Physics Based Protein Ionization and pK Estimation in Discovery Studio

June 19, 2008
Start time: 7am PST and 10am PST

Francisco Hernandez-Guzman, Ph.D.
Lead Solutions Scientist
fhernandez@accelrys.com

This presentation and/or any related documents contains statements regarding our plans or expectations for future features, enhancements or functionalities of current or future products (collectively “Enhancements”). Our plans or expectations are subject to change at any time at our discretion. Accordingly, Accelrys is making no representation, undertaking no commitment or legal obligation to create, develop or license any product or Enhancements. The presentation, documents or any related statements are not intended to, nor shall, create any legal obligation upon Accelrys, and shall not be relied upon in purchasing any product. Any such obligation shall only result from a written agreement executed by both parties. In addition, information disclosed in this presentation and related documents, whether oral or written, is confidential or proprietary information of Accelrys. It shall be used only for the purpose of furthering our business relationship, and shall not be disclosed to third parties.
Agenda:

• What is Discovery Studio?

• Protein ionization and pK prediction
  – Why do we care about accurate pK prediction?
    • General methodology
  – Overall workflow for pK prediction
    • System preparation
    • Parameter and topology definition
  – General Validation
  – Application areas

• Demo
  – Setting up a pK prediction job
  – Analysis of the results

• Validation test case OMTKY3

• Q&A
What is Discovery Studio?

• Discovery Studio is a unified Life Science integrated environment to perform molecular modeling and simulations calculations such as:
  – In silico Structure Based Drug Design
  – Rational Drug Design
  – Protein Modeling
  – Sequence Analysis
  – Molecular Dynamics Simulations
  – X-ray crystallography refinement

• Discovery Studio is tightly integrated to the Pipeline Pilot platform technology for superior technology development and customization
Why do we care about accurate pK prediction?

General Background:

- Titratable residues: exist in protonated and deprotonated forms
- Protonation state can be predicted from a titration curve (pK)
- Electrostatic properties of proteins change as a function of pH

Scientific implications in:

- Electrostatic component of protein-ligand, protein-protein binding energy
- pH dependent folding stability
- Unusual titration curves → functional relevant residue
- Stable MD simulations
- Site directed mutagenesis studies
- pI estimation for crystallography
Protein ionization and pK prediction method

• Exp. pK determination of penta-peptides ⇒ model compounds

  N-terminal –Ala- Ala – X – Ala – Ala - C-terminal

  \[ X = \text{Asp, Glu, Arg, Lys, His, Tyr, or Lys} \]

• N titratable sites ⇒ \(2^N\) titration states (E.g. 3 sites ⇒ 8 titration states)

• Consider interaction of the titratable sites ⇒ GBORN energy cutoff
  – Interaction of remaining titratable sites ⇒ mean-field potential\(^1\)

• Iterate through all N sites & calculate partial protonation until converge\(^1\)

Protein ionization and pK prediction method (cnt’d)

• New protocol to “Calculate Protein Ionization and pK” a.k.a. GBpK

  – Based on **CHARMm Generalized-Born** methods
  
  – Creates titration curves and pK estimates for residues using 3D environment of protein structure
  
  – **Automatically** sets the protonation state of each residue at given pH, based on pK
  
  – Calculates pH dependent electrostatic energy
  
  – Calculates pH dependent folding energy

Reference including validation: Spassov. et al, 2008, manuscript submitted (available upon request)
pK prediction of selected Proteins

- Comparison of experimental pK\(_{1/2}\) with calculated values for select PDB files
- All computations about 1 minute per system on a single CPU

<table>
<thead>
<tr>
<th>PDB code</th>
<th>System</th>
<th>Resolution A</th>
<th>Sites</th>
<th>GBpK Polar H</th>
<th>GBpK All H</th>
<th>PROPKA*</th>
<th>MCCE** ((\varepsilon = 8))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 4pti</td>
<td>Trypsin</td>
<td>1.50</td>
<td>14</td>
<td>0.36</td>
<td>0.36</td>
<td>0.6</td>
<td>0.47</td>
</tr>
<tr>
<td>2 2lzt</td>
<td>Lysozyme</td>
<td>1.97</td>
<td>21</td>
<td>0.45</td>
<td>0.57</td>
<td>0.66</td>
<td>0.74</td>
</tr>
<tr>
<td>3 2m2</td>
<td>Ribonuclease H (E.Coli)</td>
<td>1.48</td>
<td>25</td>
<td>0.59</td>
<td>0.68</td>
<td>0.72</td>
<td>0.87</td>
</tr>
<tr>
<td>4 3m3</td>
<td>Ribonuclease A (Bovine)</td>
<td>1.45</td>
<td>16</td>
<td>0.47</td>
<td>0.71</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>5 1pga</td>
<td>Streptococcal Prot. G B1 binding domain -</td>
<td>2.07</td>
<td>15</td>
<td>0.50</td>
<td>0.57</td>
<td>0.72</td>
<td>0.63</td>
</tr>
<tr>
<td>6 3icb</td>
<td>Calcium binding protein (bovine) -</td>
<td>2.30</td>
<td>10</td>
<td>0.33</td>
<td>0.35</td>
<td>0.9</td>
<td>0.38</td>
</tr>
<tr>
<td>7 1hng</td>
<td>Cell adhesion molecule CD2 -</td>
<td>2.80</td>
<td>14</td>
<td>0.55</td>
<td>0.53</td>
<td>0.83</td>
<td>0.76</td>
</tr>
<tr>
<td>8 1a2p</td>
<td>Barnase</td>
<td>1.50</td>
<td>14</td>
<td>0.86</td>
<td>0.94</td>
<td>0.68</td>
<td>0.89</td>
</tr>
<tr>
<td>9 1omu</td>
<td>OMTKY3</td>
<td>50 mods</td>
<td>15</td>
<td>0.64</td>
<td>0.70</td>
<td>0.44</td>
<td>1.10</td>
</tr>
<tr>
<td>10 9rt</td>
<td>Ribonuclease T1 -</td>
<td>1.50</td>
<td>15</td>
<td>0.54</td>
<td>0.65</td>
<td>1.51</td>
<td></td>
</tr>
<tr>
<td>11 3nrt</td>
<td>Ribonuclease T1 -</td>
<td>1.80</td>
<td>4</td>
<td>0.28</td>
<td>0.59</td>
<td></td>
<td>0.54</td>
</tr>
<tr>
<td>12 1bi6</td>
<td>Bromelain Inhibitor IV - heavy chain</td>
<td>1 mod</td>
<td>18</td>
<td>0.54</td>
<td>0.53</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>13 1bi6</td>
<td>Bromelain Inhibitor IV - light chain</td>
<td>1 mod</td>
<td>4</td>
<td>0.18</td>
<td>0.27</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>14 1rgg</td>
<td>Hydrolase Guanyloribonuclease</td>
<td>1.20</td>
<td>24</td>
<td>0.84</td>
<td>0.99</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>

Total sites: 206 206 206 184

Average RMSD: 0.52 0.60 0.74 0.72

RMSD = SQRT{SUM [ (pKexp(i) - pKcalc(i))^2 ] / Nsites}
Comparison of the computed pK_{half} values with the experimental pKa values of the acidic and basic groups of hen-egg white lysozyme (PDB ID 2lzt). Experimental data from Kuramitsu and Hamaguchi, 1980 and Bratik et al., 1994.

<table>
<thead>
<tr>
<th>Residue</th>
<th>pK_{half}</th>
<th>pK_{half}</th>
<th>pK_{half}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experimental</td>
<td>CHARMM polar H</td>
<td>CHARMM</td>
</tr>
<tr>
<td>NTR1</td>
<td>7.9</td>
<td>7.81</td>
<td>8.00</td>
</tr>
<tr>
<td>LYS1</td>
<td>10.6</td>
<td>10.01</td>
<td>10.01</td>
</tr>
<tr>
<td>GLU7</td>
<td>2.9</td>
<td>3.17</td>
<td>3.39</td>
</tr>
<tr>
<td>LYS13</td>
<td>10.3</td>
<td>10.49</td>
<td>10.56</td>
</tr>
<tr>
<td>HIS15</td>
<td>5.4</td>
<td>6.20</td>
<td>5.87</td>
</tr>
<tr>
<td>ASP18</td>
<td>2.7</td>
<td>2.87</td>
<td>3.11</td>
</tr>
<tr>
<td>TYR20</td>
<td>10.3</td>
<td>10.85</td>
<td>11.18</td>
</tr>
<tr>
<td>TYR23</td>
<td>9.8</td>
<td>10.16</td>
<td>10.87</td>
</tr>
<tr>
<td>LYS33</td>
<td>10.4</td>
<td>10.58</td>
<td>10.79</td>
</tr>
<tr>
<td>GLU35</td>
<td>6.2</td>
<td>5.05</td>
<td>5.90</td>
</tr>
<tr>
<td>ASP48</td>
<td>2.5</td>
<td>2.96</td>
<td>2.91</td>
</tr>
<tr>
<td>ASP52</td>
<td>3.7</td>
<td>4.32</td>
<td>4.67</td>
</tr>
<tr>
<td>TYR53</td>
<td>&gt;12</td>
<td>11.71</td>
<td>&gt;12</td>
</tr>
<tr>
<td>ASP66</td>
<td>&lt;2.0</td>
<td>2.15</td>
<td>2.87</td>
</tr>
<tr>
<td>ASP87</td>
<td>2.1</td>
<td>2.43</td>
<td>2.97</td>
</tr>
<tr>
<td>LYS96</td>
<td>10.7</td>
<td>11.18</td>
<td>11.42</td>
</tr>
<tr>
<td>LYS97</td>
<td>10.1</td>
<td>10.79</td>
<td>10.85</td>
</tr>
<tr>
<td>ASP101</td>
<td>4.1</td>
<td>3.89</td>
<td>3.92</td>
</tr>
<tr>
<td>LYS116</td>
<td>10.2</td>
<td>10.12</td>
<td>10.09</td>
</tr>
<tr>
<td>ASP119</td>
<td>3.2</td>
<td>3.08</td>
<td>3.28</td>
</tr>
<tr>
<td>CTR129</td>
<td>2.8</td>
<td>2.73</td>
<td>2.83</td>
</tr>
<tr>
<td>RMSD</td>
<td><strong>0.45</strong></td>
<td><strong>0.57</strong></td>
<td></td>
</tr>
</tbody>
</table>

1. Results are from “A Fast and Accurate Computational Approach to Protein Ionization”, Spassov et al, submitted
The pH-dependence of lysozyme total charge at different ionic strengths. Open circles and solid line: experiment and computed values at 0.1 M ionic strength; filled circles and broken line: the same at 1.0 M ionic strength.

The pH dependent contribution to the binding energy of KNI-272 inhibitor to HIV-1 protease, compared to the experimental values of association constant Ka taken from Velazquez-Campoy et al. The solid line represents the computed values of $-\Delta \Delta G_{\text{bind}}$. The triangles correspond to the values of $2.303RT \log Ka$

1. Results are from “A Fast and Accurate Computational Approach to Protein Ionization”, Spassov et al, submitted
2. Protein Sci. 2000 Sep;9(9):1801-9
Comparative Validation

<table>
<thead>
<tr>
<th></th>
<th>Ngr</th>
<th>GB/IMC</th>
<th>MCCE</th>
<th>Const. pH</th>
<th>FD/DH</th>
<th>SCP</th>
<th>PROPKA</th>
</tr>
</thead>
<tbody>
<tr>
<td>4pti</td>
<td>14</td>
<td>0.36</td>
<td>0.47</td>
<td>NA</td>
<td>0.35</td>
<td>0.33</td>
<td>0.6</td>
</tr>
<tr>
<td>2lzt</td>
<td>21</td>
<td>0.45</td>
<td>0.76</td>
<td>0.6</td>
<td>0.47</td>
<td>0.49</td>
<td>0.66</td>
</tr>
<tr>
<td>2m2</td>
<td>25</td>
<td>0.59</td>
<td>0.87</td>
<td>0.9</td>
<td>1.17</td>
<td>0.57</td>
<td>0.72</td>
</tr>
<tr>
<td>3m3</td>
<td>16</td>
<td>0.44</td>
<td>0.66</td>
<td>1.2</td>
<td>0.87</td>
<td>0.55</td>
<td>0.94</td>
</tr>
<tr>
<td>1pga</td>
<td>15</td>
<td>0.42</td>
<td>0.63</td>
<td>NA</td>
<td>0.80</td>
<td>0.59</td>
<td>0.72</td>
</tr>
<tr>
<td>3icb</td>
<td>10</td>
<td>0.33</td>
<td>0.38</td>
<td>NA</td>
<td>0.37</td>
<td>0.39</td>
<td>0.9</td>
</tr>
<tr>
<td>3rnt</td>
<td>4</td>
<td>0.28</td>
<td>0.54</td>
<td>NA</td>
<td>NA</td>
<td>0.41</td>
<td>NA</td>
</tr>
<tr>
<td>average</td>
<td></td>
<td>0.41</td>
<td>0.63</td>
<td>-</td>
<td>0.67</td>
<td>0.49</td>
<td>0.76</td>
</tr>
</tbody>
</table>


comparison of the accuracy of Accelrys pK predictions with other methods.

MCCE – FDPB based method with Monte Carlo

Const. pH – Molecular Dynamics

FD/DH – Finite Difference/Debye-Hückel interactions

SCP – Sigmoidally Screened Columb potentials

PROPKA – Empirical relationships of position and chemistry of residues closed to pK sites

The CPU time used to calculate the ionization of proteins with different chain lengths. The data were generated on an Intel Pentium4 3.0 GHz machine.

1. Results are from “A Fast and Accurate Computational Approach to Protein Ionization”, Spassov et al, submitted.
Application 1: Accurate Protein-Ligand Binding Energies

• Estimate the electrostatic contribution to binding energy for protein-ligand and protein-protein docking
  – Protonation states of proteins may differ in docked and undocked form
    • This method predicts protonation states accurately taking into account of the local environment changes upon ligand docking
  – Faster than DelPhi methods
    • Only two calculations required

Fast and accurate electrostatic component of binding energy due to consideration of proper protonation states

pH-Dependent binding energy of HIV protease
Application 2: pH Dependent Folding Energy

- Calculation done for HIV protease

- Optimal pH for protein stability
  - Experimental ~ 5
  - Predicted ~ 4.6

- Folding free energy change from pH 3.5 to 5
  - $\Delta\Delta G_{\text{calc}}(\text{pH}=3.4\rightarrow5.0) = 3.5$ kcal/mol
  - $\Delta\Delta G_{\text{exp}}(\text{pH}=3.4\rightarrow5.0) \sim 4.5$ kcal/mol

Optimal pH for stability critical for all biochemical assays
Application 2: pH Dependent Folding Energy

- **Calculation done for HIV protease**

- **Optimal pH for protein stability**
  - Experimental ~ 5
  - Predicted ~ 4.6

- **Folding free energy change from pH 3.5 to 5**
  - $\Delta\Delta G_{\text{calc}}(\text{pH}=3.4->5.0) = 3.5$ kcal/mol
  - $\Delta\Delta G_{\text{exp}} (\text{pH}=3.4->5.0) \sim 4.5$ kcal/mol

Optimal pH for stability critical for all biochemical assays
Application 3: Active Site Residue prediction

• Predict potential active sites based on analysis of titration curve for specific residues
  – A cluster of two or more perturbed residues in physical proximity is a reliable predictor of active site location\(^1\)

• pK prediction experiment with Triose Isomerase (TIM) structure (PDB ID 1tph) – (demo)

Protein Ionization and pK prediction

Overall workflow for pK prediction

1. System preparation
   - Extremely rare to find experimentally determined protein structures ready for simulations calculations (main reason – missing hydrogens)
   - Protein Report useful to find out what is “missing”
   - Automatic cleaning available
     • Need to optimize newly built side-chains
   - Parameter and topology definition – “Typing”
     • Addition of hydrogens
     • Atom type assignment

2. Experiment options and submission
   - Protein dielectric constant
   - Ionic strength

   ![Parameter Table]
   
<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Parameter Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Typed Protein Molecule</td>
<td>T1nh:F1nh</td>
</tr>
<tr>
<td>Protein Dielectric Constant</td>
<td>10</td>
</tr>
<tr>
<td>Solvent Dielectric Constant</td>
<td>80</td>
</tr>
<tr>
<td>Ionic Strength</td>
<td>0.145</td>
</tr>
<tr>
<td>pH Range</td>
<td></td>
</tr>
<tr>
<td>From pH</td>
<td>0</td>
</tr>
<tr>
<td>To pH</td>
<td>14</td>
</tr>
<tr>
<td>Protonate Molecule</td>
<td>True</td>
</tr>
<tr>
<td>At pH</td>
<td>7.4</td>
</tr>
<tr>
<td>Advanced</td>
<td></td>
</tr>
</tbody>
</table>
– Setting up a pK prediction job
  • Report, cleaning and typing
– Analysis of the results
  • Titration plots
  • 1tph example
• Turkey Ovomucoid Third Domain (OMTKY3) is a small natural serine protease inhibitor.
  
  
  
  
  – Lots of experimentally available pKa data for most of its ionizable residues (PPD Protein pKa Database). Very stable at high pH.
### Accuracy - Experimental vs Calculated (GBpK)

- **Experimental pKa data available for 1omu.pdb (NMR structure – 50 models)**

<table>
<thead>
<tr>
<th>Titratable Residue</th>
<th>Experimental</th>
<th>Calculated (CHARMm) – model 01</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.015M</td>
<td>0.2M</td>
</tr>
<tr>
<td>N-term</td>
<td>7.96</td>
<td>7.70</td>
</tr>
<tr>
<td>ASP 7</td>
<td>2.7</td>
<td>2.73</td>
</tr>
<tr>
<td>GLU 10</td>
<td>4.2</td>
<td>3.91</td>
</tr>
<tr>
<td>TYR 11</td>
<td>10.16</td>
<td>10.13</td>
</tr>
<tr>
<td>LYS 13</td>
<td>9.86</td>
<td>9.87</td>
</tr>
<tr>
<td>GLU 19</td>
<td>3.2</td>
<td>3.57</td>
</tr>
<tr>
<td>TYR 20</td>
<td>11.07</td>
<td>11.1</td>
</tr>
<tr>
<td>ARG 21</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ASP 27</td>
<td>2.3</td>
<td>3.36</td>
</tr>
<tr>
<td>LYS 29</td>
<td>11.12</td>
<td>10.91</td>
</tr>
<tr>
<td>TYR 31</td>
<td>12.5</td>
<td>12.26</td>
</tr>
<tr>
<td>LYS 34</td>
<td>10.13</td>
<td>10.12</td>
</tr>
<tr>
<td>GLU 43</td>
<td>4.8</td>
<td>4.52</td>
</tr>
<tr>
<td>HIS 52</td>
<td>7.5</td>
<td>6.65</td>
</tr>
<tr>
<td><strong>LYS 55</strong></td>
<td>11.1</td>
<td>10.79</td>
</tr>
<tr>
<td>CYS 56</td>
<td>2.5</td>
<td>2.92</td>
</tr>
</tbody>
</table>

**Bold** indicates predictions >1.5 pH units from experiment.
Exp to Calc pKa prediction (0.015M) - model 01
CHARMm all H

Observation:
Higher variation amongst basic groups…

1omu.pdb (model 01/50)

R = 0.973

$R^2 = 0.947$

Slope = 1.006
Y-intercept = 0.261

RMSD = 0.90 pH units
Accuracy - Experimental vs Calculated (GBpK) cnt’d

1omu.pdb (model 01/50)

R = 0.985
R^2 = 0.971
Slope = 1.017
Y-intercept = 0.046
RMSD = 0.66 pH units

Observation:
Better correlation with CHARMM polar H
• Other interesting observations 1omu.pdb (model 01):

  – Significant pKa shifts: (>1)

<table>
<thead>
<tr>
<th>Residue</th>
<th>Standard pKa*</th>
<th>Experimental pKa (15mM)</th>
<th>Predicted Forsyth, et al. 1998</th>
<th>GBpK (CHARMm H)</th>
<th>GBpK (CHARMm polar-H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asp 7</td>
<td>3.67</td>
<td>2.7</td>
<td>3.3</td>
<td>2.73</td>
<td>2.63</td>
</tr>
<tr>
<td>Glu 19</td>
<td>4.25</td>
<td>3.2</td>
<td>2.8</td>
<td>3.57</td>
<td>3.52</td>
</tr>
<tr>
<td>Asp 27</td>
<td>3.67</td>
<td>2.3</td>
<td>3.4</td>
<td>3.36</td>
<td>3.13</td>
</tr>
<tr>
<td>His 52</td>
<td>6.54</td>
<td>7.5</td>
<td>6.2</td>
<td>6.77</td>
<td>6.42</td>
</tr>
</tbody>
</table>

GBpK (full H and polar-H) predicts 3/4 correctly within 1 pH unit!!!

* Updated values according to Thurlkill, et al (2006)
LYS 55 – QM/MM vs GBpK (model 01)

- Experimental value for LYS 55 = 11.10

- QM/MM calculated value = 11.00 (model 01)

- GBpK calculated value (model 01): CHARMm full H = 11.17
  CHARMM polar H = 11.24

- 10 days computation time (4GB, 3/4 nodes RS/6000 44P 270 workstation)

- ~30 secs/run computation time (2GB, 1 node, 2.8Ghz, Win XP IBM workstation)
LYS 55 –GBpK (CHARMm, all models)

- Experimental value for LYS 55 = 11.10

**Histogram - LYS 55 (GBpK)**

Average = 10.99 ± 0.27

**Histogram - LYS 55 (GBpK - polar-H)**

Average = 11.02 ± 0.23
Conclusion

• GBpK is a GBORN-based method that estimates pK within just a few minutes
  – More accurate and rigorous than template-based methods
  – Faster than existing Poison-Boltzmann methods

• Very simple to use
  – Requires a known structure
  – Only one parameterized variable: dielectric constant ($\varepsilon$) of protein interior

• Method can be used to identify optimal pH for proteins

• Correct protein protonation based on calculated pK can improve quality of:
  – Molecular dynamics simulations
  – Protein-ligand docking
  – Protein-ligand binding energy calculation (electrostatic part)

• Methods can be used to identify and characterize active site residues

• Predictions (OMTKY3) show high correlation with exp. data & good RMSD deviation (< than 1 pH unit)

• Calculations for OMTKY3 - LYS 55 are in good agreement with experimental data and QM/MM predicted values

• Unusual pK shifts of OMTKY3 are well predicted
Acknowledgements

• Velin Spassov
• Lisa Yan
• Paul Flook
• Dipesh Risal
• Shikha O’Brien
• Thank You for attending today’s webinar. If you have any further questions please e-mail me at: fhernandez@accelrys.com

• You can also contact us using the form on our website: http://accelrys.com/company/contact/

• We will be exhibiting at the following upcoming events:
  – CHI Protein Kinase Targets (June 23 – 25, Boston, Booth #4)
  – CHI Structure Based Design (June 25 – 27, Boston, Booth #7)
  – Drug Discovery Technology and Development (August 4 – 7, Boston, Booth #512)
  – ACS Fall 2008 (August 17 – 21, Philadelphia, Booth #211)

• Reminder: the next webinar in this series will be:

  “Towards Increased Accuracy in Computational Drug Discovery with QM/MM”

  June 26, 2008 at 7am PST and 10am PST
References:


• Spassov, V. et al. “LOOPER: A Molecular Mechanics Based Algorithm for Protein Loop Prediction”, Accepted, *Protein Engineering, Design and Selection*

• Spassov, V. and Yan, L. “A Fast and Accurate Computational Approach to Protein Ionization” *Submitted*, 2008