A novel approach to modeling pH-sensitive regions within proteins

Ryan M Harrison & Jeffrey J Gray

Chemical & Biomolecular Engineering, Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218, USA

11 November 2004

Abstract

Due to the limited speed of experimental protein structure determination, genomic scale proteomic research requires the use of computational protein structure prediction algorithms. Though still less accurate than their experimental counterparts, both the speed and accuracy of computational predictions is increasing. A substantial literature review yielded no protein prediction algorithms capable of modeling pH-sensitive regions within proteins. However, modeling pH-sensitive regions within proteins is a necessary step towards the further development of computational protein prediction algorithms. I have developed and implemented pH-sensitive side-chain modeling in Rosetta, a leading protein structure prediction algorithm. Rosetta has achieved significant results in protein docking, folding and design predictions using its unique combination of physical and statistical scoring. The third domain of Turkey Ovomucoid Inhibitor (Omtky3), Ribonuclease A (RNase A) and a simple α-helical nano-gel complex were modeled at various pH values using Rosetta’s new pH-implementation. In addition to correctly modeling pH-dependent energetic and conformational changes in the α-helical nano-gel complex, Rosetta accurately predicted pK_a shifts and conformational realignments in both Omtky3 and RNaseA. The root-mean squared deviation (rmsd) of Rosetta’s pK_a simulations were 0.77 and 0.83 respectively. These results indicate that Rosetta can now capture both energetic and conformational changes induced by pH. Proteins previously outside the scope of Rosetta, particularly membrane fusion peptides and enzymes, are now included in Rosetta’s modeling arsenal. This opens a new realm of possibilities including docking pH-sensitive enzymes, protein folding under variable pH conditions, redesigning proteins to exhibit pH-sensitive functions, simulating the pH of optimum protein stability and modeling pH-denaturization. pH-sensitive modeling will be an integral part of biomolecular engineering, improving drug delivery systems, nanotechnology, pharmaceuticals, and even the creation of novel protein functions and motifs.
Introduction

Modeling pH-sensitive regions within proteins is a necessary step in the further development of computational protein prediction algorithms.\textsuperscript{1,2,3,4} Nearly a million proteins are encoded in the 25,000 genes recently sequenced by the Human Genome Project.\textsuperscript{5,6} However, the functions, interactions and structures of over 95\% of these proteins are still unknown.\textsuperscript{5,6} Many are suspected to have pH-sensitive regions which are fundamental to their structure and function.\textsuperscript{1,2,3,4,7,8} The Protein Databank, the world’s largest freely assessable repository of molecular structures, currently contains approximately 27,000 experimentally determined protein or protein-related molecular structures: most of which are from non-human organisms (\url{http://www.rcsb.org/pdb/}). Proteomics, the study of all proteins expressed by the genome of a cell, is the next logical extension of genomics.\textsuperscript{5,6} Proteomic research has a myriad of applications ranging from biomolecular engineering to drug delivery systems, nanotechnology, pharmaceuticals and even the creation of new proteins.\textsuperscript{5,9,10,11,12,13,14,15} Before protein-based nanotechnology or cures for devastating genetically linked diseases (such as obesity, type-2 diabetes, cancer and cardiovascular disease) become a reality, a deeper understanding of many biological and environmental parameters must be obtained. As a result, both the importance and demand for protein structure and function determination is expected to increase well into the foreseeable future.\textsuperscript{16} Due to their relatively fast speed and low cost, computational methods have an advantage in determining protein structure and function when compared to experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR), and neutron diffraction.\textsuperscript{15,17,18,19,20}

Rosetta:

One of the most successful and accurate computational techniques is the Rosetta protein prediction algorithm.\textsuperscript{15,17,18,19,20,21} The independent labs of the Rosetta community are at the forefront of computational protein design, docking and folding prediction.\textsuperscript{15,17,18,19,20} Rosetta’s modules (RosettaDesign, RosettaDock and RosettaFold) have yielded excellent results in CAPRI (Critical
Assessment of PRediction of Interactions), CASP (Critical Assessment of techniques for Protein Structure Prediction) and other protein design simulations.\textsuperscript{15,17,18,19,20} In a recent 54 target protein-docking benchmark, Rosetta achieved an exceptional accuracy of 10Å or better root-mean squared deviation (rmsd) in 50% of all predictions.\textsuperscript{18} Other successes include the redesigning of an \textit{E.coli} maltose receptor to bind TNT, predictions of 6 out of 8 CAPRI targets with less than 3.2Å interface rmsd and the release of Robetta, a fully automated web based protein tertiary structure prediction client powered by Rosetta (http://lanl.bakerlab.org/).\textsuperscript{14}

\textbf{Applications:}

Despite unprecedented advances in Rosetta and other computational techniques, computational methods are still far from replacing their slower but, substantially more accurate experimental counterparts. At present, the two central roles of computational techniques has been simulating atomic-level details and elucidating structural and functional detail in previously unknown complexes.\textsuperscript{14,17,18,19,22} Rosetta’s applications are far reaching. Current topics of research in the Rosetta community include: investigating and modeling of allosteric proteins, crucial in understanding and eventually correcting the defective signaling pathways of cancerous cells; predicting the structure of therapeutic antibodies, useful in designing and implementing new and more effective drugs; modeling of protein-surface interactions, an important step towards creating self-assembling bio-electronic devices and tissue specific medical implants; investigating the energetics of protein complexes, fundamental to a deeper understanding of protein function from an energetic standpoint; and genomic scale protein modeling, indispensable to the rapid structure determination of entire proteomes. My project, investigating and modeling pH-sensitive regions within proteins, not only improve the overall accuracy of protein docking, design and folding predictions; but will also greatly extend Rosetta’s modeling capabilities. Proteins previously outside the scope of Rosetta, particularly membrane fusion peptides and enzymes, are now included in Rosetta’s modeling arsenal. In short, I have opened a new realm of possibilities for Rosetta and the larger field of Computational Proteomics.
A novel approach to modeling pH-sensitive regions within proteins

**pH-sensitive regions within proteins:**
Principal determinants of the atomic-level environment of proteins include hydrogen-bonding, van der waals, electrostatic and solvation interactions. In Rosetta, a variety of scoring terms, functions and weights are used to evaluate these and other protein energetic properties. When compounded, the thousands of discrete energetic contributions within a protein help to determine its structure and function. pH-sensitive regions exist in many proteins. These regions are central in determining the local/global energetics, structure and function of proteins. There have been many studies on the energetic, structural and functional effects of pH upon proteins; however, none have attempted incorporation of their techniques into a viable protein prediction algorithm. Now that some of the problems related to computational modeling of proteins have been adequately accounted for, expanding the range and accuracy of computational protein prediction techniques will depend upon the modeling of pH-sensitive regions within proteins. A substantial literature review yielded no protein prediction algorithms capable of modeling pH-sensitive regions within proteins. In many proteins, ignoring pH-sensitive regions leads to an inaccurate depiction of atomic-level structure. This has strong negative consequences for both the accuracy and reliability of protein predictions.

**Objectives:**
To extend the range, accuracy and long-term viability of computational protein prediction algorithms, I have developed and implemented pH-sensitive modeling in Rosetta. Before the introduction of my pH-sensitive side-chain packing and minimization routines, Rosetta could not accommodate the variability of protein energetics or structure with respect to pH (Rosetta assumes pH 7.0). The implementation of Rosetta’s pH-sensitivity was sub-divided into two major phases: (1) a preliminary un-dynamic ionization model (single-ionization state per amino acid type) designed to gauge the feasibility of implementation and (2) a more rigorous full-scale multi-ionization state model. The objective of the preliminary model, which only incorporated the most essential elements of pH-sensitivity, was to detect the electrostatic pH-
sensitivity of a simple $\alpha$-helical nano-gel. The overall objective of the multi-ionization state model is to expand the range and accuracy of protein docking, design and folding in Rosetta. Before this aim can be achieved, Rosetta must first be able to simulate $pK_a$ shifts within proteins. $pK_a$ shifts are changes from the standard or “IpK$_a$” caused by local interactions. Thus, simulations for experimentally determined $pK_a$s are being conducted to gauge Rosetta’s accuracy in predicting both $pK_a$ derived energetic and structural effects. $pK_a$ values were predicted for the third domain of turkey ovomucoid inhibitor (OMTKY3) and ribonuclease A (RnaseA); these two well characterized proteins contain a total of 19 experimentally determined $pK_a$ values.

**Methods**

Sets of pH-sensitive potential side-chain conformations or “rotamers” were constructed to simulate the presence of alternate ionization states. Each alternate ionization state requires parameters for solvation, electron density, reference energy, explicit hydrogen and hydrogen-bonding corresponding to its appropriate ionization state. The energetically most favorable ionization states were determined on the basis of their probability of occurrence at a given pH, self-energy, pair-wise energies and global energetic effects.$^{19,33}$ The preliminary model, which forced all amino acids of the same type to occupy the same ionization state and incorporated only the most rudimentary electrostatic score function, was used to estimate the electrostatic pH-sensitivity of a simple $\alpha$-helical nano-gel. A second rough approximation of the pH-sensitivity of electrostatics, solvation and hydrogen bonding with the multi-state model was conducted on the same $\alpha$-helical model system. A more rigorous test, predicting the $pK_a$ values of ionizable residues, was then conducted on more complex protein structures: the third domain of turkey ovomucoid inhibitor (OMTKY3) and ribonuclease A (RnaseA).$^{1,30}$
A novel approach to modeling pH-sensitive regions within proteins

**pK$_a$ simulation:**

Two trials, using either a coulombic distance-dependent dielectric or modified generalized born electrostatic model, were conducted on both OMTKY3 and RnaseA. Both trials consisted of 140 independent simulations, each representing an interval of 0.1 pH-units in a pH range of 1-14. In each simulation of independently optimized and minimized protein structures both charged and uncharged protonation states for pH-sensitive amino acids were sampled. Each independently optimized and minimized protein structure, named a “decoy”, undergoes a side-chain packing/optimization process, then a rotamer-based minimization. Each pK$_a$ was defined as the pH at which a residue changed ionization state. The accuracy of the pK$_a$ simulations was determined through comparison of simulated pK$_a$s to theoretical approximations, IpK$_a$s and experimental pK$_a$s.

**Optimization and minimization of protein structures:**

In the simulations, a two-stage approach of protein scoring and side-chain conformation determination is used. First, the protein structure was optimized through side-chain packing. Rotamers are used to represent side-chain conformational freedom. Every rotamer of every amino acid was sampled and scored. The lowest scoring rotamers for every amino acid was then minimized in a rotamer trial protocol.$^{18,19}$ In rotamer trials, pair-wise interactions are calculated between all amino acid rotamers within an 8Å shell. In both the optimization and minimization protocol, rotamers for all defined tautomers and protonation states are sampled. Using the dunbrack energy$^{33}$, the ionization state equilibrium, the chemical equilibrium between water, H$_3$O and protein side-chain, was biased toward the standard distribution of a theoretical titration curve. The standard theoretical titration curve is given by:

$$f_{\text{protonated}} = \frac{1}{10^{pH-IpK_f} + 1}$$

where $f_{\text{protonated}}$ is the fraction protonated and pH is the negative logarithm of the H$_3$O concentration. The Dunbrack energy is given by:

$$\Delta G_{\text{Dunbrack}} = -\log(rprob)$$

Harrison 5
where $\Delta G_{Dunbrack}$ is the dunbrack energy and $r_{\text{prob}}$ is the probability of the given rotamer conformation given the current backbone conformation.\textsuperscript{33} Therefore, the rotamer probability of a protonated side-chain, assuming the rotamer probability distribution is equal for both charged and uncharged states, is given by:

$$r_{\text{prob}}^{\text{protonated}} = r_{\text{prob}} \times f^{\text{protonated}}$$

where $r_{\text{prob}}^{\text{protonated}}$ is the rotamer probability of the protonated state. Thus, the rotamer probability of the deprotonated state of a side-chain would be given by:

$$r_{\text{prob}}^{\text{deprotonated}} = r_{\text{prob}} \times (1 - f^{\text{protonated}})$$

where $r_{\text{prob}}^{\text{deprotonated}}$ is the rotamer probability of the uncharged ionization state. Through this method, $\Delta G_{Dunbrack}$ was modified to vary with pH. If other energetic factors do not further influence the amino acid, then distribution of charged and uncharged ionization states will follow their standard theoretical model. However, if other local energetic factors (normally electrostatic or solvation) impact the amino acid, the bias toward a theoretical distribution can be overcome. This will lead to a pK$_a$ shift, a shift in the simulated pK$_a$ away from the standard theoretical model. As a result of this simple and novel approach, pK$_a$ shifts in Rosetta can be simulated.

**Rosetta energy functions:**

Solvation, electron density, reference energy, and hydrogen bonding are the major physical amino acid attributes affected by pH.\textsuperscript{1,5,23,24,25,28,29} Changes in these attributes are reflected through Rosetta’s energy functions.\textsuperscript{17,18,19} Energy functions within Rosetta utilize the parameters defined by amino acid templates to calculate a variety of biophysically relevant energies. Therefore, modification of template energetic parameters will affect the local and global energetic landscape of a protein.
A novel approach to modeling pH-sensitive regions within proteins

Table I. Solvation Parameters

<table>
<thead>
<tr>
<th>Atom types</th>
<th>Volume</th>
<th>ΔGi_{ref}</th>
<th>ΔGi_{free}</th>
<th>ΔHi_{ref}</th>
<th>ΔCpi_{ref}</th>
</tr>
</thead>
<tbody>
<tr>
<td>NH1</td>
<td>4.4</td>
<td>-5.950</td>
<td>-8.90</td>
<td>-9.059</td>
<td>-8.80</td>
</tr>
<tr>
<td>NR</td>
<td>4.4</td>
<td>-3.820</td>
<td>-4.00</td>
<td>-4.654</td>
<td>-8.80</td>
</tr>
<tr>
<td>NH2</td>
<td>11.2</td>
<td>-5.450</td>
<td>-7.80</td>
<td>-9.028</td>
<td>-7.00</td>
</tr>
<tr>
<td>NH3</td>
<td>11.2</td>
<td>-20.000</td>
<td>-20.00</td>
<td>-25.000</td>
<td>-18.00</td>
</tr>
<tr>
<td>OH1</td>
<td>10.8</td>
<td>-5.920</td>
<td>-6.70</td>
<td>-9.264</td>
<td>-11.20</td>
</tr>
<tr>
<td>O</td>
<td>10.8</td>
<td>-5.330</td>
<td>-5.85</td>
<td>-5.787</td>
<td>-8.80</td>
</tr>
<tr>
<td>OC</td>
<td>10.8</td>
<td>-10.000</td>
<td>-10.00</td>
<td>-12.000</td>
<td>-9.40</td>
</tr>
</tbody>
</table>

Table obtained directly from Lazaridis & Karplus, 2000 [24]

||Volumes in Å, free energies and enthalpies in kcal/mol, heat capacities in cal/mol · K. For polar hydrogens, the parameters in this table are zero. All free energy, enthalpy, and heat capacity values refer to 298.15 K.

Atom types: NH1: amide nitrogen, NR: aromatic nitrogen with no hydrogens, NH2: nitrogen bound to two hydrogens, NH3: nitrogen bound to three hydrogens, OH1: hydroxyl oxygen, O: carbonyl oxygen, OC: carboxyl oxygen [24].

ΔGi_{ref} and ΔHi_{ref} obtained from [24] Table 1. The enthalpies for these groups are given a value larger in magnitude than the free energy to account for a negative solvation entropy; (b) proline N, for which no parameters were available; it is given values smaller in magnitude than for the other, more polar nitrogens. For the carbonyl group, all solvation free energy is assigned to the O.

ΔGi_{ref} is determined by increasing the magnitude of ΔGi_{ref} until the solvation free energy of deeply buried groups becomes zero [24].

Rosetta’s solvation models are used to capture the energetic penalty of high solvent exposure of uncharged and aromatic groups and the energetic favorability of high solvent exposure for charged and polar groups.18,19 Both an implicit Gaussian solvent-exclusion model24 and an implicit hydrophobic surface area calculation25 represent solvation and hydrophobic effects in Rosetta. These energy models utilize the solvation parameters from template amino acids (Table I).24 A condensed version of the Gaussian solvent-exclusion model24 is given by:

\[ \Delta G_{i}^{slv} = \Delta G_{i}^{ref} - \sum_{j \neq i} f(r_{ij})V_{j} \]

where \( \Delta G_{i}^{slv} \) is the solvation energy in kcal mol\(^{-1}\), \( \Delta G_{i}^{ref} \) is the amino acid reference energy, \( f(r_{ij}) \) is the

Harrison 7
solvent free energy density and $V_j$ is the volume of atom $j$.

![Figure I: Simplified partial charge diagrams for Glutamate and Lysine. Parameters for the charged states of each amino acid are derived from CHARM 19. Parameters for uncharged states were approximated using principles from organic chemistry and trial/error.](image)

The electron density parameters are used for electrostatic calculations (Figure I). Either a coulombic distance-dependent dielectric model or a modified version of the generalized born electrostatic model was used for electrostatic calculations. The equation for the coulombic distance-dependent dielectric model is given by:

$$\Delta G_{elec} = \sum_{j=1}^{n} \sum_{i=1}^{n} \frac{332qiqj}{\epsilon r}$$

where $\Delta G_{elec}$ is the electrostatic energy in kcal mol$^{-1}$, $\epsilon$ is the distance dependent dielectric, $n$ is the number of atoms and $q$ is the electron density of atom $i$ or $j$.

Reference energies, the intrinsic energy of an amino acid, help stabilize the equilibrium between charged and uncharged ionization states. Uncharged amino acid ionization states tend to have low or negative reference energies, while charged amino acid ionization states tend to have higher reference energies. This is due to the intrinsic energetic bias against charge concentrations within proteins; however, this can be compensated for by favorable electrostatic interactions or high solvent accessibility. The reference energies used in Rosetta were taken from several experimental $\Delta \Delta G$ experiments, and then parameterized to prevent over-expression of any amino acid within RosettaDesign mode.

Hydrogen bonding in Rosetta is based on an orientation-dependent potential derived from high-resolution
A novel approach to modeling pH-sensitive regions within proteins.

PDB structures.\textsuperscript{22} A condensed version of the hydrogen bonding score function is given by:

\[ \Delta G_{\text{hbond}} = - \sum_{i=1}^{n} kT \ln(hbprob_i) \]

where \( \Delta G_{\text{hbond}} \) is the hydrogen bonding energy in kcal mol\(^{-1}\), \( hbprob \) is the hydrogen orientation dependent hydrogen bond probability and \( n \) is the number of hydrogen bonds in the protein. Van der Waals interactions, which are not substantially pH-sensitive, dominate the energetic landscape within Rosetta.\textsuperscript{18,19} They are calculated using a modified Lennard-Jones potential.\textsuperscript{18,19}

**Amino Acid Templates:**

Originally, Rosetta contained one amino acid template for each of the twenty common l-form amino acids; later, an additional template was added for histidine’s tautomer. The first step in incorporating pH-sensitive side-chain conformational variability into Rosetta was the construction of templates for alternate ionization states of the amino acids: aspartic acid, glutamic acid, histidine, lysine, arginine and tyrosine. (Although cystine is also an ionizable residue and its disulfide forming properties vary with pH, implementation would have severely conflicted with technical aspects of disulfide formation in Rosetta. It is being considered for later implementation.) Template coordinates for alternate ionization state rotamers were derived from PyMOL, which is based on the CHARMM19 coordinate set. Energetic parameters, such as solvation, electron density, reference energy, and hydrogen bonding, were then added to each of the alternate ionization state templates. Solvation and hydrogen bonding parameters were taken directly from the literature.\textsuperscript{24,22} Both electron density and reference energy parameters were estimated from pre-existing parameters within Rosetta.\textsuperscript{23,27}
Results & Discussion

Single Ionization State (Preliminary) pH-implementation:

![Graph showing pH vs Electrostatic Energy](image)

**Figure II.** $\Delta G_{elec}$ as a function of solvent pH. Positive-negative electrostatic interactions at pH 7.0 cause $\alpha$-helix stabilization resulting in a gel state.

Previous experimental data for the self-assembling bioactive nano-gel developed in the Harden Lab\textsuperscript{11,13} indicates a state transition from gel, at pH 7.0, to solution, at pH 3.5 and 11.0.\textsuperscript{11} This gel-to-solution transition can be attributed to weakened electrostatic interactions between the acidic and basic $\alpha$-helical caps at extremes of pH.\textsuperscript{11} At a pH of 7.0, glutamic acid and lysine exhibit strong electrostatic attraction, resulting in a low $\Delta G_{elec}$, electrostatic energy, as given by Coulomb’s Law. When the ionization state of glutamic acid or lysine changes from its standard physiological state; neutral-positive or negative-neutral electrostatic interactions ensue across the interface. These interactions are far weaker than the negative-positive charge interactions observed at pH 7.0; thus, the $\Delta G_{elec}$ at pH 7.0 should be lower than that of pH 3.5 and pH 11.0. My computational results matched experimental trends, indicating the successful
prediction of a rough correlation between pH and electrostatic energy (Figure II). On average, $\Delta G_{\text{elec}}$ at pH 7.0 was 27% (~28.69 kcal/mole) lower than that of pH 3.5 and 74% (~78.65 kcal/mole) lower than that of pH 11.0. The 49.96 kcal/mole discrepancy between the electrostatic energies at pH 3.5 and pH 11.0 was attributed the greater solvent accessibility of lysine. For many lysine residues, both solvent and surrounding glutamate residues stabilize lysine, while glutamate residues are normally only stabilized by one or the other. (Figure III). At pH 7.0, less than half of the lysine residues had exclusive interactions with solvent; however, at pH 3.5 nearly all the lysine residues interacted solely with solvent. A similar, but far less dramatic phenomenon occurs at pH 11.0 when charged lysine is absent. This is most certainly due to the absence of charged glutamate residues at such an acidic pH. Given the success of Rosetta’s single-state implementation, implementation of a more complex and encompassing pH-sensitive packing model was feasible. Thus, the next logical step in Rosetta’s pH-implementation was to develop and test multiple-ionization state models.

**Multi-Ionization State pH-implementation:**

**$\alpha$-helical Nano-gel:**

**Figure III:** Acidic-Basic $\alpha$-helical model system
A. Leucine zipper of acidic (red) and basic (blue) Alpha-Helical Fragments at pH 7.0.
B. Negative Glutamates (red) and Positive Lysines (blue) of Alpha-Helical Fragments at pH 7.0.

The first step in approximating the accuracy of the multi-ionization state pH-implementation was to obtain a coarse approximation of its functionality. The same $\alpha$-helical nano-gel (Figure III) used in the
coarse approximation of the single-ionization state was used; resulting in a similar, but more accurate, depiction of the gel to solution transition indicated in the single-state simulation.\textsuperscript{11} Multi-ionization state functionality allowed for the changes induced by pH to be observed over a greater and more dynamic range of pH values (Figure IV). Additionally, hydrogen bonding, solvation and electrostatic parameters were updated between the single and multi-state versions of the pH-implementation. From Figure IV, it is clear that electrostatics was the most pH-sensitive of Rosetta’s energy functions. The minimum electrostatic energy was expected at pH 7.0; however, there is actually an increased electrostatic energy at pH 7.0 versus the electrostatic minimum of -107 kcal/mol at pH 8.0. This was attributed to two lysine residues shifting into a less solvent-exposed side-chain conformation at pH 7.0. In addition to decreasing solvation energy, this conformational change disrupted attractive charge interaction with neighboring glutamate residues. Apparently, this action relieved some unfavorable charge interaction with a nearby leucine.

**pH-Sensitive Protein Energetics**

![Figure IV: Free energy of total score, solvation, cumbic electrostatic and short ranged hydrogen bonding as a function of pH. The dramatic decrease in electrostatic energy in the range of pH 5.0 – pH 10.0 is caused by the favorable interaction of interface GLU and LYS residues. The slight change in solvation energy is due to the high solvent accessibility of charged polar groups. The very small hydrogen-bonding energy decrease is due to a weak hydrogen bond caused by the reorientation of a protonated LYS.](image)
The solvation energy varied with pH because of fluxing degrees of solvent accessibility. There seems to be a fine balance between maximum solvent exposure and interaction with oppositely charge residues. At the most optimal conformations, around pH 8.0, many residues are able to obtain stabilization from both these factors. Below pH 5.0 and above pH 10.0, solvation energy is lowered by the lesser solvent affinity of the uncharged forms of lysine and glutamate. Contrary to expectation, side-chain hydrogen bonding did not play a significant role. Only one lysine residue was involved as in a weak hydrogen bond, as a donor to a nearby carboxyl group. This was probably due to the greater favorability of bulk solvent interaction. Unfortunately, only qualitative results could be obtained for this system due to the lack of experimentally determined pKₐ values. More detailed characterization of this nano-gel system, especially its ability to house bioactive sites, is currently underway. (Note: The predicted radius of gyration for this model system aided in further experimental characterization of the bioactive sites of this model system.)

**Third Domain of Turkey Ovomucoid Inhibitor (OMTKY3):**

*Figure V:* Rosetta prediction of X-ray crystal structure of a complex of human leukocyte elastase (PMN elastase) and the third domain of turkey ovomucoid inhibitor at pH 7.0 (PDB: 1PPF [30]). **Green:** Third domain of turkey ovomucoid inhibitor; **Red:** De-protonated Glutamate residues; **Orange:** De-protonated Aspartate residues

Because it has been well characterized by several experimental and computational approaches, the turkey ovomucoid enzyme-inhibitor complex (Figure V) was an ideal sample system. Of its 16
ionizable residues, 15 have experimentally measured pK\(_a\) shifts.\(^{28}\) Both GLU19 and ASP27 demonstrate minor conformational changes between their two ionization states. ASP27 seems to have realigned to take advantage of a very favorable electrostatic interaction with HIS52. ASP7, GLU19 and ASP27 demonstrate substantially downshifted pK\(_a\) values, 2.7, 3.2 and 2.3 respectively (Table II). While Rosetta did correctly simulate the pK\(_a\) downshift, the magnitude of the downshift was underestimated. Only ASP27 had a predicted pK\(_a\) greater than 1 pH unit away from the experimental value. I attribute this to an inaccurate assessment of the strength of the interaction between ASP27 and HIS52. Since the weight on the Generalized-Born electrostatic score term, \(\Delta G_{\text{gbelc}}\), was very low, 0.041, other energetically favorable electrostatic contributions were probably underestimated. This suggests that the score function weights require slight calibration and/or the partial charge approximations for acidic residues are insufficiently accurate. Since Rosetta’s standard score weights were unaltered in this simulation, improper weighting of the \(\Delta G_{\text{gbelc}}\) score term probably accounted for most of the error. For both GLU10 and GLU43, Rosetta incorrectly predicted a downshift in pK\(_a\). However, for GLU10 the magnitude of the downshift was comparable to that of other models; GLU43 missed its experimental pK\(_a\) by 0.8 pH units. This would suggest that improperly weighted the \(\Delta G_{\text{gbelc}}\) for Rosetta might also underestimate repulsive electrostatics. Due to lysine’s less dramatic pKa shifts and greater degree of solvent accessibility, lysine pK\(_a\)’s were predicted more accurately than glutamic and aspartic acid pK\(_a\)’s.

Several other groups have calculated pK\(_a\) shifts in the third domain of turkey ovomucoid inhibitor. Amongst the most accurate were a modified Tanford-Kirkwood electrostatic potential\(^{32}\) and a FDPB electrostatic potential\(^1\) with root mean squared (rms) errors of 0.61 and 0.59 pH units respectfully. Both of these approaches lacked stringent solvation models and ignored hydrogen bonding. Rosetta’s overall rmsd was 0.15 pH units less than the null value; 0.78 pH units with coulombic electrostatics and 0.77 pH units with the generalized born electrostatics. This suggests that in this complex the accuracy of the two electrostatic models was approximately equal. If this were the case for the majority of protein complexes, it would be more efficient to use the less computational expensive coulombic electrostatics.
A novel approach to modeling pH-sensitive regions within proteins

Table II: pKₐ values for OMTKY3

<table>
<thead>
<tr>
<th>Group</th>
<th>Rosetta*</th>
<th>Coulomb</th>
<th>gen. born</th>
<th>IpKₐ</th>
<th>Experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASP 7</td>
<td>3.4</td>
<td>3.4</td>
<td>4.0</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>GLU 10</td>
<td>3.8</td>
<td>3.9</td>
<td>4.4</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>LYS 13</td>
<td>10.6</td>
<td>10.6</td>
<td>10.4</td>
<td>9.9</td>
<td></td>
</tr>
<tr>
<td>GLU 19</td>
<td>4.2</td>
<td>4.2</td>
<td>4.4</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>ASP 27</td>
<td>3.7</td>
<td>3.7</td>
<td>4.0</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>LYS 29</td>
<td>10.7</td>
<td>10.8</td>
<td>10.4</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>LYS 34</td>
<td>10.6</td>
<td>10.6</td>
<td>10.4</td>
<td>10.1</td>
<td></td>
</tr>
<tr>
<td>GLU 43</td>
<td>4.0</td>
<td>4.0</td>
<td>4.4</td>
<td>4.8</td>
<td></td>
</tr>
<tr>
<td>LYS 55</td>
<td>10.5</td>
<td>10.5</td>
<td>10.4</td>
<td>11.1</td>
<td></td>
</tr>
</tbody>
</table>

RMSD 0.78 Max abs 1.4

|$\text{Experimental measured } pK_\alpha \text{ values [28]}

$\text{Intrinsic model compound } pK_\alpha \text{ values [1]}

*Simulated pKₐ values determined by using backbone coordinates from the X-ray crystal structure of the third domain of turkey ovomucoid inhibitor (PDB: 1PPF [30]). Rotamer Conformations for all side-chains were relaxed to their free-energy minima though full atom optimized and minimized. See methods: Optimization and minimization of protein structures.

$\text{Coulombic distance dependent dielectric electrostatic model [23]}

§Modified generalized born electrostatic model [27]

Ribonuclease A (RnaseA):

Figure VI: Columbic Rosetta prediction of X-ray crystal structure of Ribonuclease A at pH 5.2 (a) and pH 7.1 (b) (PDB: 1KF2-8 [28]). Gray: Experimental structure; Magenta: Rosetta prediction; Red: Experimental position of tritratable amino acids; Green: Rosetta prediction position of tritratable amino acids.
A novel approach to modeling pH-sensitive regions within proteins

Ribonuclease A (Figure VI), a well characterized RNA degradation enzyme, was selected as a sample system because of its relatively high degree of difficulty and the availability of high-resolution crystal structures at multiple pHs. The high connectivity and pKₐ coupling of this complex is difficult to capture with most computational techniques. Due to undetected coupling, the two predicted pKₐ values for GLU2 and ASP14 are inaccurate to more than 1 pH unit (Table III). Despite this, the overall pKₐ rmsd was 0.87 with coulombic electrostatics and 0.83 with generalized born electrostatics; more accurate than the null value rmsd. The discrepancy between the coulombic and generalized born electrostatic models suggests that the more computationally expensive generalized born model is more accurate in this sample system. All fullatom rmsds are less than the 1.20 Å resolution of the crystal structures (Table IV). This suggests that Rosetta’s sidechain packing minimizations are sufficiently accurate to capture the local energetic effects that dramatically shift pKₐs.

Table III: pKₐ values for RNaseA

<table>
<thead>
<tr>
<th>Group</th>
<th>Rosetta*</th>
<th>coulomb¶</th>
<th>gen. born§</th>
<th>IpKₐ d</th>
<th>Experimental₪</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLU 2</td>
<td>4.0</td>
<td>3.9</td>
<td>4.4</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>GLU 9</td>
<td>4.0</td>
<td>3.9</td>
<td>4.4</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>ASP 14</td>
<td>4.1</td>
<td>3.9</td>
<td>4.0</td>
<td>&lt;2.0</td>
<td></td>
</tr>
<tr>
<td>ASP 38</td>
<td>3.6</td>
<td>3.6</td>
<td>4.0</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>GLU 49</td>
<td>4.1</td>
<td>3.9</td>
<td>4.4</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>ASP 53</td>
<td>3.7</td>
<td>3.7</td>
<td>4.0</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>ASP 83</td>
<td>3.9</td>
<td>3.9</td>
<td>4.0</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>GLU 86</td>
<td>4.1</td>
<td>4.1</td>
<td>4.4</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>GLU 111</td>
<td>4.1</td>
<td>4.1</td>
<td>4.4</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>ASP 121</td>
<td>3.8</td>
<td>3.9</td>
<td>4.0</td>
<td>3.1</td>
<td></td>
</tr>
</tbody>
</table>

| RMSD  | 0.87     | 0.83     | 0.98       |
| Max abs | 2.10    | 1.90     | 2.00       |

¶ Exponentially measured pKₐ values [28]
§ Intrinsic model compound pKₐ values [1]
*Simulated pKₐ values determined by using backbone coordinates from the X-ray crystal structure of the third domain of turkey ovomucoid inhibitor (PDB: 1KF5 [27]). Rotamer Conformations for all side-chains were relaxed to their free-energy minima though full atom optimized and minimized. See methods: Optimization and minimization of protein structures.
¶Coulombic distance dependent dielectric electrostatic model [23]
§Modified generalized born electrostatic model [27]
Table IV: RMSD values for RnaseA±

<table>
<thead>
<tr>
<th>pH</th>
<th>All Residues</th>
<th>Tritratable Residues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fullatom RMSD</td>
<td>Sidechain RMSD</td>
</tr>
<tr>
<td>pH 5.2</td>
<td>1.08 Å</td>
<td>1.56 Å</td>
</tr>
<tr>
<td>pH 7.1</td>
<td>0.99 Å</td>
<td>1.43 Å</td>
</tr>
</tbody>
</table>

± RMSD (Root Mean Squared Deviation) between experimental and Rosetta predicted structure at given pH. The resolution of the experimental structures is 1.20 Å.

Conclusions

In addition to correctly modeling pH-dependent energetic and conformational changes in the α-helical nano-gel complex, Rosetta accurately predicted both pKₐ shifts and conformational realignments in both OMTKY3 and RNaseA. Results from simulations involving the α-helical Nano-gel were the best that could possibly be expected given the lack of a large body of experimental work. The rmsd error of Rosetta’s pKₐ simulation on OMTKY3 was 0.77, 0.15 pH units more accurate than the IpKₐ. A similar result was obtained for RnaseA with a rmsd error of 0.83, 0.15 pH units more accurate than the IpKₐ. These results indicate that the current implementation for modeling of pH-sensitive regions is more accurate than a traditional Rosetta simulation. To the author’s knowledge, Rosetta is now the first protein structure prediction algorithm featuring modeling of pH-sensitive regions. Rosetta’s pH-sensitive side-chain packing and minimization is a significant accomplishment given the complexity of protein structure prediction, pKₐ calculations/simulations and the novelty of the approach taken. A variety of substantial obstacles were overcome implementing pH-sensitivity into Rosetta: a FORTRAN to C++ conversion of Rosetta, understanding the inner-most workings of the extremely complex Rosetta algorithm, creating complex amino acid parameters and score functions and creating, developing and implementing the idea of pH-sensitive packing and minimization in Rosetta from initiation to completion.
Current Research

I am currently coding a completely revamped version of Rosetta’s pH-implementation. My new protonation potential combines the efficiency of the Dunbrack weighting method with the improved accuracy of slower, more thermodynamically linked methods. I am also re-calibrating Rosetta’s amino acid templates and score function weights to more accurately approximate experimental pK_a values. Recently, I have selected six well-characterized protein complexes with experimentally determined pK_a shifts (Barnase, BPTI, HIV-1 protease, ProtG, CEW lysozyme and Subtilisin inhibitor); these complexes contain over 60 experimentally determined pK_a values. This larger sample set will be a better depiction of the accuracy and effectiveness of pK_a simulations in Rosetta. Possible future directions include reexamining pH-sensitive complexes within the 54 target protein docking benchmark performed by Gray et al., docking pH-sensitive enzymes, modeling pH-denaturization, protein folding under variable pH conditions, redesigning proteins to exhibit pH-sensitive binding and simulating the pH of optimum protein stability.

References


Figures and Tables were created by the author unless otherwise stated.