclear how the pHtMD method would perform for these residues.

pK_a calculation for ALAC

α-lactalbumin is a calcium metalloprotein with 123 amino acids with several functional properties (e.g. apoptosis and induction of cell growth inhibition) and industrial applications.[70–73] From a structural point of view, it exhibits conformational fluctuations that result in its electrostatic properties similar to those of an intrinsically disordered protein structure.[68] Since FPTS relies on a rigid protein model, systems like ALAC with high structural fluctuations represent an additional challenge.

Experimental pK_as and the corresponding values calculated by the the NULL model, all-atom REX-CpHMD, PropKa and FPTS methods are compared in table 5. Unexpectedly due to its simplifications, the deviations obtained by the FPTS (RMSD=0.5) are smaller than the others [RMSD(REX-CpHMD)=0.68 and RMSD(PropKa)=1.4]. Computed pK_as by the all-atom REX-CpHMD scheme are similar to the FPTS who has the best accuracy for this system when comparing the three descriptors (1.0 x 1.6 for MAX, 0.4 x 0.7 for AAD, and 0.5 x 0.9 for RMSD). The qualitative analysis of the titration confirms these behaviour. Only five amino acids (GLU7, GLU11, GLU25, GLU49 and ASP63) have their pK_a shifts contrary to the experiments. They are shifted down while in the experimental measurements they were shifted up. The discrepancy increases to seven cases for the REX-CpHMD method. On the other hand, PropKa is the theoretical method that shows the greatest difficulties to reproduce the experimental shifts for this protein both in terms of the quantitative and qualitative descriptors. It even gives a negative r value (-0.16) and ten cases are in the wrong pK_a shift direction. All methods provide results with less accuracy than the NULL model. The FPTS is the only one whose outcomes are quite closer to the NULL model.

pK_a calculation for OMTKY3 and HEWL2

Two other proteins, the turkey ovumucoid third domain and the hen egg-white orthorhombic lysozyme, were included in this benchmark study. Rather than detailing their biological functions or structural features, the main reason to study these two systems here is the

Table 6: Calculated and experimental pK_a values of OMTKY3 and HEWL2 proteins. Salt concentrations were 10 mM for for OMTKY3, and 100mM for HEWL2. The mean and standard deviations of the calculated FPTS pK_a values for OMTKY3 were obtained from the results of all 50 NMR structures available in the PDB coordinates (PDB id 1OMU) as done in ref.[10] [a] Experimental data from refs.[74] and[75] for OMTKY3 and HEWL2, respectively. [b] The theoretical data for the hybrid neMD−MC was taken from ref.[4] For OMTKY3, the data is based on the averaged result for 7 simulations. For deviations and the linear correlation coefficients, all available experimental points (15) were used.

Residue	Experiment[a]	hybrid neMD−MC[b]	propKa	FPTS	NULL
OMTKY3					
Asp7	2.7	3.43	3.43	3.34(7)	4.0
Asp27	2.3	**4.27**	3.69	3.15(10)	4.0
Glu10	4.1	4.04	**5.03**	3.98(6)	4.4
Glu19	3.2	3.53	4.14	3.38(9)	4.4
Glu43	4.8	**4.39**	4.64	**4.12(3)**	4.4
	MAX	1.97	1.35	0.85	1.7
	AAD	0.70	0.83	0.49	0.98
	RMSD	0.97	0.92	0.57	1.12
HEWL2					
Asp18	2.66	2.74	3.39	3.2	4.0
Asp48	1.6	1.41	2.07	3.41	4.0
Asp52	3.68	3.99	**4.73**	3.56	4.0
Asp66	0.9	0.83	1.98	3.37	4.0
Asp87	2.07	3.03	2.31	3.46	4.0
Asp101	4.09	**3.57**	4.08	**3.42**	4.0
Asp119	3.2	3.06	2.99	3.49	4.0
Glu7	2.85	2.86	2.98	3.75	4.4
Glu35	6.2	**3.98**	6.37	**3.8**	4.4
His15	5.36	3.85	**6.31**	5.74	6.3
	MAX	2.22	1.08	2.47	3.1
	AAD	0.60	0.50	1.10	1.43
	RMSD	0.93	0.64	1.37	1.68
For these two proteins					
	MAX	2.2	1.4	2.5	3.1
	AAD	0.6	0.6	0.9	1.3
	RMSD	0.9	0.7	1.2	1.5
	r	0.73	0.94	0.62	0.56

possibility to compare the FPTS results with the recent developed hybrid neMDMC.[4] With regard to the fundamental theory supporting this method, it is expected to accurately sample the coupling of conformational dynamics, titratable sites and mobile explicit ions with improved pK_a convergence. Table 6 shows the outcomes from experiments and the computed pK_a values by hybrid neMD−MC, PropKa, the NULL model and FPTS. For OMTKY3, this data confirms that FPTS is able to reproduce experimental shifts (RMSD= 0.57) even better than more sophisticated methods for some systems regardless of the model approximations adopted to speed up calculations. PropKa gives deviations at an intermediate level (RMSD= 0.92) in comparison with other theoretical schemes. From this result, apparently, the convergence was probably not reached by the hybrid neMD−MC method for this specific system (RMSD= 0.97). In contrast to the trend that was observed above for buried residues, ASP-27 (0% buried) is the amino acid that seems to be the most difficult case for all three theoretical methods. Probability, it is an amino acid where the changes in the protonation state induce conformational variations that are neither described in the schemes using a rigid protein description (as the FPTS) nor enough sampled in the constant-pH MD approaches due to the slow convergence of them. All theoretical methods perform better than the NULL model.

For HEWL2, the FPTS has, to a small degree, higher deviations in comparison with the other methods. ASP-66 (13% buried) and GLU-35 (68% buried) are the two amino acids that present the maximum deviations (−2.5 and 2.4, respectively). The hybrid neMDMC method also has difficulties to reproduce the GLU-35 pK_a value. The difference between the experimental and computed pK_a is 2.2 for this amino acid. PropKa achieved the lower deviations for this protein. Curiously, the maximum deviation in PropKa data is observed for ASP-66 (−1.1) as seen for the FPTS. Despite the quantitatively small discrepancies, FPTS preserves the main protonation features of the other theoretical methods being able to describe the pK_a shift directions. Similar number of cases with pK_a shifts inverted are observed for all the theoretical methods (three for both propKa and FPTS, and four for

Table 7: Summary of the deviations observed by different theoretical methods. All points were taken from the previous tables. Following them, some protein systems were grouped together. The qualitative analysis is expressed in terms of the number of cases where pK_a shifts are in the opposite direction in comparison with experiments divided by the total number of available experimental points. The number between parenthesis expresses the percentage to fail in this qualitative criterion. [a] Data for 25^{o}C. [b] Calculated excluding some residues. See text for details. [c] The CPU time is given for a single run with 10^7 MC steps in a personal notebook [Intel i7-3630QM and 2.40GHz (4788.95/per processor) – running ubuntu 12.04] for the FPTS. [d] Some ionizable amino acids were not calculated. Missing data for 4 (for pHtMD) or 5 (for REX) residues.

				system		
Descriptor	**method**	**HP36/BBL/HEWL**	**RNase**	**SNASE**	**ALAC**[a]	**OMTKY3/HEWL2**
MAX	REX-CpHMD	2.5	1.6	1.9	1.6	
	pHtMD			**1.1**		
	neMD−MC					2.2
	GB	1.7				
	PropKa	**1.4**	3.0	4.3	2.7	**1.4**
	FPTS	2.6	**1.0**	3.4	1.0	2.5
	NULL model	2.8	2.2	2.5	**0.5**	3.1
AAD	REX-CpHMD	0.7	0.7	0.8	0.7	
	pHtMD			**0.4**		
	neMD−MC					**0.6**
	GB	**0.5**				
	PropKa	**0.5**	1.5	0.9	1.1	**0.6**
	FPTS	0.7	**0.4**	0.7	**0.4**	0.9
	NULL model	0.9	1.1	0.9	**0.4**	1.3
RMSD	REX-CpHMD	0.9	0.8	1.0	0.9	
	pHtMD			**0.6**		
	neMD−MC					0.9
	GB	0.7				
	PropKa	**0.6**	1.7	1.3	1.4	**0.7**
	FPTS	1.0	**0.5**	1.0	0.5	1.2
	NULL model	1.2	1.1	1.1	**0.4**	1.5
r	REX-CpHMD	0.87	0.88	-0.03	**0.68**	
	pHtMD			**0.85**		
	neMD−MC					0.73
	GB	0.86				
	PropKa	**0.92**	0.83	0.33/0.53[b]	-0.16	**0.94**
	FPTS	0.67	**0.95**	0.59/0.88[b]	0.58	0.62
	NULL model	0.67	0.93	0.58/0.87[b]	0.53	0.56
qualitative	REX-CpHMD	5/22 (22.7%)	3/12 (25.0%)	**0/14 (0%)**[d]	7/15 (46.7%)	
	pHtMD			1/15 (6.7%)[d]		
	neMD−MC					4/15 (26.7%)
	GB	**4/22 (18.2%)**				
	PropKa	6/22 (27.3%)	**0/12 (0%)**	10/19 (52.6%)	10/15 (66.7%)	**3/15 (20.0%)**
	FPTS	5/22 (22.7%)	2/12 (16.7%)	3/19 (15.8%)	**5/15 (33.3%)**	**3/15 (20.0%)**
	CPU costs (s)[c]	3/4/9	10	13	12	4/9

the hybrid neMDMC method). Analyzing together the two proteins reinforce this capacity. FPTS results are closer to the ones produced by the hybrid neMDMC method at much lower CPU cost. For the HEWS2 system, the CPU time for 10^7 MC steps is 11 s per experimental condition in a personal notebook (Intel i7-3630QM and 2.40GHz). As observed for ALAC, all theoretical methods are able to provide pK_a predictions with higher accuracy than the NULL model.

Summary of pK_a calculations for all studied systems

Figure 4 shows the correlation between experimental and calculated pK_as using the FPTS and PropKa methods. These scatter plots showing all available data suggests at least an equivalent predictive capacity between the FPTS and other popular theoretical methods. There is a slight tendency for the FPTS to be closer to the experimental measurements. From this Figure, one can see that the slope given by the FPTS comes nearer to the ideal line (given in dashed). The MAX, AAD and RMSD are, respectively, 5.2, 0.8 and 1.5, for FPTS using all 87 available points; 3.4, 0.6 and 0.9, for FPTS using the same 81 points as in PropKa analysis (excluding the data from table 3); and 4.9, 1.0 and 1.4, for PropKa using 81 points. As a reference to analyse these numbers, in another benchmark study, the results for PropKa using a larger set of proteins were quite similar: 5.2, 1.0 and 1.4.[46] Accordingly to this work, CpHMD produces very close performance (5.1, 1.0 and 1.4).[46] We note that outcomes for RNA systems obtained by the FPTS are slightly more precise.[45]

The results for all studied proteins are summarized in Table 7, which also includes the computation time for each system employing the FPTS. This comparison indicates that each method performs better for a specific protein. In spite of the fact that the FPTS might have difficulties to yield the actual value of a specific experimental pK_a, it gives reasonable agreement with experiments and at least equivalent performance to more elaborated methods. The best reproducibility is seen for RNase, SNASE and ALAC. RNase is a good example where the FPTS performs better than any other approach. The values reported are in within the typical range of values given by different theoretical models.[37] Except for ALAC, where the FPTS predictions have slightly less accuracy [RMSD(FPTS)=0.5 and RMSD(NULL)=0.4], FPTS performs better for all studied systems than the NULL model. In fact, for this system, the other theoretical methods are much slower and the corresponding predictions are poor [RMSD(REX-CpHMD)=0.9 and RMSD(propKa)=1.4]. A qualitative analysis of the pK_a shifts calculated by REX-CpHMD, propKa and the FPTS (the three methods with more available data in table 7 for a better statistics) verifies that the FPTS predicts experimental

trends with less errors than the others (21.7% for FPTS, 23.8% for REX-CpHMD, and 34.9% for propKa). Together with its fast convergence and cheap CPU costs,[45] this confirms the robustness and ability of FPTS to properly describe the main system's physics.

The reduction in computational time achieved by the FPTS represents a considerably advantage of this method. A single run for a given protein structure in an electrolyte solution takes $\approx$ 10s as seen in Table 7. This permits the repetition of the calculation for a much larger set of protein conformations (as done here by all low-energy NMR solution structures for OMTKY3), salt concentrations, mutations, etc., and, even more important, the study of *multi* titrating objects containing several ionizable sites. Note, however, that this single calculation provides $< z_{aa} >$ and $< Z_p >$ and not directly pK_a. Note also that the simulation has to be repeated for equilibration and production phases, at least. Therefore, the full set of simulations to compute the pK_a is higher than the numbers given in this table.

In terms of CPU performance, PropKa is by far the faster method. A typical run for a single protein chain takes less than 1s of user time in the same Linux box (data not shown in this table). Nevertheless, it provides the pK_as as its main output and not charges. PropKa does not evaluate the electrostatic interactions of the titratable site with all *extra-molecular* charges in the system. Conversely, in a MC run with the FPTS, charges are directly obtained, and can fluctuate as a function of the solution pH. Extra molecular electric fields from other charged species present in the solution are taken into account contributing to the titration acceptance criterion.[40] This gives the real opportunity to explore the important charge regulation mechanisms so relevant for protein complexation.[15,32,34,59] A further development of the present method is the coupling with a MD engine, a on going work that will contribute to bring more information for the understanding of the electrostatic world of the biomolecules.

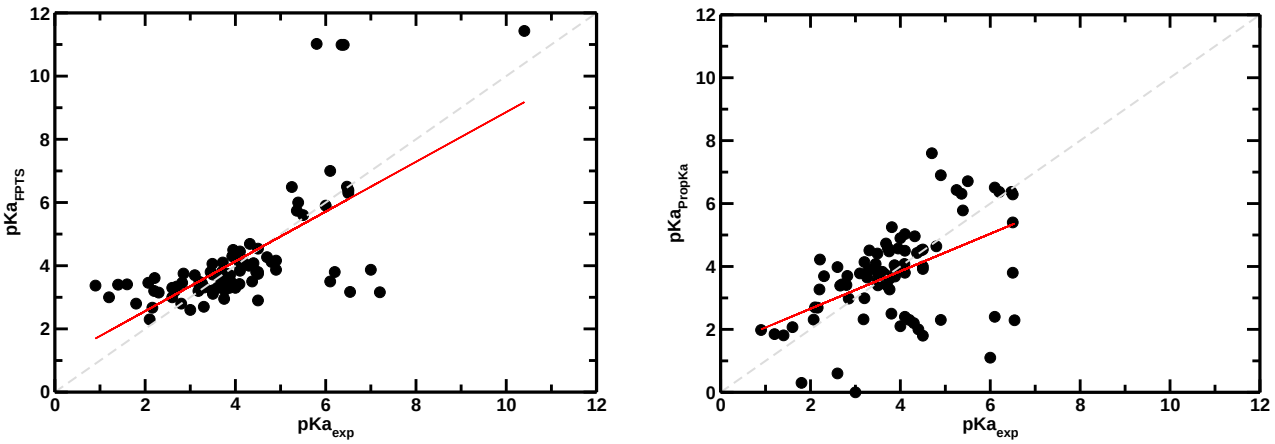


Figure 4: Correlation between experimental and calculated pK_as using the FPTS (top panel) and PropKa (bottom panel) methods. All points were taken from the previous tables (1–6) for the cases where there was both an experimental and predicted data ($n = 87$ for FPTS and $n = 81$ for PropKa). Fitted regression lines are given in red. The linear correlation coefficients and slopes are equal to 0.68 and 0.79, for FPTS; 0.51 and 0.60, for PropKa.

Conclusion

The accurate prediction of protonated amino acids by the FPTS was critically tested for several proteins with a diversity of structural features. Since experimental pK_a values compare relatively well with those calculated by the FPTS, this suggests that this fast Coarse Grained method is able to properly describe the main physics of the studied protein systems. Deviations are found to be within the typical range of values given by other theoretical models. Some ionizable groups are shifted to the acid or basic regimes as also happens with other computational methods. pK_a can be substantially shifted for deeply buried amino acids, because this scheme relies on the phenomenological assumptions given by the use of pK_0s. Features like conformational changes in response to the switch in the titration state are not incorporated in the model and can also affect the quality of the results. The average, maximum absolute and root-mean-square deviations were measured as $[0.4-0.9]$, $[1.0-5.2]$ and $[0.5-1.2]$ pH units, respectively. The present data might also contribute with the understanding of factors that affect the accuracy of theoretical pK_as leading to the improve of any biomolecular electrostatics method. It is important to stress that we rely on the available

protein structures, and the comparison is carried out with experimental data also subject to uncertainties.

uncertainties. The gain in CPU performance is clearly a great advantage of the present method. It permits its application to much larger systems and a proper sampling of the complex electrostatic coupling between the ionizable sites and other charges and also opens up opportunities for the study of multibiomolecular systems. The latter is a real achievement for the simulation of such systems. Other molecular simulation methods based on CpHMD are still prohibitive in this context even with massive computational resources due to their slow convergence that will become more critical when more biomolecules are included in the simulation box.

Acknowledgement

This work has been supported in part by the *Fundação de Amparo à Pesquisa do Estado de São Paulo* [Fapesp 2015/16116-3 (FLBDS)] and the *University College Dublin* (UCD) through a visiting professors grant (Seed Funding). FLBDS thanks also the support of the University of São Paulo through the NAP-CatSinQ (Research Core in Catalysis and Chemical Synthesis) and the hospitality of the UCD School of Physics, UCD Institute for Discovery and CECAM-IRL. DMK acknowledges the support of the European Commission through H2020 project number 676531 (E-CAM).

References

(1) Baptista, A. M.; Marte, P. J.; Petersen, S. B. Simulation of Protein Conformational Freedom as a Function of pH: Constant-pH Molecular Dynamics Using Implicit Titration. *Proteins: Struc., Func., and Genetics* **1997**, *27*, 523–544.

(2) Goh, G. B.; Knight, J. L.; Brooks, C. L. pH-dependent dynamics of complex RNA macromolecules. *J. Chem. Theory Comput.* **2013**, *9*, 935–943.

(3) Chen, W.; Morrow, B. H.; Shi, C.; Shen, J. K. Recent development and application of constant pH molecular dynamics. *Mol. Sim.* **2014**, *40*, 830–838.

(4) Chen, Y.; Roux, B. Constant-pH Hybrid Nonequilibrium Molecular DynamicsMonte Carlo Simulation Method. *J. Chem. Theory Comput.* **2015**, *11*, 3919–3931.

(5) Socher, E.; Stich, H. Mimicking titration experiments with MD simulations: A protocol for the investigation of pH-dependent effects on proteins. *Scientific Reports* **2016**, *22523*, 1–12.

(6) Donnini, S.; Ullmann, R. T.; Groenhof, G.; Grubmll̈er, H. Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer. *J. Chem. Theory Comput.* **2016**, *12*, 1040–1051.

(7) Stigter, D.; Dill, K. A. Charge effects on folded and unfolded proteins. *Biochemistry* **1990**, *29*, 1262–1271.

(8) Garcia-Moreno, B. Probing structural and physical basis of protein energetics linked to protons and salt. *Methods in Enzymnology* **1995**, *259*, 512–538.

(9) Harano, Y.; Kinoshita, M. On the physics of pressure denaturation of proteins. *J. Phys.: Condens. Matter* **2006**, *18*, L107L113.

(10) Tang, C. L.; Alexov, E.; Pyle, A. M.; Honig, B. Calculation of pKas in RNA: on the structural origins and functional roles of protonated nucleotides. *J. Mol. Biol.* **2007**, *366*, 1475–1496.

(11) Thaplyal, P.; Bevilacqua, P. C. Experimental Approaches for Measuring pKas in RNA and DNA. *Methods in Enzymology* **2014**, *549*, 189–219.

(12) Warshel, A. Electrostatic basis of structure-function correlation in proteins. *Acc. Chem. Res.* **1981**, *14*, 284–290.

(13) Roca, M.; Messer, B.; Warshel, A. Electrostatic contributions to protein stability and folding energy. *FEBS letters* **2007**, *581*, 2065–2071.

(14) Lizatović Robert,; Aurelius Oskar,; Stenström Olof,; Drakenberg Torbjörn,; Akke Mikael,; Logan Derek T.,; André Ingemar, A De Novo Designed Coiled-Coil Peptide with a Reversible pH-Induced Oligomerization Switch. *Structure* **2016**, *24*, 946–955.

(15) Da Silva, F. L. B.; Jönsson, B. Polyelectrolyte-protein complexation driven by charge regulation. *Soft Matter* **2009**, *5*, 2862–2868.

(16) Stoll, S. Computer Simulations of Soft Nanoparticles and Their Interactions with DNA-like Polyelectrolytes. Soft Nanoparticles for Biomedical Applications. Londres, 2014; pp 342–371.

(17) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153–159.

(18) Lund, M.; Jönsson, B. Charge regulation in biomolecular solution. *Quarterly Reviews of Biophysics* **2013**, *46*, 265–281.

(19) Delboni, L.; da Silva, F. L. B. On the complexation of whey proteins. *Food Hydrocolloids* **2016**, *55*, 89–99.

(20) Ye, K.; Malinina, L.; Patel, D. Recognition of small interfering RNA by a viral suppressor of RNA silencing. *Nature* **2003**, *426*, 874–878.

(21) Koukiekolo, R.; Sagan, S. M.; Pezacki, J. P. Effects of pH and salt concentration on the siRNA binding activity of the RNA silencing suppressor protein p19. *FEBS Letters* **2007**, *581*, 3051–3056.

(22) Chen, K.; Xu, Y.; Rana, S.; Miranda, O. R.; Dubin, P. L.; Rotello, V. M.; Sun, L.; Guo, X. Electrostatic Selectivity in Protein-Nanoparticle Interactions. *Biomacromolecules* **2011**, *12*, 2552–2561.

(23) Steiner, E.; Gastl, M.; Becker, T. Protein changes during malting and brewing with focus on haze and foam formation: a review. *Eur. Food Res. Technol.* **2011**, *232*, 191–204.

(24) Egan, T.; O'Riordan, D.; O'Sullivan, M.; Jacquier, J.-C. Cold-set whey protein microgels as pH modulated immobilisation matrices for charged bioactives. *Food Chemistry* **2014**, *156*, 197–203.

(25) da Silva, F. L. B.; Pasquali, S.; Derreumaux, P.; Dias, L. G. Electrostatics analysis of the mutational and pH effects of the N-terminal domain self-association of the Major Ampullate Spidroin. *Soft Matter* **2016**, *12*, 5600–5612.

(26) Wagoner, T.; Vardhanabhuti, B.; Foegeding, E. A. Designing Whey ProteinPolysaccharide Particles for Colloidal Stability. *Annu. Rev. Food Sci. Technol.* **2016**, *7*, 93–116.

(27) Nozaki, Y.; Tanford, C. Examination of Titratation Behavior. *Methods Enzymol.* **1967**, *11*, 715–734.

(28) Bashford, D. In *Scientific Computing in Object-Oriented Parallel Environments: First International Conference, ISCOPE 97 Marina del Rey, California, USA December 8–11, 1997 Proceedings*; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 1997; pp 233–240.

(29) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10037–10041.

(30) Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pKa values. *Proteins: Structure, Function, and Bioinformatics* **2005**, *61*, 704–721.

(31) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulation. *Nucleic Acids Res.* **2012**, *40*, E537–541.

(32) Kirkwood, J. G.; Shumaker, J. B. Forces Between Protein Molecules in Solution Arising from Fluctuations in Proton Charge and Configuration. *Proc. Natl. Acad. Sci. USA* **1952**, *38*, 863–871.

(33) da Silva, F. L. B.; Lund, M.; Jönsson, B.; Åkesson, T. On the Complexation of Proteins and Polyelectrolytes. *J. Phys. Chem. B* **2006**, *110*, 4459–4464.

(34) da Silva, F. L. B.; Boström, M.; Persson, C. Effect of Charge Regulation and IonDipole Interactions on the Selectivity of ProteinNanoparticle Binding. *Langmuir* **2014**, *30*, 4078–4083.

(35) Wallace, J. A.; Shen, J. K. Unraveling A Trap-and-Trigger Mechanism in the pH-Sensitive Self-Assembly of Spider Silk Proteins. *J. Phys. Chem. Lett.* **2012**, *3*, 658–662.

(36) Dashti, D. S.; Roitberg, Y. M. A. E. pH-Replica Exchange Molecular Dynamics in Proteins Using a Discrete Protonation Method. *J. Phys. Chem. B* **2012**, *116*, 8805–8811.

(37) Chen, W.; Wallace, J. A.; Yue, Z.; Shen, J. K. Introducing Titratable Water to All-Atom Molecular Dynamics at Constant pH. *Biophys. J.* **2013**, *105*, L15–L17.

(38) Mahadevan, T. S.; Garofalini, S. H. Dissociative Water Potential for Molecular Dynamics Simulations. *J. Phys. Chem. B* **2008**, *111*, 8919–8927.

(39) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. Benchmarking pKa prediction. *BMC Biochemistry* **2006**, *7*, 1–12.

(40) Teixeira, A. A.; Lund, M.; da Silva, F. L. B. Fast Proton Titration Scheme for Multiscale Modeling of Protein Solutions. *Journal of Chemical Theory and Computation* **2010**, *6*, 3259–3266.

(41) Tanford, C.; Kirkwood, J. G. Theory of Protein Titration Curves I. General Equations for Impenetrable spheres. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

(42) da Silva, F. L. B.; Jönsson, B.; Penfold, R. A critical investigation of the Tanford-Kirkwood scheme by means of Monte Carlo simulations. *Prot. Sci.* **2001**, *10*, 1415–1425.

(43) Persson, B.; Lund, M.; Forsman, J.; Chatterton, D. E. W.; Åkesson, T. Molecular evidence of stereo-specific lactoferrin dimers in solution. *Biophys Chem.* **2010**, *3*, 187–189.

(44) Kurut, A.; Dicko, C.; Lund, M. Dimerization of Terminal Domains in Spiders Silk Proteins Is Controlled by Electrostatic Anisotropy and Modulated by Hydrophobic Patches. *ACS Biomater. Sci. Eng.* **2015**, *1*, 363–371.

(45) da Silva, F. L. B.; Derreumaux, P.; Pasquali, S. Fast coarse-grained model for RNA titration. *J. Chem. Phys.* **2017**, *146*, 035101+.

(46) Stanton, C. L.; Houk, K. N. Benchmarking pKa Prediction Methods for Residues in Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 951–966.

(47) Warwicker, J.; Watson, H. C. Calculation of the electric potential in the active site cleft due to α-helix dipoles. *J. Mol. Biol.* **1982**, *157*, 671–679.

(48) Harvey, S. C. Treatment of Electrostatic Effects in Marcomolecular Modeling. *Proteins: Struc., Func. and Genetics* **1989**, *5*, 78–92.

(49) Havranek, J. J.; Harbury, P. B. Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11145–11150.

(50) de Carvalho, S. J.; Ghiotto, R. T.; da Silva, F. L. B. Monte Carlo and Modified Tanford-Kirkwood Results for Macromolecular Electrostatics Calculations. *J. Phys. Chem. B* **2006**, *110*, 8832–8839.

(51) Hill, T. L. *Statistical Mechanics*; McGraw-Hill: New York, 1956.

(52) Beresford-Smith, B.; Chan, D. Y. C. Electrical double-layer interactions in concentrated colloidal systems. *Faraday Disc. Chem. Soc.* **1983**, *76*, 65–75.

(53) Lin, S.-C.; Lee, W. I.; Shurr, J. M. Brownian Motion of Highly Charged Poly(L-lysine). Effects of salt and polyion concentration. *Biopolymers* **1978**, *17*, 1041–1064.

(54) Schmitz, K. S., Ed. *Macro-ion Characterization: From Dilute Solutions to Complex Fluids*; American Chemistry Society: Washington, 1994.

(55) Kjellander, R.; Ulander, J. Effective ionic charges, permittivity and screening length: dressed ion theory applied to 1:2 electrolyte solutions. *Mol. Phys.* **1996**, *95*, 495–505.

(56) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of *Rhodobacter sphaeroides*. *Proc. Natl. Acad. Sci. USA,* **1991**, *88*, 5804–5808.

(57) Baptista, A. M.; Soares, C. M. Some theoretical and computational aspects of the inclusion of proton isomerism in the protonation equilibrium of proteins. *J. Phys. Chem. B* **2001**, *105*, 293–309.

(58) Wang, L.; Li, L.; Alexov, E. pKa predictions for proteins, RNAs, and DNAs with the Gaussian dielectric function using DelPhi pKa. *Proteins* **2015**, *83*, 2186–2197.

(59) Lund, M.; Jönsson, B. On the Charge Regulation of Proteins. *Biochemistry* **2005**, *44*, 5722–5727.

(60) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28*, 235–242.

(61) Metropolis, N. A.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1097.

(62) Frenkel, D.; Smit, B. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: San Diego, 1996.

(63) Stenqvist, B.; Thuresson, A.; Kurut, A.; Vácha, R.; Lund, M. Faunus – A flexible framework for Monte Carlo simulation. *Mol. Sim.* **2013**, *39*, 1233–1239.

(64) Olsson, M. H.; Sondergard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa predictions. *J. Chem. Theory and Computation* **2011**, *7*, 525–537.

(65) Schutz, C. N.; Warshel, A. What Are the Dielectric "constants" of Proteins and How To Validate Electrostatic Models? *Proteins: Struc., Func., and Genetics* **2001**, *44*, 400–417.

(66) Carstensen, T.; Farrell, D.; Huang, Y.; Baker, N. A.; Nielsen, J. E. On the development of protein pka calculation algorithms. *Proteins* **2011**, *79*, 3287–3298.

(67) Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A.; Huang, Y.; Milletti, F.; Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. PROGRESS IN THE PREDICTION OF pKa VALUES IN PROTEINS. *Proteins* **2011**, *79*, 3260–3275.

(68) Wallace, J. A.; Shen, J. K. Predicting pKa Values with Continuous Constant pH Molecular Dynamics. *Methods in Enzymology* **2009**, *466*, 455–475.

(69) Borkovec, M.; Jönsson, B.; Koper, G. In *Surface and Colloid Science*; Matijević, E., Ed.; Surface and Colloid Science; Springer US, 2001; Vol. 16; pp 99–339.

(70) Fast, J.; Mossberg, A.-K.; Svanborg, C.; Linse, S. Stability of HAMLET – A kinetically trapped α-lactalbumin oleic acid complex. *Prot. Sci.* **2005**, *14*, 329–340.

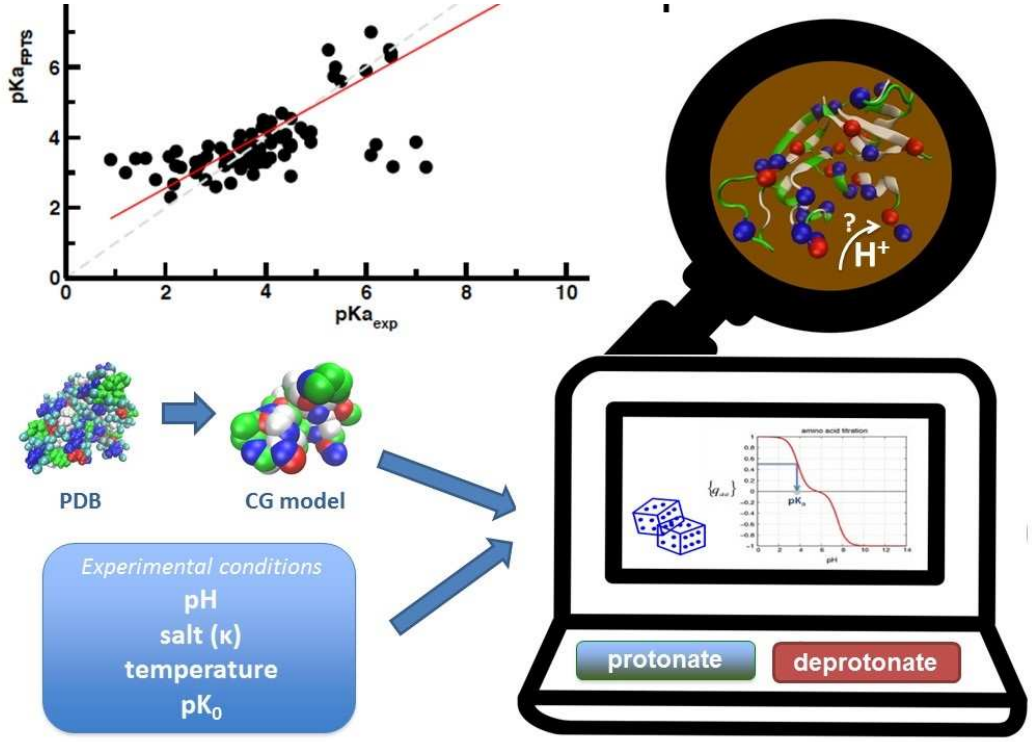
(71) Kuwajima, K. The molten globule state of alpha-lactalbumin. *FASEB J.* **1996**, *10*, 102–109.

(72) Svensson, M.; Hakansson, A.; Mossberg, A.-K.; Linse, S.; Svanborg, C. Conversion of alpha-lactalbumin to a protein inducing apoptosis. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 4221–4226.

(73) Ryan, K. N.; Foegeding, E. A. Formation of soluble whey protein aggregates and their stability in beverages. *Food Hydrocolloids* **2015**, *43*, 265–274.

(74) Schaller, W.; Robertson, A. D. pH, Ionic Strength, and Temperature Dependences of Ionization Equilibria for the Carboxyl Groups in Turkey ovomucoid Third Domain. *Biochemistry* **1995**, *34*, 4714–4723.

(75) Bartik, K.; Redfield, C.; Dobson, C. M. Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional 1H NMR. *Biophys. J.* **1994**, *66*, –145.


pKa_FPTS
pKa_exp
PDB
CG model
Experimental conditions
pH
salt (κ)
temperature
pK0
protonate
deprotonate

TOC