

# H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations

Ramu Anandakrishnan<sup>1</sup>, Boris Aguilar<sup>1</sup> and Alexey V. Onufriev<sup>2,\*</sup>

<sup>1</sup>Department of Computer Science and <sup>2</sup>Department of Computer Science and Physics, Virginia Tech, Blacksburg, VA 24061, USA

Received February 16, 2012; Revised April 10, 2012; Accepted April 12, 2012

## ABSTRACT

The accuracy of atomistic biomolecular modeling and simulation studies depend on the accuracy of the input structures. Preparing these structures for an atomistic modeling task, such as molecular dynamics (MD) simulation, can involve the use of a variety of different tools for: correcting errors, adding missing atoms, filling valences with hydrogens, predicting pK values for titratable amino acids, assigning predefined partial charges and radii to all atoms, and generating force field parameter/topology files for MD. Identifying, installing and effectively using the appropriate tools for each of these tasks can be difficult for novice and time-consuming for experienced users. H++ (<http://biophysics.cs.vt.edu/>) is a free open-source web server that automates the above key steps in the preparation of biomolecular structures for molecular modeling and simulations. H++ also performs extensive error and consistency checking, providing error/warning messages together with the suggested corrections. In addition to numerous minor improvements, the latest version of H++ includes several new capabilities and options: fix erroneous (flipped) side chain conformations for HIS, GLN and ASN, include a ligand in the input structure, process nucleic acid structures and generate a solvent box with specified number of common ions for explicit solvent MD.

## INTRODUCTION

Molecular modeling and simulations are routinely used to study the structure, function and activity of biomolecules (1–6). The accuracy of such modeling and simulations depends critically on the accuracy of the input structure (7). Approximately 88% of the structures in the protein

data bank (PDB) (8) are determined by X-ray crystallography, which can not, in general, resolve positions of most hydrogen atoms. While the positions for most of the missing hydrogen atoms can be easily estimated using simple chemical rules, predicting the protonation states of titratable groups such as the side chains of ASP, GLU, ARG, LYS, TYR, HIS or CYS is non-trivial. Their protonation states depend on complex electrostatic interactions between these groups themselves and with the surrounding environment.

Many theoretical methods have been developed for predicting protonation states of titratable groups (9–23). While these methods vary in the details of the underlying physical models, they share one common feature—computational and algorithmic complexity. Hence, the computational process usually involves multiple non-trivial steps. There is often an additional complication arising from errors and inconsistencies in the input PDB structures, requiring significant ‘pre-processing’ of structures. As a result, predicting protonation equilibria (pK), adding missing protons to PDB structures and generating the necessary input files for molecular dynamics (MD) simulation can involve the use of multiple tools from different packages. The complexity of this computational ‘pipe-line’ makes it difficult for novice and even moderately experienced users to identify, install, learn and effectively use the appropriate tools, increasing the likelihood of errors and creating unnecessary barriers for the adoption of these tools. Even for experts, the careful setup required for MD simulations takes a considerable amount of time.

H++ (<http://biophysics.cs.vt.edu/H++>) is a freely available open-source web server for automating the addition of missing protons to the PDB structures based on predicted pKs of titratable groups, and generating basic parameter/topology and coordinate files for MD simulations (currently only in AMBER format). The calculations are based on the standard continuum solvent methodology (11), within the framework of the Poisson Boltzmann (PB) model (21,24). The server is intended for both

\*To whom correspondence should be addressed. Tel: +1 540 231 4237; Fax: +1 540 231 6075; Email: alexey@cs.vt.edu

experts and non-experts. Typical default values are used for processing parameters and options, although a limited subset of parameters and options can be modified on the input screen.

The original implementation of *H++* was described in Gordon *et al.* (25). The present work summarizes the key features of *H++* and describes in additional detail significant new enhancements since the original implementation, see the ‘Materials and Methods’ section below. The accuracy of the current version of *H++* is presented in the ‘Results’ section.

## MATERIALS AND METHODS

The input structure to be processed by *H++* can be either in the PDB format, or alternatively in the PQR format, which is similar to the PDB format with atomic charges and radii included. Figure 1 shows the five main processing steps in *H++*, which are described below.

### Step 1: pre-process the input structure

Input PDB files can contain numerous errors and format inconsistencies, such as missing heavy atoms, suboptimal residue conformations and non-standard atom names. *H++* attempts to make automatic, albeit conservative corrections for many of these problems when possible. Otherwise the errors are identified for possible manual correction. For example, the N and O atoms in the amide groups of ASN and GLN, and the N and C atoms in the imidazole ring of HIS cannot be easily distinguished from electron density maps. Thus, the assignment of these atoms in the PDB file may be (optionally) ‘flipped’ using the `reduce` algorithm that is based on an analysis of van der Waals contacts and H-bonding (26).

An example of errors that are identified for manual correction are missing residues in the middle of protein chains. Input PDB files may also contain HETATM entries for solvent and ligand molecules; *H++* removes these entries. Solvent molecules are removed by default because they are treated implicitly by the continuum solvent methodology used. Non-protein ligands are removed by default, but an option is now available to manually include many ligands and specific buried water molecules for processing, as described on the *H++* site. Inclusion of buried waters has been shown to improve the accuracy of computed pK of nearby groups (27). For peptide, protein, DNA and RNA ligands, current AMBER force field parameters are used to add H atoms and assign atomic partial charges. For other organic ligands, *H++* uses OpenBabel (28) to add H atoms, and atomic partial charges are assigned using the `ante-chamber` module from AmberTools (29) and the generalized AMBER force field (GAFF) parameter set. PDB structures may also contain residues with partial occupancy representing multiple possible conformations. Without manual intervention from the user, *H++* selects the ‘A’ conformation and ignores all others.

An input PQR file, on the other hand, is assumed to have already been validated (e.g. in order to compute the atomic charges and radii included in the PQR file). Therefore, most error and consistency checks are bypassed for input PQR files. In addition, *H++* requires AMBER compatible atom and residue names in the input PQR file.

### Step 2: add missing atoms

In this step, missing heavy atoms and protons are added (assuming standard protonation states of all titratable

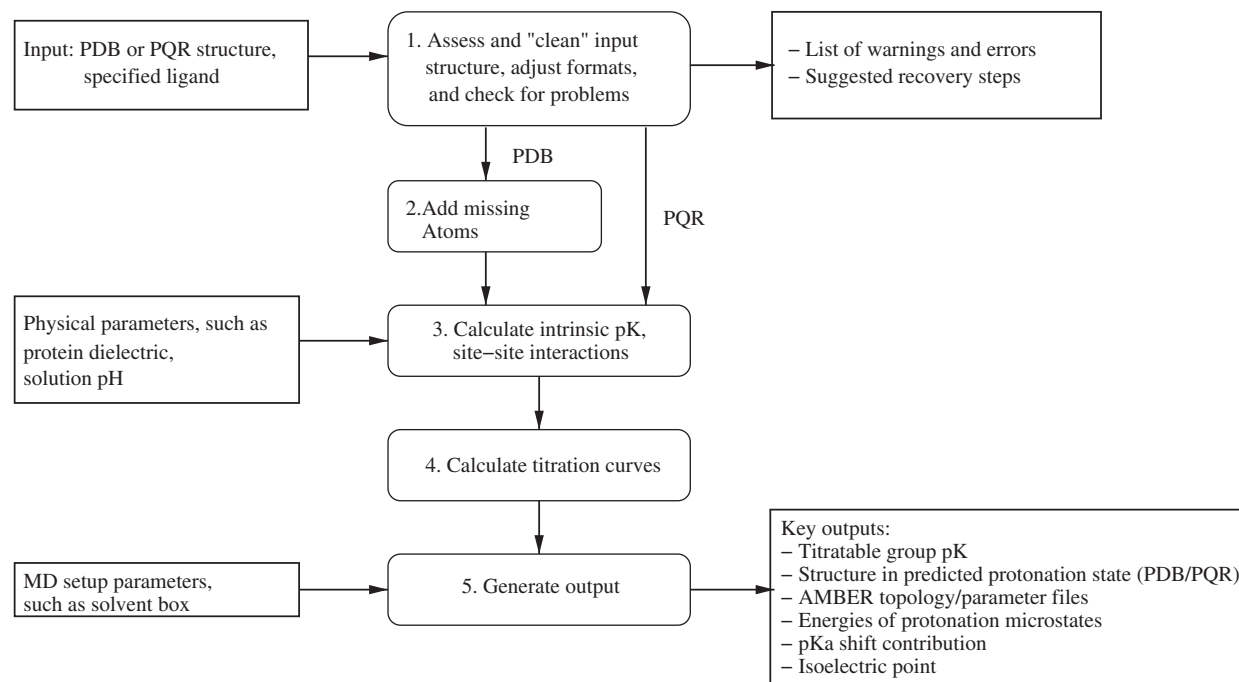


Figure 1. Flowchart of computations performed by the *H++* server described here.

groups at this stage), and atomic partial charges and radii [Bondi (30)] are assigned using the LEAP modules in AmberTools. The position of added atoms, and assignment of atomic partial charges and radii are based on the latest AMBER force field parameters (ff10). This step is bypassed for PQR files, (Figure 1) assuming that missing atoms have been added, and other consistency checks have been performed by the user.

### Step 3: calculate interaction energies between titratable groups

The interaction energies between all titratable groups and other “fixed” charges in the molecule are required to calculate the titration curves in the next step. The finite difference Poisson-Boltzmann (FDPB) continuum electrostatics methodology (11,31), implemented in MEAD (32), is used to calculate the energetics of proton transfer. See Gordon *et al.* (25) for additional details.

A key input to the FDPB computations are the atomic partial charges in the protonated and deprotonated states of titratable groups. The partial charges from the AMBER force field parameters are used for the protonated and deprotonated states of ASP, CYS, GLU, HIS, LYS, and the deprotonated state of ARG and TYR. Since AMBER does not contain parameters for the rarely occurring protonated (neutral) state of ARG and TYR, these partial charges were obtained from Ryde (33) and MEAD, respectively.

*H++* uses the single-conformer version of MEAD. However, conformational variability is partially accounted for by the ‘smear charge’ representation of titratable groups, and the optimization of the positions of added protons. In the case of ASP and GLU, the charge of the additional proton in the protonated state is equally divided between the two O atoms to which it may be attached. In the case of ARG, HIS and LYS, the charge reduction in the deprotonated state is equally divided between the protons that may be removed. While not the most systematic or exhaustive way of incorporating conformational variability, we believe, based on previous experiences (18,31,34,35), that this particular model is a reasonable balance between speed and accuracy, and is therefore a good choice for a web-based calculation.

### Step 4: calculate titration curves

The titration curve for a titratable group is the protonation probability as a function of solution pH. The protonation probability can be calculated as the Boltzmann average over the protonation states for the set of interaction energies calculated above (11). However, the exact computation of these statistical averages involves summations over all possible  $2^N$  microstates, where  $N$  is the number of titratable groups. To make the computation of Boltzmann averages tractable, two different approximations are used. For structures with less than 80 titratable groups, a clustering algorithm is employed to approximate the Boltzmann averages (36,37). For larger structures, the full Monte Carlo method as implemented in MCTI (12), is used. The  $pK$  of each group is computed as the mid-point of the

corresponding titration curve, that is  $pK = pK_{1/2}$ . Users are encouraged to check shapes of titration curves displayed by *H++*, as  $pK_{1/2}$  loses its meaning for titration curves that strongly deviate from the standard sigmoidal shape (38).

### Step 5: generating outputs

*H++* produces several outputs useful for molecular modeling and simulations. These include the estimated  $pK$  value for each titratable group, a PDB and the corresponding PQR file in the predicted protonation state, AMBER format topology and force field parameter files for explicit and implicit solvent simulations, energies of protonation microstates (for structures with <25 titratable groups), breakdown of contribution to  $pK$  shift of each group and the computed isoelectric point.

For efficiency, the titration curve is only calculated in the experimentally accessible pH range from 0 to 12. In some cases, the pH value at which the calculated protonation probability is 0.5 may be outside this range. For cases where the 0.5 protonation probability occurs outside the above pH range, the  $pK$  value is reported as ‘<0’ or ‘>12’.

The input PDB file is updated with titratable groups in their predicted protonation states based on estimated  $pK$  values and the user specified pH value—protonated if  $pH < pK$ , deprotonated otherwise. The PDB file is updated using the LEAP module in AmberTools which also adds atomic radii and partial charges producing a PQR format file. The LEAP module also produces AMBER force field parameter and coordinate files for running molecular dynamics simulations using AMBER or NAMD. These files can optionally include a cubic or octahedral solvent box with the user-specified number and type of (commonly used) ions, for explicit solvent simulations.

## RESULTS

The accuracy of *H++* was evaluated for a set of 23 high-quality structures (X-ray structures with resolution  $\leq 2.50 \text{ \AA}$  and no missing residues in the middle of the protein sequence), for which experimentally determined  $pK$  values are available (39) (Table 1). The set includes 201 titratable groups, of which 66 have large  $pK$  shifts ( $\Delta pK \geq 1.0$ ) and 135 have small  $pK$  shifts ( $\Delta pK < 1.0$ ). For comparison, predictions of a ‘Null’ model are also shown. Additional details for the results shown in Table 1, including PDB codes of the test proteins, can be found at <http://biophysics.cs.vt.edu/faq.php>.

Most of the test sets used for benchmarking  $pK$  prediction methods, including the one described above, lack groups with extreme  $pK$  shifts. To address this deficiency, we added to our test set a membrane-embedded protein bacteriorhodopsin (BR), a system with functionally relevant extreme  $pK$  shifts (40–42) (Table 2). The trends of the predicted  $pK$  shifts are in reasonable agreement with experiment. However, for the deeply buried groups, the use of a protein dielectric constant of 4, lower than



**Table 1.** RMS error in pK values computed by *H++* relative to experiment, based on 23 protein structures with a total of 201 titratable groups

$\Delta pK$	RMS error		Prot. accuracy (%) at pH 7		$\Delta pK$ direction accuracy (%)	
	<i>H++</i>	Null model	<i>H++</i>	Null model	<i>H++</i>	Null model
$\geq 1.0$	1.44	2.05	98	88	92	50
$< 1.0$	1.39	0.51	97	95	75	50

Results are broken down into two sets. One with large ( $>1.0$ ) pK shifts (66 groups) and the other with small ( $\leq 1.0$ ) pK shifts (135 groups). Table shows the RMS error, accuracy of predicted protonation at pH 7.0 and accuracy of predicted direction of pK shift, compared to experimentally determined values. The Null model assumes a pK shift of 0 from the model compound pK for the purpose of computing RMS error and predicted protonation state. For direction of the predicted pK shift the Null model assumes equal probabilities of  $\pm$  shift.

**Table 2.** Experimental and computed pK values of the membrane-embedded Bacteriorhodopsin

Site	Experiment	<i>H++</i>	
		$\epsilon_{in} = 6$ (default)	$\epsilon_{in} = 4$
ARG 82	$>13.8$	$>12$	$>12$
ASP 85	2.6	2.3	1.8
ASP 86	$>12$	7.6	$>12$
ASP 115	$>9.5$	8.0	10.4
ASP 212	$<2.5$	$<0$	$<0$
Schiff Base (216)	$>12$	$>12$	$>12$

We used two values of protein dielectric ( $\epsilon_{in}$ ) constant. The structure corresponds to the BR resting state of the photocycle, and is constructed from PDB ID 1QJH, as described in Onufriev *et al.* (34).

the default value  $\epsilon_{in} = 6$ , results in better agreement with experiment.

There are many factors that limit the accuracy of pK predicted by *H++*. Arguably, the more important factors are: shortcomings of the semi-microscopic, 2-dielectric PB implicit solvent model; incomplete accounting of conformational flexibility; and the absence of coupling between protonation and conformational states. Additional factors that affect the accuracy of pK prediction in general, can be found in (20) and (43). The PB implicit solvent model in *H++* uses a fixed dielectric value for the solute (protein), even though the appropriate dielectric value can vary considerably depending on the location of the group in question within the protein (16,20,27,44,45). Significant improvement in pK prediction have also been reported when conformational flexibility is taken into account (46,47). Moreover, constant pH molecular dynamics studies have shown a strong correlation between conformation and protonation states (48,49). Inclusion of variable dielectric, conformational flexibility and the coupling between protonation states using currently available methods involve significant additional computational costs, which is counter to our goal of a fast web-based application.

## CONCLUSION

The free open-source *H++* web server described here automates the process of pK estimation and many of the steps required to prepare biomolecular structures for atomistic simulations. Input PDB structures are protonated according to the computed pKs, and force-field parameter/topology and coordinate files compatible with AMBER/NAMD formats are generated. The server is intended for both novice and expert users alike. A distinctive feature of *H++* is that it includes extensive input consistency checks, safeguards, and pre- and post-processing warning messages that can help users make informed decisions when using the outputs from the server. The pK calculations in *H++* are based on the established PB continuum electrostatics methodology that has been successfully used for this purpose for over two decades. The accuracy and usefulness of *H++* has been improved in the latest version which includes several new capabilities: such as fixing erroneous (flipped) residue conformations; including a ligand in the input structure; processing nucleic acid structures; and generating a solvent box with specified number of common ions for explicit solvent MD. In addition, upgrades to the *H++* server resulted in a  $3\times$  speedup relative to *H++* 1.0 (25). We continue to enhance *H++*, with future plans to automate ligand processing and to add the ability to generate force field parameter/topology files for other popular modeling packages. Since its original release in August 2005, over 38 000 structures have been uploaded to *H++* for processing by 2287 registered users and thousands of anonymous users, proving the usefulness of this tool to the molecular modeling and simulation community. The source code for *H++* is also available, upon request, for offline batch processing of a large number of structures.

## ACKNOWLEDGEMENTS

We thank John Gordon for helping to maintain the *H++* server during the first 3 years of its operation.

## FUNDING

Funding open access charge: National Institutes of Health [GM076121 to A.V.O.].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Dodson, G.G., Lane, D.P. and Verma, C.S. (2008) Molecular simulations of protein dynamics: new windows on mechanisms in biology. *EMBO Rep.*, **9**, 144–150.
2. Freddolino, P.L., Arkhipov, A.S., Larson, S.B., McPherson, A. and Schulten, K. (2006) Molecular dynamics simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure*, **14**, 437–449.
3. Karplus, M. and McCammon, J.A. (2002) Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.*, **9**, 646–652.
4. Karplus, M. and Kuriyan, J. (2005) Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6679–6685.
5. Wang, W., Donini, O., Reyes, C.M. and Kollman, P.A. (2001) Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein,

- and protein-nucleic acid noncovalent interactions. *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 211–243.
6. Schlick, T., Guevara, R.C., Halvorsen, L.A., Jung, S. and Xiao, X. (2011) Biomolecular modeling and simulation: A field coming of age. *Quart. Rev. Biophys.*, **44**, 191–228.
  7. Garcia-Viloca, M., Poulsen, T.D., Truhlar, D.G. and Gao, J. (2004) Sensitivity of molecular dynamics simulations to the choice of the X-ray structure used to model an enzymatic reaction. *Protein Science*, **13**, 2341–2354.
  8. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.
  9. Tanford, C. and Kirkwood, J. (1957) Theory of Protein Titration Curves. *J. Am. Chem. Soc.*, **79**, 5333–5339.
  10. Tanford, C. and Roxby, R. (1972) Interpretation of Protein Titration Curves. *Biochemistry*, **11**, 2192–2198.
  11. Bashford, D. and Karplus, M. (1990)  $pK_a$ 's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry*, **29**, 10219–10225.
  12. Beroza, P., Fredkin, D.R., Okamura, M.Y. and Feher, G. (1991) Protonation of interacting residues in a protein by Monte Carlo method. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 5804–5808.
  13. Takahashi, T., Nakamura, H. and Walda, A. (1992) Electrostatic forces in two lysozymes: Calculations and measurements of histidine  $pK_a$  values. *Biopolymers*, **32**, 897–909.
  14. Yang, A.-S., Gunner, M.R., Sampogna, R., Sharp, K. and Honig, B. (1993) On the Calculation of  $pK_a$ 's in Proteins. *Proteins*, **15**, 252–265.
  15. DelBuono, G., Figueirido, F. and Levy, R. (1994) Intrinsic  $pK_a$ 's of ionizable residues in proteins: An explicit solvent calculation for lysozyme. *Proteins*, **20**, 85–97.
  16. Demchuk, E. and Wade, R.C. (1996) Improving the continuum dielectric approach to calculating  $pK_a$ 's of ionizable groups in proteins. *J. Phys. Chem.*, **100**, 17373–17387.
  17. Sham, Y.Y., Chu, Z.T. and Warshel, A. (1997) Consistent Calculations of  $pK_a$ 's of Ionizable Residues in Proteins: Semi-microscopic and Microscopic Approaches. *J. Phys. Chem.*, **101**, 4458–4472.
  18. Ullmann, G.M. and Knapp, E.-W. (1999) Electrostatic models for computing protonation and redox equilibria in proteins. *Eur. Biophys. J.*, **28**, 533–551.
  19. Spassov, V.Z. and Bashford, D. (1999) Multiple-Site Ligand Binding to Flexible Macromolecules. *J. Comp. Chem.*, **20**, 1091–1111.
  20. Antosiewicz, J., McCammon, J.A. and Gilson, M.K. (1994) Prediction of pH-dependent Properties of Proteins. *J. Mol. Biol.*, **238**, 415–436.
  21. Nielsen, J.E. and Vriend, G. (2001) Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based  $pK_a$  calculations. *Proteins*, **43**, 403–412.
  22. Georgescu, R., Alexov, E. and Gunner, M. (2002) Combining Conformational Flexibility and Continuum Electrostatics for Calculating  $pK_a$ 's in Proteins. *Biophysical Journal*, **83**, 1731–1748.
  23. Mongan, J., Case, D.A. and McCammon, J.A. (2004) Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. *J. Comp. Chem.*, **25**, 2038–48.
  24. Baker, N.A., Bashford, D. and Case, D.A. (2006) Implicit solvent electrostatics in biomolecular simulation. In: Leimkuhler, B., Chipot, C., Elber, R., Laaksonen, A., Mark, A., Schlick, T., Schütte, C. and Skeel, R. (eds), *New Algorithms for Macromolecular Simulation*, Vol. 49. Springer Berlin/Heidelberg, pp. 263–295, Chapter 15.
  25. Gordon, J.C., Myers, J.B., Folta, T., Shoja, V., Heath, L.S. and Onufriev, A.V. (2005) H<sup>++</sup>: a server for estimating  $pK_a$ 's and adding missing hydrogens to macromolecules. *Nucleic Acids Res.*, **33**, 68–71.
  26. Word, M.J., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, **285**, 1735–1747.
  27. Harms, M.J., Schlessman, J.L., Chimenti, M.S., Sue, G.R., Damjanović, A. and García-Moreno, B. (2008) A buried lysine that titrates with a normal  $pK_a$ : role of conformational flexibility at the protein-water interface as a determinant of  $pK_a$  values. *Protein sci.*, **17**, 833–845.
  28. O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T. and Hutchison, G. (2011) Open Babel: An open chemical toolbox. *J. Cheminform.*, **3**, 33.
  29. Case, D.A., Cheatham, T.E., Darden, T., Gohlke, H., Luo, R., Merz, K.M., Onufriev, A., Simmerling, C., Wang, B. and Woods, R.J. (2005) The Amber biomolecular simulation programs. *J. Comp. Chem.*, **26**, 1668–1688.
  30. Bondi, A. (1964) van der Waals Volumes and Radii. *J. Phys. Chem.*, **68**, 441–451.
  31. Bashford, D. and Gerwert, K. (1992) Electrostatic Calculations of the  $pK_a$  Values of Ionizable Groups in Bacteriorhodopsin. *J. Mol. Biol.*, **224**, 473–486.
  32. Bashford, D. (1997) An object-oriented programming suite electrostatic effects in biological molecules. An experience report on the MEAD project. Scientific Computing in Object-Oriented Parallel Environments. Vol. 1343. Springer Berlin / Heidelberg, of *Lecture Notes in Computer Science*, chapter 30 pp. 233–240.
  33. Sigfridsson, E. and Ryde, U. (1998) Comparison of methods for deriving atomic charges from the electrostatic potential and moments. *J. Comput. Chem.*, **19**, 377–395.
  34. Onufriev, A., Smondryev, A. and Bashford, D. (2003) Proton Affinity Changes During Unidirectional Proton Transport in the Bacteriorhodopsin Photocycle. *J. Mol. Biol.*, **332**, 1183–1193.
  35. Spassov, V.Z. and Yan, L. (2008) A fast and accurate computational approach to protein ionization. *Protein Sci.*, **17**, 1955–1970.
  36. Anandakrishnan, R. and Onufriev, A. (2008) Analysis of Basic Clustering Algorithms for Numerical Estimation of Statistical Averages in Biomolecules. *J. Comput. Biol.*, **15**, 165–184.
  37. Myers, J., Grothaus, G., Narayanan, S. and Onufriev, A. (2006) A simple clustering algorithm can be accurate enough for use in calculations of  $pK_a$ 's in macromolecules. *Proteins*, **63**, 928–938.
  38. Onufriev, A., Case, D.A. and Ullmann, G.M. (2001) A novel view of pH titration in biomolecules. *Biochemistry*, **40**, 3413–3419.
  39. Grimsley, G.R., Scholtz, J.M. and Pace, C.N. (2009) A summary of the measured  $pK_a$  values of the ionizable groups in folded proteins. *Prot. Sci.*, **18**, 247–251.
  40. Spassov, V.Z., Luecke, H., Gerwert, K. and Bashford, D. (2001)  $pK_a$  calculations suggest storage of an excess proton in a hydrogen-bonded water network in bacteriorhodopsin. *J. Mol. Biol.*, **312**, 203–219.
  41. Kono, M., Misra, S. and Ebrey, T.G. (1993) pH dependence of light-induced proton release by bacteriorhodopsin. *FEBS Lett.*, **331**, 31–34.
  42. Balashov, S.P., Imasheva, E.S., Govindjee, R. and Ebrey, T.G. (1996) Titration of aspartate-85 in bacteriorhodopsin: what it says about chromophore isomerization and proton release. *Biophys. J.*, **70**, 473–481.
  43. Castañeda, C.A., Fitch, C.A., Majumdar, A., Khangulov, V., Schlessman, J.L. and García-Moreno, B.E. (2009) Molecular determinants of the  $pK_a$  values of Asp and Glu residues in staphylococcal nuclease. *Proteins*, **77**, 570–588.
  44. Schutz, C.N. and Warshel, A. (2001) What are the dielectric constants of proteins and how to validate electrostatic models? *Proteins*, **44**, 400–417.
  45. Isom, D.G., Castañeda, C.A., Cannon, B.R., Velu, P.D. and García-Moreno, B.E. (2010) Charges in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci. USA*, **107**, 16096–16100.
  46. Song, Y., Mao, J. and Gunner, M.R. (2009) MCCE2: Improving protein  $pK_a$  calculations with extensive side chain rotamer sampling. *J. Comput. Chem.*, **30**, 2231–2247.
  47. Powers, N. and Jensen, J.H. (2006) Chemically accurate protein structures: Validation of protein NMR structures by comparison of measured and predicted  $pK_a$  values. *J. Biomol. NMR*, **35**, 39–51.
  48. Lee, M.S., Salisbury, F.R. and Brooks, C.L. (2004) Constant-pH molecular dynamics using continuous titration coordinates. *Proteins*, **56**, 738–752.
  49. Mongan, J., Case, D.A. and McCammon, J.A. (2004) Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.*, **25**, 2038–2048.