Development of theoretical methods for describing the protonation states of solvated molecules based on the integral equation theory of liquids

藤木, 涼

https://hdl.handle.net/2324/5068165

出版情報：Kyushu University, 2022, 博士（理学）, 課程博士
バージョン：
権利関係：

Development of theoretical methods for describing the protonation states of solvated molecules based on the integral equation theory of liquids

Ryo Fujiki

Development of theoretical methods for describing the protonation states of solvated molecules based on the integral equation theory of liquids

Ryo Fujiki

Department of chemistry

Graduate School of Science

Kyushu University

2022

Contents

1 General introduction

1.1 General introduction

The protonation state of molecules is one of the fundamental elements determining reactivity, structures, and functions in chemical and biological processes. Therefore, the protonation state of molecules in solution is attracting much attention in the field of chemistry, biochemistry, biophysics, and pharmacy. Changes in the protonation state give rise to changes in various features, such as the determination of the reaction path, the structural change of molecules, and the rearrangement of hydrogen bond network. For example, in biochemistry, the presence or absence of dissociative hydrogen affects the formation of higher-order structures of proteins or DNA. All these reactions are strongly affected by the surrounding solution environment therefore, the relationship between the protonation state of target molecules and solution environments is a very interesting research subject.

The equilibrium constant for the deprotonation reaction is known as the acid dissociation constant (K_a). It indicates the abundance ratio of protonated and deprotonated forms at the equilibrium state of a molecule. In experiments, nuclear magnetic resonance (NMR) spectroscopy is widely used to determine the K_a value by measuring the chemical shift at various pH conditions [1-4]. The transfer of protons can be directly observed as chemical shifts. However, if the target molecule has multiple allogeneic dissociative residues as in proteins, assignment of the obtained chemical shifts is very difficult because of the similarity of each type of dissociative residue. Consequently, Holmes et al. have reported a useful approach sensitive to hydrogen atom positions, to compliment diffraction methods [5]. The attribution based on the environment around the target residue is required for accurate analysis. The neutron diffraction (ND) method is also useful; it can detect protons directly [6]. However, for obtaining a good analysis of detailed properties, a crystal that is very pure is required. However, the latter is often difficult to achieve because of the crystal's fragile nature.

Faced with difficulties encountered in the experimental approach, a theoretical approach to determine the protonation state should be considered. PROPKA is one of the most successful methods for the prediction of protonation states of amino acids in proteins [7]. The method is based on empirical parameters, depending on protein fragment structure data and it can estimate $\mathrm{p}K_\mathrm{a}$ values with good accuracy. However, it may be better to use nonempirical methods for systems that have large structural fluctuations where there is significant electronic structure reorganization, because it is difficult to apply empirical parameters.

The most straightforward way to determine the protonation states of solvated molecules based on nonempirical methods is to estimate the free energy change of the deprotonation reaction. The protonation states are affected by the surrounding molecules, including the solvent. Therefore, the solvation models are essential for the free energy computation. In consideration of the solvent effect, there are two major approaches that exist, namely, the explicit and implicit solvation models. The former is a method of arranging a finite number of solvent molecules explicitly and then considering their correlation. It is used in the framework, such as molecular dynamics (MD) and Monte Carlo (MC) simulation [8]. The latter is a method in which the solvent effect is considered by approximating solvents as a continuum medium and characterizing it by the macroscopic parameters such as the dielectric constant [9].

Explicit solvent models can consider the microscopic intermolecular interactions directly and assist in revealing the effect of the molecular motion of solutes and solvents on the chemical changes of solute molecules. On the other hand, as the size of the system to be handled is limited, it is difficult to estimate how much sampling for the configurational space is required to obtain an appropriate ensemble average.

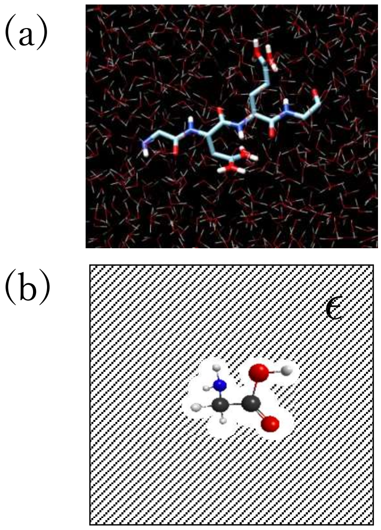

Figure 1.1. The illustration of (a) explicit and (b) implicit solvent model.

Implicit solvent dielectric continuum models, such as generalized Born (GB) or the polarizable continuum model (PCM) are well known that treat solvent molecules as continuum media. Their advantage is the low computational cost, based on approximation in which the solvent environment is regarded as a polarizable dielectric continuum model characterized by a dielectric constant (ε). For example, Cramer et al. have suggested the optimized continuum model SM8, which has been shown to be useful for various indicators [10]. Takano et al. developed a conductor-like PCM (CPCM) and applied it to the base-catalyzed hydrolysis of methyl acetate in water [11]. However, continuum models are inadequate when considering microscopic solvent effects, such as hydrogen bonding.

Another approach to taking solvent effects into account is the integral equation theory of molecular liquids. This theory is also an implicit solvation model but it is able to consider explicit intermolecular interaction because solvent molecules are treated as interaction sites. Three-dimensional reference interaction site model (3D-RISM) theory is one of the most successful integral equation theories of liquids [12-14]. As such treatment offers lower calculation costs, 3D-RISM theory can be applied to large molecular systems, such as proteins or DNA in solution. Furthermore, the molecular properties of solvents are considered even in highly anisotropic environments, such as inside the clefts of proteins, where the macroscopic dielectric constant cannot be determined. 3D-RISM theory gives the distribution function of solvent molecules around a solute, and more detailed microscopic effects between solute and solvent are obtained, compared with PCM. The integral equation theories including the reference interaction site model (RISM) and 3D-RISM have been applied to estimate the K_a value of water, amino acids and drug-like molecules [15-18]. However, it is suggested that the results are not quantitative because of the difficulties in considering the structural fluctuation and estimating the free energy of excess protons.

In this thesis, the linear fitting method based on reference data set is adopted for development of quantitative prediction and the hybrid method of constant pH MD (CpHMD) and 3D-RISM is suggested for consideration of structural fluctuation.

Chapter 2 describes the development of a prediction method of the protonation state in water based on the hybrid theory, the 3D-RISM self-consistent field (3D-RISM-SCF) [19][20][21] and linear fitting correction (LFC) scheme (LFC/3D-RISM-SCF), and then application to amino acids. Chapter 3 describes the application of the LFC/3D-RISM-SCF method for the prediction of the protonation state in methanol. Chapter 4 describes the development of a prediction method of the protonation state based on CpHMD simulation coupled with 3D-RISM theory and application to polypeptides. Chapter 5 is devoted to general conclusions.

References

[1] M. Tanokura, M. Tasumi and T. Miyazawa, *Biopolymers*,1976, **15**, 393–401.

[2] K. Bartik, C. Redfield and C. M. Dobson, *Biophys*. J., 1994, **66**, 1180–1184.

[3] M. D. Joshi, A. Hedberg and L. P. McIntosh, *Protein Sci.*,1997, **6**, 2667–2670.

[4] A. Jadhav, V. S. Kalyani, N. Barooah, D. D. Malkhede and J. Mohanty, *Chem. Phys. Chem.*, 2015, **16**, 420–427.

[5] J. B. Holmes, V. Liu, B. G. Caulkins, E. Hilario, R. K. Ghosh, V. N. Drago, R. P. Young, J. A. Romero, A. D. Gill, P. M. Bogie, J. Paulino, X. Wang, G. Riviere, Y. K. Bosken, J. Struppe,

A. Hassan, J. Guidoulianov, B. Perrone, F. Mentink-Vigier, C. A. Chang, J. R. Long, R. J. Hooley, T. C. Mueser, M. F. Dunn, L. J. Mueller, *Proceedings of the National Academy of Sciences*, 2022, **2**, 119.

[6] Pertsemlidis, A.; Saxena, A. M.; Soper, A. K.; Head-Gordon, T.; Glaeser, R. M. *Proc. Natl. Acad. Sci. U.S.A.*,1996, 93, 10769.

[7] H. Li, A.D. Robertson, J.H. Jensen, *Proteins Struct. Funct. Genet.* 2005, **61**, 704.

[8] F. Jensen, “Introduction to Computational Chemistry”, Wiley (2007)

[9] J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999–3093.

[10] C. J. Cramer, D. G. Truhlar, *Acc. Chem. Res.*, 2008, **41**, 760.

[11] Y. Takano, K. N. Houk, *J. Chem. Theory Comput.*, 2005, **1**, 70.

[12] D. Beglov, B. Roux, *J. Phys. Chem. B*, 1997, **101**, 7821–7826.

[13] D. Beglov, B. Roux, *J. Chem. Phys.*, 1996, **104**, 8678–8689.

[14] A. Kovalenko, F. Hirata, *Chem. Phys. Lett.*, 1998, **290**, 237–244.

[15] Y. Seno, N. Yoshida, H. Nakano, *J. Mol. Liquids*, 2016, **217**, 93-98.

[16] Sato, H.; Hirata, F., *J. Phys. Chem. B*, 1999, **103**, 6596－6604

[17] K. Kido, H. Sato, S. Sakaki, *Int. J. Quantum. Chem.* 2012, **112**, 103-112

[18] N. Yoshida, R. Ishizuka, H. Sato, F. Hirata, *J. Phys. Chem. B* 2006, **110**, 16, 8451–8458

[19] S. Ten-No, F. Hirata, S. Kato, *Chem. Phys. Lett.*, 1993, **214**, 391–396.

[20] A. Kovalenko, F. Hirata, *J. Chem. Phys.*, 1999, **110**, 10095–10112.

[21] H. Sato, A. Kovalenko, F. Hirata, *J. Chem. Phys.*, 2000, **112**, 9463–9468.

2 Development of quantitative prediction method of protonation state based on quantum chemical calculation and integral equation theory of liquids

2.1 Introduction

Protonation and deprotonation reactions are fundamental in various fields related to biological systems. The protonated state of a molecule is an important factor involved in the formation of higher-order structures and the strength of intramolecular interactions. For the intricate parts such as those found in proteins, they may behave completely differently from protonation in bulk water [1-3]. This is because the environment inside a protein is hydrophobic and dissociation is less likely to occur. The acid dissociation constant (K_a) or its logarithmic value ($\mathrm{p}K_\mathrm{a}$) is usually determined for the index of protonation by experiment and measured by nuclear magnetic resonance (NMR) with titration or neutron diffraction (ND) [4-6]. However, each method has some form of disadvantage. Assignment of the signals of multiple protons obtained by NMR is difficult and a pure and large crystal of the target molecule is required for ND. Hence, more recently, theoretical approaches for the determination of the $\mathrm{p}K_\mathrm{a}$ value are attracting increasing interest.

The $\mathrm{p}K_\mathrm{a}$ value is expressed by the Gibbs free energy difference of deprotonation reaction ($\Delta\mathrm{G}$); the relationship is given by,

$$\mathrm{p}K_\mathrm{a} = \frac{\Delta G}{(\ln 10)RT} \tag{1}$$

where R and T are the gas constant and absolute temperature, respectively, and

$$\Delta\mathrm{G} = \mathrm{G}(\mathrm{A}^-) + G(\mathrm{H}^+) - G(\mathrm{HA}) \tag{2}$$

where $G(\mathrm{X})$ denotes a Gibbs energy of species X. Here, HA and A^- represent the protonated and deprotonated states of acid A.

One of the major approaches for obtaining the Gibbs energy of a solvated molecule is calculation using ab initio molecular dynamics methods [7][8]. However, as they require substantial computational costs, it is impractical to apply them to complex molecular systems.

More recently, the hybrid quantum mechanics and molecular mechanics (QM/MM) method is commonly used for $\mathrm{p}K_\mathrm{a}$ evaluation [9][10]. In this approach, only the reactive moiety is treated by the ab initio molecular orbital (MO) or Kohn–Sham density functional theory (KS-DFT). The remaining parts are treated by classical molecular mechanics. As a more approximate and efficient method, hybrid methods with the implicit solvation models such as the PCM [48], or the statistical mechanics integral equation theory of liquids, such as the reference interaction site model (RISM), or three-dimensional RISM (3D-RISM) theory, have attracted wide focus [11][12]. An advantage of these methods is the qualitative evaluation of solvation free energy within a reasonable computational time.

However, treatment of the Gibbs energy of the proton, $\mathrm{G(H^+)}$, is difficult for the quantitative evaluation of the $\mathrm{p}K_\mathrm{a}$ value, because the degree of association of water molecules in the solution is unclear, and there is not major way to handle it. One of the common approaches is treatment of the excess proton that exists as the hydronium ion ($\mathrm{H_3O^+}$) or a water n-mer, but there may be a factor of quantitative error in the $\mathrm{p}K_\mathrm{a}$ value. For this reason, adopting empirical parameters and corrections for proton free energy is a well-known method to evaluate the $\mathrm{p}K_\mathrm{a}$ value in water solution [13-15]. While empirical parameters are easy to understand and useful for tracking the behavior of water, it should be noted that they cannot respond to the subtleties in a specific environment [16][17]. Matsui et al. proposed a scheme based on the linear relationship between the $\mathrm{p}K_\mathrm{a}$ value and the Gibbs energy difference between HA and $\mathrm{A^-}$ [18-20]. What should be noted about the method is the low computational cost but high accuracy. In this chapter, this method is referred to the linear fitting correction (LFC) scheme as a generic term.

In the LFC scheme, the $\mathrm{p}K_\mathrm{a}$ values of target molecules are determined from calculations with fitted parameters. These are obtained from the least squares method to the experimental values of training set molecules. By employing this scheme, the problematic free energy calculation of protons is avoided. This scheme has been successfully applied to the evaluation of $\mathrm{p}K_\mathrm{a}$ values of amino acids. The results show good agreement with experimental observations. The LFC scheme of earlier research employs the PCM for consideration of the solvent effect on the electronic structure. The PCM and related methods are widely used to investigate chemical processes in solution. However, as the solvent environment is handled uniformly, because of treatment as a dielectric continuum characterized by a dielectric constant, it is difficult to reproduce the local molecular interactions, such as hydrogen bonding and to define the dielectric constant in a heterogeneous environment, such as inside a protein.

In this chapter, a new scheme based on the LFC scheme employing the 3D-RISM as a solvent model is proposed. Hybrid methods of the 3D-RISM theory and the quantum chemical theory, such as KS-DFT and ab initio MO, have been proposed by Kovalenko, Sato, and Hirata. These methods are referred to as KS-DFT/3D-RISM or three-dimensional reference interaction site model self-consistent field (3D-RISM-SCF) [21][22]. 3D-RISM-SCF has been applied to various chemical processes in a solution, including the $\mathrm{p}K_\mathrm{a}$ shift of drug molecules [23][24]. The method allows us to treat a highly anisotropic solvent environment, such as inside a cavity and channel of a protein. Therefore, by employing 3D-RISM-SCF, we expect to establish a method that is applicable to complex biological systems.

The layout of this section is as follows. Section 2.2 provides an introduction to theoretical methods. Section 2.3 gives computational details. Section 2.4 describes results and discussions. In Section 2.4.1, the parameters for the LFC scheme are determined by least squares fitting based on the Gibbs energy of the training set molecules calculated by 3D-RISM-SCF and the corresponding experimental $\mathrm{p}K_\mathrm{a}$ values. In Section 2.4.2, the basis set dependency on the performance of the scheme is also examined. In Section 2.4.3, the scheme is applied to amino acids to assess the transferability of the fitted parameters. The chapter is concluded with a summary in Section 2.5.

2.2 Methods

2.2.1 Linear fitting correction method with empirical parameters

The $\mathrm{p}K_\mathrm{a}$ value is related to the Gibbs energy difference of the acid dissociation reaction, ΔG, equation (1) is rewritten by introducing the scaling factor s,

$$\mathrm{p}K_\mathrm{a} = \frac{s\{\mathrm{G}(\mathrm{A}^-) - G(\mathrm{HA})\}}{(\ln 10)RT} + \frac{s\{G(\mathrm{H}^+)\}}{(\ln 10)RT} = k\Delta G_0 + C_0 \tag{3}$$

where,

$$k = \frac{s}{(\ln 10)RT} \tag{4}$$

$$\Delta G_0 = \mathrm{G}(\mathrm{A}^-) - G(\mathrm{HA}) \tag{5}$$

$$C_0 = \frac{s\{G(\mathrm{H}^+)\}}{(\ln 10)RT} \tag{6}$$

The scaling factor s should be unity when the calculated Gibbs energy values are identical to exact values, and $\mathrm{k} = 0.733\ \mathrm{mol\,kcal^{-1}}$ when $s = 1$ at 298.15 K. The scaling factor s is an adjustable parameter, which corresponds to the activity coefficient of deprotonation reaction and corrects the systematic error of the computational method. The parameters k and C_0 were determined by the least square fitting to minimize the errors of $\mathrm{p}K_\mathrm{a}$ values,

$$\varepsilon = \sum_i \left\{\mathrm{p}K_{\mathrm{a}i}^{\mathrm{expt}} - \left(k\Delta G_{0,i} + C_0\right)\right\}^2 \tag{7}$$

where $\mathrm{p}K_{\mathrm{a}i}^{\mathrm{expt}}$ is an experimental $\mathrm{p}K_\mathrm{a}$ value of molecule i and the summation over i is taken for all molecules in the training set that have the same dissociative chemical group and those $\mathrm{p}K_\mathrm{a}$ values are already known. $\Delta G_{0,i}$ is evaluated using ab initio MO or KS-DFT with a solvation model such as the PCM. The parameters k and C_0 are determined for each of the dissociative chemical groups, such as carboxyl, amine, alcohol, thiol, phenol, and imidazole.

2.2.2 3D-RISM-SCF theory

In the original LFC methods, the PCM/DFT is adopted for ΔG_0 calculation. In this study, 3D-RISM-SCF is employed instead of the PCM/DFT for consideration the solvation effect. As the details of the 3D-RISM-SCF method can be found in the literature, we only provide a brief explanation of the theory here.

The Gibbs energy of the solute molecule in the solvent at infinite dilution is defined as the sum of the solute electronic energy (E_0), solvation free energy ($\Delta\mu$), and the kinetic free energy (G_{kin})

$$G = E_0 + \Delta\mu + G_{\mathrm{kin}} \tag{8}$$

where E_0 given by

$$E_0 = \langle \Psi | \hat{H}_0 | \Psi \rangle \tag{9}$$

and where $\hat{H}_0$ and Ψ denote the Hamiltonian of the isolated molecules and the electronic wave function of solute molecules. The kinetic free energy (G_{kin}) includes the vibrational, rotational and translational energies, which are obtained in a usual quantum mechanics manner after the normal mode analysis. In the present study, the kinetic term, G_{kin},is ignored because the change in this term caused by the deprotonation reaction is rather small and the adjustable parameter can absorb the error emerging from this approximation. The solvation free energy is given by

$$\Delta\mu = k_B T \sum_{i}^{\mathrm{solvent}} \rho_i \int \left[\frac{1}{2} h_i(\boldsymbol{r})^2 \Theta\left(-h_i(\boldsymbol{r})\right) - c_i(\boldsymbol{r}) - \frac{1}{2} h_i(\boldsymbol{r}) c_i(\boldsymbol{r}) \right] d\boldsymbol{r} \tag{10}$$

where i runs over the solvent interaction sites. Θ, k_B, T, and ρ_i denote the Heaviside step function, the Boltzmann constant, the absolute temperature, and the number density of the solvent site i, respectively. $h_i(\boldsymbol{r})$ and $c_i(\boldsymbol{r})$ are total and direct correlation functions, obtained by solving the 3D-RISM equation coupled with the Kovalenko–Hirata closure [25]

$$h_i(\boldsymbol{r}) = \sum_{j}^{\text{solvent}} c_j(\boldsymbol{r}) * X_{ij}(\boldsymbol{r}) \tag{11}$$

$$h_i(\boldsymbol{r}) = \begin{cases} \exp\left(\mathrm{d}_i(\boldsymbol{r})\right) - 1 \text{ for } \mathrm{d}_i(\boldsymbol{r}) < 0 \\ \quad -\mathrm{d}_i(\boldsymbol{r}) \qquad \text{for } \mathrm{d}_i(\boldsymbol{r}) \geq 0 \end{cases} \tag{12}$$

$$\mathrm{d}_i(\boldsymbol{r}) = -\frac{1}{k_{\mathrm{B}}T} u_i(\boldsymbol{r}) + h_i(\boldsymbol{r}) - c_i(\boldsymbol{r}) \tag{13}$$

where * denotes a convolution integral. $X_{ij}(\boldsymbol{r})$ is a solvent susceptibility function, obtained by solving the RISM equation for pure solvent systems prior to 3D-RISM-KH calculation. $u_i(\boldsymbol{r})$ is an interaction potential function between a solute molecule and solvent molecules at position r. In the 3D-RISM-SCF framework, $u_i(\boldsymbol{r})$ is given by

$$u_i(\boldsymbol{r}) = 4 \sum_{j}^{\text{solute}} \varepsilon_{ij} \left\{ \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} + \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right\} + q_i \sum_{j}^{\text{solute}} \frac{Z_j}{r_{ij}} - q_i \int \frac{|\Psi(r')|^2}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}' \tag{14}$$

where ε_{ij} and σ_{ij} are the Lennard-Jones parameters (with usual meanings), and q_i denotes the point electronic charge on the solvent site i. Z_j is a nuclear charge of atom j.

2.3 Computational Details

In the present study, the parameters for six chemical groups, alcohol, amine, imidazole, thiol, phenol, and carboxyl were determined. Table 2.1.1 - 2.1.3 summarize the training data sets for parameter fitting.

Prior to the Gibbs energy calculation, the structure optimization of protonated (HA) and deprotonated (A^-) states was performed at the B3LYP/6-31++G(d,p) level, in water, with the PCM, for all the training set molecules. For the Gibbs energy calculation, two different sizes of basis sets were employed, 6-31++G(d,p) and 6-31G, to examine the basis set dependency of the parameter fitting.

The parameters used in the 3D-RISM calculation were temperature of 298.15 K and density of solvent water of 1.0 $\mathrm{g\,cm^{-3}}$. The Lennard-Jones parameters for solute molecules were taken from the general Amber force field (GAFF) parameter set with antechamber software [26]. The extended simple point charge model (SPC/E) parameter set for the geometrical and potential parameters for the solvent water was employed with modified hydrogen parameters (σ = 1.0 Å, and ε = 0.056 $\mathrm{kcal\,mol^{-1}}$) [27][28]. The grid spacing for the 3D grid was 0.5 Å and the number of grid points on each axis was 128. All calculations were performed with a modified version of the GAMESS program package, for which the 3D-RISM-SCF program has been implemented [29-32].

Table 2.1. training data sets for parameter fitting in Carboxyl and Amine

Carboxyl	Molecule	pK_a	Ref.	Amine	Molecule	pK_a	Ref.
	CHOCOOH	3.32	(33)		$Ph(CH_2)_2NH_3^+$	9.83	(37)
	trans-CH_3CH=CHCOOH	4.69			$PhCH_2NH_3^+$	9.34	
	$Ph(OH)_2COOH$	4.48	(34)		$PhNH_3^+$	4.58	
	H_2C=$CHCH_2COOH$	4.42			$CH_3(CH_2)_3NH_3^+$	10.58	
	$(CH(OH)COOH)_2$	1.14			$CH_3CH_2NH_3^+$	10.67	
	$CHOHCH_3COOH$	3.86			$HO(CH_2)_2NH_3^+$	9.50	
	CH_3COCH_2COOH	3.58			$HONH_3^+$	5.96	
	$CH_3COCOOH$	2.50			NH_4^+	9.21	(36)
	$CHCl_2COOH$	1.29	(35)		H_2C=$CHCH_2NH_3^+$	9.49	
	CH_2FCOOH	2.66			$CH_3(CH_2)_2NH_3^+$	10.53	
	NO_2CH_2COOH	1.68			Cyclohexylamine	10.64	
	$PhNO_2COOH$	2.45	(36)		Cyclohexylmethyl amine	10.49	
					Isopropylamine	10.63	
					Methoxyamine	4.60	
					γ -Phenylpropyl amine	10.20	
					neo-Pentylamine	10.21	
					sec-Butylamine	10.56	

Table 2.1.2 training data sets for parameter fitting in Imidazole and Thiol

Imidazole	Molecule	pK_a	Ref.	Thiol	Molecule	pK_a	Ref.
	2-Methyl-4-hydroxy-aminobenzimidazole	6.65	(38)		$C_2H_5OCH_2CH_2SH$	9.38	(39)
	2-Methylbenzimidazole	6.10			$C_2H_5OCOCH_2SH$	7.95	
	2-Methylimidazole	7.75			$C_6H_5CH_2SH$	9.43	
	4-Hydroxy-6-aminobenzimidazole	5.90			$CH_2{=}CHCH_2SH$	9.96	
	4-Hydroxy benzimidazole	5.30			$HOCH_2CHOHCH_2SH$	9.51	
	4-Methoxy benzimidazole	5.10			n-C_3H_7SH	10.65	
	4-Nitroimidazole	1.50			n-C_4H_9SH	10.66	
	6-Nitrobenzimidazole	3.05			t-$C_5H_{11}SH$	11.21	
	Benzimidazole	5.40			2-Mercaptoethanol	9.50	(40)
	Imidazole	6.95			2-Mercaptoethylamine	8.60	
	2-Methyl-4-hydroxy-6-nitrobenzimidazole	3.90			Thioglycolic acid	10.31	
	4-Hydroxy-6-nitrobenzimidazole	3.05			Thiophenol	7.8	(41)
	4-(2-4-dihydroxy phenyl)-imidazole	6.45			o-Aminothiophenol	6.59	
	4-Methyl-imidazole	7.45			3-Mercaptopropionicacid	10.27	
	6-Aminobenzidazole	6.00					
	Histamine	6.00					

Table 2.1.3 training data sets for parameter fitting in Alcohol and Phenol

Alcohol	Molecule	pK_a	Ref.	Phenol	Molecule	pK_a	Ref.
	CCl_3CH_2OH	11.80	(42)		2Cl-4NO_2-phenol	5.42	(45)
	$CHF_2CF_2CH_2OH$	11.34			$C_3H_5CH_2O_2C$-phenol	8.41	(46)
	CH_2=$CHCH_2OH$	15.10			m-CH_3CO-phenol	9.19	
	CH_3CH_2OH	15.90	(43)		m-CH_3O-phenol	9.65	
	$CH_3OCH_2CH_2OH$	14.80			m-F-phenol	9.28	
	CH_3OH	15.54			m-$HOCH_2$-phenol	9.83	
	$CHCCH_2OH$	13.55			m-NH_2phenol	9.87	
	$CHCl_2CH_2OH$	12.89			o-OCH-phenol	6.79	
	$HOCH_2CF_2CH_2OH$	11.00			p-Br-phenol	9.34	
	CH_3OCH_2OH	14.80			p-$C_2H_5O_2C$-phenol	8.50	
	$C(CH_2OH)_4$	14.10			p-C_6H_5-phenol	9.51	
	$HOCH_2CHOHCH_2OH$	14.40			p-CH_3O_2C-phenol	8.47	
	C_2H_5OH	16.00			p-CH_3S-phenol	9.53	
	CF_3CH_2OH	12.37			p-CH_3SO_2-phenol	7.83	
	$HOCH_2CH_2OH$	14.77			p-HO-phenol	9.96	
	$CF_3C(CH_3)_2OH$	11.60			p-NC-phenol	7.95	
	$CF_3CH(OH)CH_3$	11.80	(44)		p-O_2C-phenol	9.39	
					p-$(CH_3)_3N^+$phenol	8.00	

2.4 Results and Discussion

2.4.1 Fitting parameter determination

Table 2.2 summarizes the determined parameters by fitting. The properties of computed and experimental $\mathrm{p}K_\mathrm{a}$ values are also shown in Fig. 2.1 with the $\mathrm{p}K_\mathrm{a}$ values computed without using the LFC scheme. Then, the acid dissociation reaction for the $\mathrm{p}K_\mathrm{a}$ calculation without the LFC is assumed as below,

$$\mathrm{HA} + \mathrm{H_2O} \leftrightarrow \mathrm{A^-} + \mathrm{H_3O^+} \tag{12}$$

and the associated $\mathrm{p}K_\mathrm{a}$ formula

$$\mathrm{p}K_\mathrm{a} = \frac{\mathrm{G}(\mathrm{A^-}) + G(\mathrm{H_3O^+}) - G(\mathrm{HA}) - G(\mathrm{H_2O})}{(\ln 10)RT} \tag{13}$$

where the Gibbs energy of each molecule is calculated by 3D-RISM-SCF method. In this study, we refer to this treatment as a direct 3D-RISM-SCF scheme.

The improvement of accuracy of the computed $\mathrm{p}K_\mathrm{a}$ value is obvious from Fig. 2.1 for all the chemical groups by combination of the LFC scheme and 3D-RISM-SCF method. We refer to this as the LFC/3D-RISM-SCF scheme. LFC/3D-RISM-SCF method shows enough reasonable score for both the root mean square error (RMSE), 0.709, and the correlation factor, $\mathrm{r} = 0.978$. Although the direct 3D-RISM-SCF scheme also shows good correlation with the experimental value, $\mathrm{r} = 0.912$, its $\mathrm{p}K_\mathrm{a}$ values are overestimated (RMSE is 18.43). It indicates that the Gibbs energy of the reaction is overestimated by direct 3D-RISM-SCF. Why such overestimation is suppressed in LFC/3D-RISM-SCF is the scaling factor s, because the range of s is 0.43 - 0.67.

Table 2.2 shows the good contribution of the Gibbs energy of the proton ($G(\mathrm{H^+})$). Table 2.2 summarizes the $G(\mathrm{H^+})$ ($= C_0/k$) values. The range of $G(\mathrm{H^+})$ values are from -255 to -248 $\mathrm{kcal\,mol^{-1}}$, and these results are comparable with those obtained in the previous LFC approach by Matsui et al., which ranged from -268 to -246 $\mathrm{kcal\,mol^{-1}}$. It is also similar to the experimental and other theoretical approaches, which ranged from −264 to −259 kcal mol^{-1}.

Table 2.2. Determined parameters of each chemical group and RMSE

	k (mol kcal^{-1})	C_0 (kcal mol^{-1})	RMSE	r	$G(\mathrm{H}^+)$ (kcal mol^{-1})
Alcohol	0.443	−112.460	1.175	0.698	−254.0
Amine	0.396	−98.600	0.469	0.973	−248.9
Imidazole	0.338	−84.960	0.629	0.927	−251.1
Thiol	0.490	−123.403	0.821	0.750	−252.0
Phenol	0.317	−78.788	0.423	0.931	−248.5
Carboxyl	0.319	−81.552	0.661	0.832	−255.3
Total			0.709	0.978	

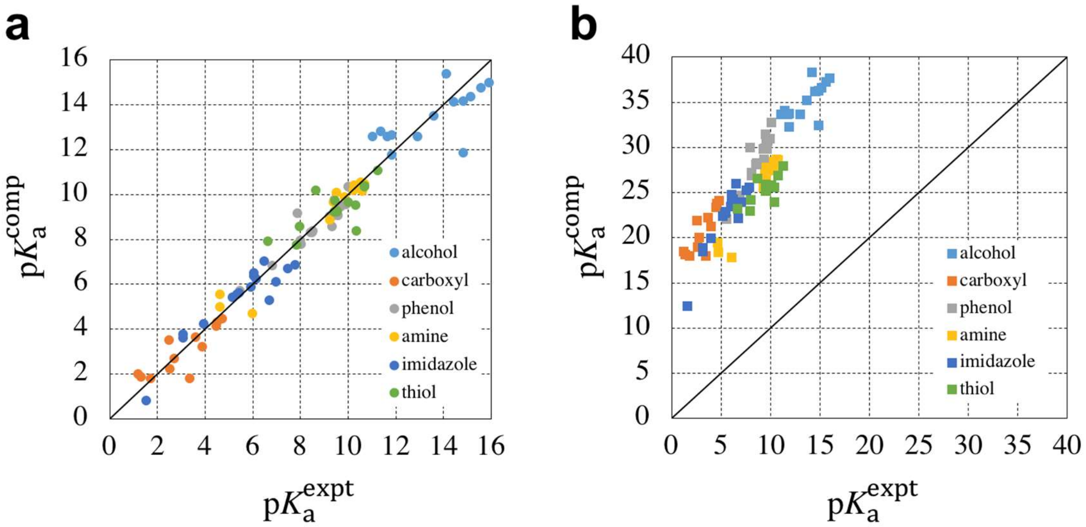

Figure 2.1. Comparison between the computed and experimental pK_a values, based on (a) LFC/3D-RISM-SCF and (b) direct 3D-RISM-SCF.

Fig. 2.2 shows the comparisons of computed and the experimental pK_a values. The correlation in alcohol and thiol is less than others, and they show large RMSE. In the case of alcohol group, a methoxyethanol (CH_3OCH_2OH) shows largest deviation (the experimental pK_a value is 14.8 and the LFC/3D-RISM-SCF value is 12.1). If the parameters are determined again with the training set except methoxyethanol, the RMSE and correlation are improved (RMSE is 0.88 and r is 0.87). On the other hand, in the thiol group, a mercaptoethylamine, and a thioglycolic acid show large deviations. These errors suggest that they have other factors affected to the fitting parameters other than the chemical group. For example, these molecules of thiol group have multiple dissociative groups, which might have some contribution to the accuracy. Therefore, for improvement of accuracy, it may be necessary to adopt additional parameters or factors.

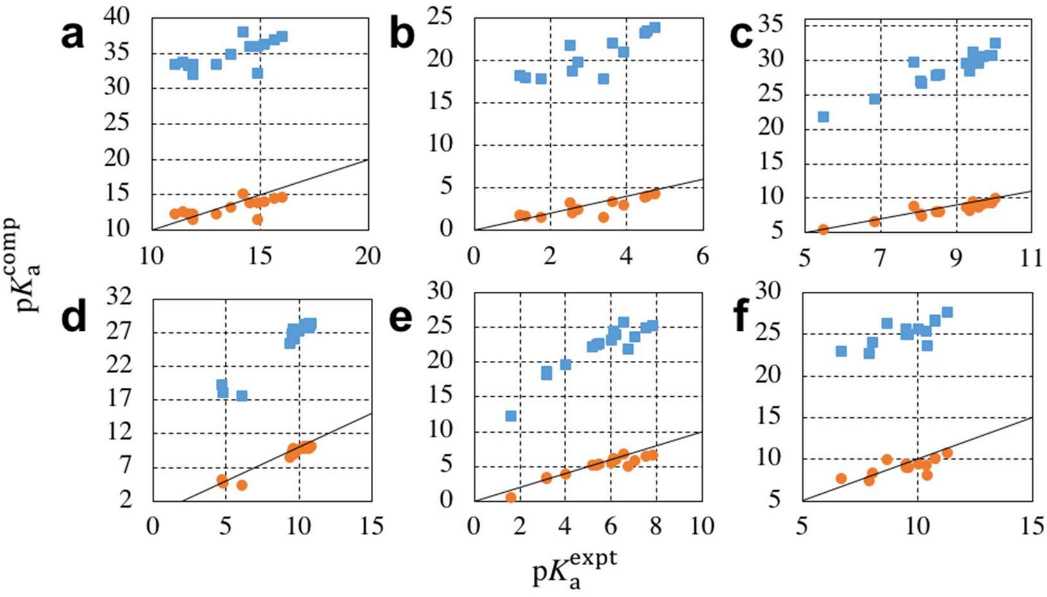

Figure 2.2. Comparison of the computed pK_a values with the experimental values in separate panels. (a) alcohol, (b) carboxyl, (c) phenol, (d) amine, (e) imidazole, and (f) thiol. The squares and circles denote the pK_a values determined by direct 3D-RISM-SCF and LFC/3D-RISM-SCF, respectively.

2.4.2 Investigation of basis set dependency

To investigate the basis set dependence of the parameters and their accuracy, another basis set, 6-31G, was examined. Table 2.2 shows the results of the fitted parameters and the experimental and computed pK_a values are compared in Fig. 2.3. While the parameter k of 6-31++G(d,p) is in the range of 0.3 – 0.48, that of 6-31G is in the range of 0.23 – 0.42. The RMSE and correlation values of 6-31G are slightly worse than that of 6-31++G(d,p). Though the parameters or errors of 6-31G are a little inferior to the results of 6-31++G(d,p), the accuracy of the results determined using LFC/3D-RISM-SCF is acceptable.

The direct 3D-RISM-SCF results of the thiol group is also remarkable. Its computed values shift a little close to experimental values. This may be because the description of the inadequate electronic structure by using a small basis set. Such an irregular behavior of a specific chemical group can be compensated by the parameters in the LFC scheme. This result clearly indicates that LFC/3D-RISM-SCF allows us to use the computationally cheaper basis set, thereby providing a significant advantage when the scheme is applied to large molecular systems such as biomolecules.

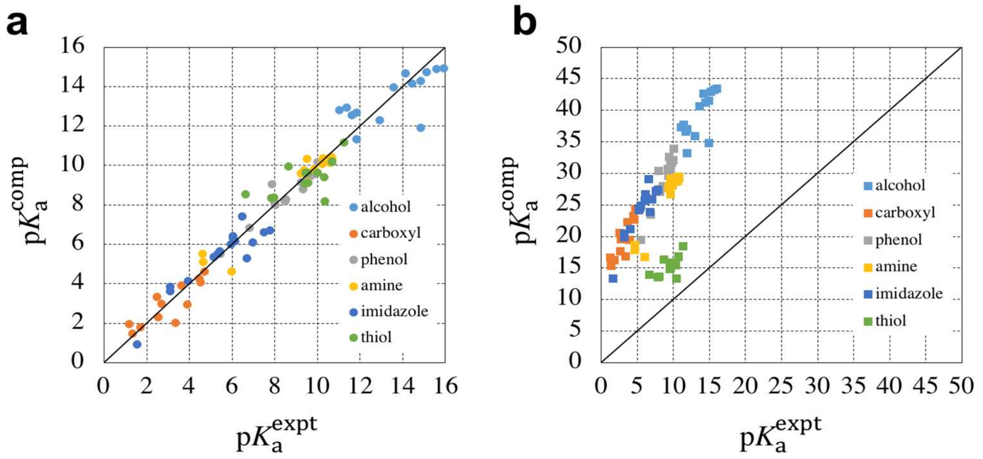

Figure 2.3. Comparison of the computed pK_a values using the 6-31G basis set with the experimental values, using (a) LFC/3D-RISM-SCF and (b) direct 3D-RISM-SCF. The references for the experimental values are given in Table 2.1.

2.4.3 pK_a calculation of amino acid

For the evaluation of transferability of the LFC/3D-RISM-SCF method to the biomolecules, the pK_a calculations of the dissociative amino acids by LFC/3D-RISM-SCF were examined. Table 2.3 and Fig. 2.4 respectively compare the experimental and computed pK_a values of several amino acid side chains. An aspartic acid (Asp), glutamic acid (Glu), cysteine (Cys), histidine (His), lysine (Lys), and tyrosine (Tyr) are examined. The computed pK_a values by LFC/3D-RISM-SCF show quantitative agreement with the experimental data. On the other hand, direct 3D-RISM-SCF shows much larger deviation than LFC scheme (the RMSE of LFC/3D-RISM-SCF is 0.39, that of direct 3D-RISM-SCF is 18.7). These results indicate that LFC/3D-RISM-SCF have good transferability and that it can be used for the pK_a prediction of proteins.

Table 2.3. Computed and experimental pK_a values of amino acids.[47]

		pK_a		
Amino acid	Chemical group	LFC/3D-RISM-SCF	Direct 3D-RISM-SCF	Expt.
Asp	Carboxyl	3.92	22.86	3.86
Cys	Thiol	9.14	25.07	8.33
Glu	Carboxyl	3.94	22.92	4.25
His(D/E)^b	Imidazole	6.31/6.27	22.92/24.39	6.04
Lys	Amine	10.69	24.29	10.53
Tyr	Phenol	9.64	28.95	10.07

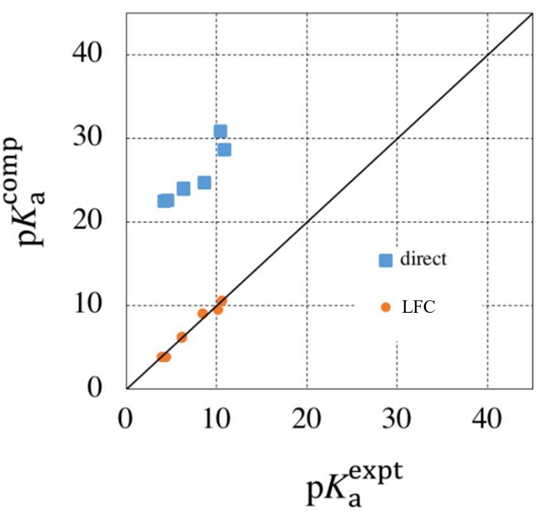

Figure 2.4. Comparison of the computed $\mathrm{p}K_\mathrm{a}$ values for amino acids with values determined experimentally. The filled squares and circles denote the computed $\mathrm{p}K_\mathrm{a}$ values by direct 3D-RISM-SCF and LFC/3D-RISM-SCF, respectively.

2.4.4 Solvent model dependency

In this section, the results of calculations with LFC/3D-RISM-SCF is compared with the results with PCM for the same training set for the assessment of the solvent model dependencies. Fig. 2.5 shows that the computed $\mathrm{p}K_\mathrm{a}$ values by the LFC and direct schemes are compared with the experimental $\mathrm{p}K_\mathrm{a}$ values. Table 2.1 summarized the determined parameters by fitting, RMSE, and correlation factors. While the correlation and RMSE of the direct PCM values with the experimental values is worse than those of 3D-RISM-SCF, correlation factor is 0.80 and total RMSE is 27.9, the computed $\mathrm{p}K_\mathrm{a}$ values by the LFC/PCM scheme archived high accuracy and good correlation, correlation factor is and 0.98 and the total RMSE and is 0.72. These are shown in Fig. 2.5a and b. In comparison with LFC/PCM, LFC/3D-RISM-SCF shows slightly better values in the RMSE and correlation factor than LFC/PCM. In the application to amino acids, excellent transferability of LFC/PCM is suggested from Fig. 2.5c and Table 2.1. The RMSE for the amino acids by LFC/PCM is 1.03, and that by LFC/3D-RISM-SCF is 0.39. This result indicates that LFC/3D-RISM-SCF has better transferability of the LFC scheme to biomolecules than the PCM.

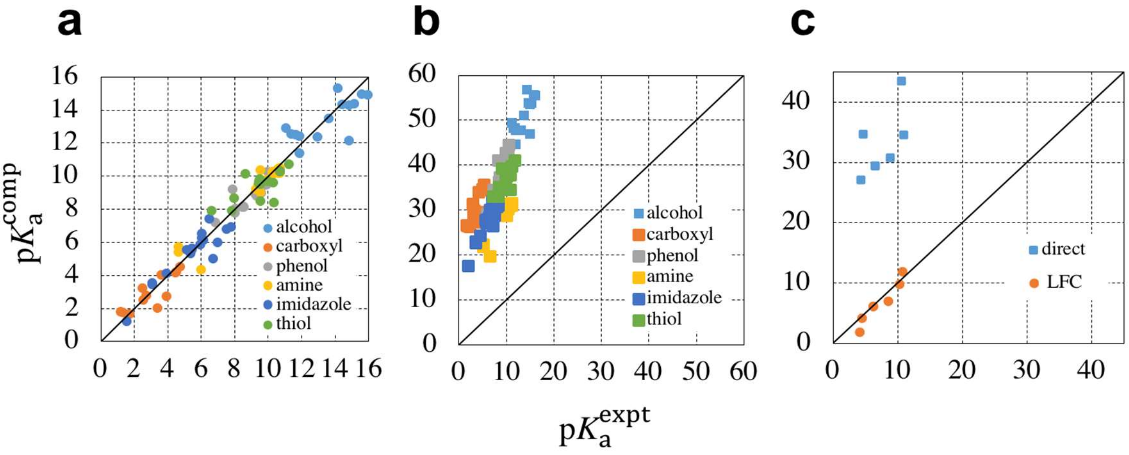

Figure 2.5. Comparison of the computed $\mathrm{p}K_\mathrm{a}$ values with the experimental values. (a) Comparison of the LFC/PCM values with the training set molecules. (b) Comparison of the direct PCM values with the training set molecules. (c) Comparison of LFC/PCM and direct PCM values with the amino acids. The filled squares and filled circles denote the direct and LFC values, respectively.

2.5 Summary

A scheme for computing $\mathrm{p}K_\mathrm{a}$ values based on 3D-RISM-SCF with the LFC scheme was proposed. The $\mathrm{p}K_\mathrm{a}$ value was computed by utilizing the linear relationship between the $\mathrm{p}K_\mathrm{a}$ value and the Gibbs energy difference between the protonated and deprotonated states of target molecules in this scheme. The parameters were determined by the least square fitting for the experimental values of a training set for each chemical group, and these parameters corresponded to the Gibbs energy of the excess proton and the scaling factor. The problem about excess proton in water is solved by adopting parameters, and the errors from the computational condition such as basis sets for electronic structure calculations are well dealt with. In addition, the computationally inexpensive basis set can be used for pK_a calculations in this scheme. After that, the parameters were applied to the amino acid molecules and the application shows a good performance. Furthermore, LFC/3D-RISM-SCF shows better behavior than the LFC/PCM scheme, especially in terms of the transferability of the parameters.

These features may allow us to use this scheme for the prediction of pK_a values of amino acids in biological systems. It will be very strong tool for analysis of protonation states. To apply the LFC/3D-RISM-SCF scheme to amino acids in proteins, a method taking account of environment other than water, such as surrounding residue and ions, which are not currently considered, is necessary. Previously, we proposed the use of advanced methods of 3D-RISM-SCF, in combination with quantum chemical methods, applicable to the biomolecular systems, which we referred to as the quantum mechanics/molecular mechanics/RISM (QM/MM/RISM) and the fragment molecular orbital/3D-RISM (FMO/3D-RISM) methods. The combination of the proposed scheme and QM/MM/RISM or FMO/3D-RISM may be a powerful tool to deal with the problems related to the protonation and deprotonation of dissociated amino acid residues in biological systems.

References

[1] A. A. Gorfe, P. Ferrara, A. Caflisch, D. N. Marti, H. R. Bosshard and I. Jelesarov, *Proteins*, 2002, **46**, 41-60.

[2] E. L. Mehler, M. Fuxreiter, I. Simon and E. B. Garcia-Moreno, *Proteins*, 2002, **48**, 283-292.

[3] D. G. Isom, C. A. Castaneda, B. R. Cannon and B. E. Garcia-Moreno, *Proc. Natl. Acad. Sci. USA*, 2011, **108**, 5260-5265.

[4] M. Tanokura, M. Tasumi and T. Miyazawa, *Biopolymers*, 1976, **15**, 393-401.

[5] K. Bartik, C. Redfield and C. M. Dobson, *Biophys. J.*, 1994, **66**, 1180-1184.

[6] M. D. Joshi, A. Hedberg and L. P. McIntosh, *Protein Sci.*, 1997, **6**, 2667-2670.

[7] A. K. Tummanapelli, S. Vasudevan, *J. Phys. Chem. B,* 2014, **118**, 13651-13657.

[8] A. K. Tummanapelli, S. Vasudevan, *Phys. Chem. Chem. Phys.*, 2015, **17**, 6383-6388.

[9] A. Warshel, *J. Phys. Chem.*, 1979, **83**, 1640-1652.

[10] A. Warshel, Biochemistry, 1981, **20**, 3167-3177.

[11] H. Sato and F. Hirata, J. Phys. Chem. A, 1998, **102**, 2603-2608.

[12] N. Yoshida, R. Ishizuka, H. Sato and F. Hirata, *J. Phys. Chem. B,* 2006, **110**, 8451-8458.

[13] H. Li, A. D. Robertson and J. H. Jensen, *Proteins*, 2005, **61**, 704-721.

[14] J. H. Jensen, H. Li, A. D. Robertson and P. A. Molina, *J. Phys. Chem. A,* 2005, **109**, 6634-6643.

[15] J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin, M. Uchimaya, J. Comput. *Aided Mol. Des.*, 2007, **21**, 681-691.

[16] T. Kesvatera, B. Jonsson, E. Thulin, S. Linse, *J. Mol. Biol.*, 1996, **259**, 828-839.

[17] M. Tollinger, J. D. Forman-Kay and L. E. Kay, *J. Am. Chem. Soc.*, 2002, **124**, 5714-5717.

[18] T. Matsui, A. Oshiyama and Y. Shigeta, *Chem. Phys. Lett.,* 2011, **502**, 248 -252

[19] D. Riccardi, P. Schaefer and Q. Cui, *J. Phys. Chem. B*, 2005, **109**, 17715 -17733

[20] T. Matsui, Y. Shigeta and K. Morihashi, *J. Chem. Theory Comput.*, 2017, **13**, 4791 -4803

[21] A. Kovalenko and F. Hirata, *J. Chem. Phys.*, 1999, **110**, 10095-10112.

[22] H. Sato, A. Kovalenko and F. Hirata, *J. Chem. Phys.*, 2000, **112**, 9463-9468.

[23] A. Kovalenko and F. Hirata, *J. Mol. Liq.*, 2001, **90**, 215-224.

[24] S. Gusarov, T. Ziegler, A. Kovalenko, 2005, **108**, 207-208.

[25] A. Kovalenko and F. Hirata, *Chem. Phys. Lett.*, 2001, **349**, 496-502.

[26] J. Wang, W. Wang, P. A. Kollman and D. A. Case, *J Mol Graph Model,* 2006, **25**, 247-260.

[27] H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269-6271.

[28] S. Ten-No, F. Hirata and S. Kato, *J. Chem. Phys.*, 1994, **100**, 7443-7453.

[29] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.*, 1993, **14**, 1347-1363.

[30] N. Yoshida and F. Hirata, *J. Comput. Chem.*, 2006, **27**, 453-462.

[31] N. Yoshida, Y. Kiyota and F. Hirata, *J. Mol. Liq., 2011,* **159**, 83-92.

[32] N. Yoshida, *J. Chem. Phys.,* 2014, **140**, 214118.

[33] J.F.J. Dippy, S.R.C. Hughes, A. Rozanski, *J. Chem. Soc.* 1959, 2492-2498

[34] R.M.C. Dawson, et al., Data for Biochemical Research, Oxford, Clarendon Press, 1959

[35] J. March, et al, *Advanced Organic Chemistry,* 3rd Ed, 1985

[36] H.C. Brown, et al., in E.A. Braude, and F.C. Nachod, *Academic Press,* New York, 1955

[37] H.K. Hall Jr., *J. Am. Chem. Soc.* 1957, **79**, 5441-5444

[38] T.C. Bruice, G.L. Schmir, *J. Am. Chem. Soc.* 1958, **80**, 148-156

[39] M.M. Kreevoy, et al. *J. Am. Chem. Soc.* 1960, **82**, 4899-4902.

[40] J.T. Edsall, Wyman and Jeffries, Biophysical Chemistry, Academic Press, Inc., New York,

1958

[41] J.P. Danehy, Noel, C.J. *J. Am. Chem. Soc.* 1960, **82**, 2511-2515

[42] R.N. Haszeldine, *J. Chem. Soc.* 1953, 1748-1757

[43] P. Ballinger, F.A. Long, *J. Am. Chem. Soc.* 1960, **82**, 795-798

[44] P. Ballinger, F.A. Long, *J. Am. Chem. Soc.* 1959, **81**, 1050-1053

[45] V.E. Bower, R.A. Robinson, *J. Phys. Chem.* 1960, **64**, 1078-1079

[46] R. Williams, pK_a data compiled by R. Williams. [Online accessed on]. http://www.chem.wisc.edu/areas/organic/index-chem.htm. 2011. [14/01/2022].

[47] R. M. C. Dawson, D. C. Elliott, W. H. Elliott and K. M. Jones, *Data for biochemical research*, Clarendon Press Oxford, 1969.

[48] J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999–3093

3 Application of LFC/3D-RISM-SCF for quantitative prediction of protonation state in methanol

3.1 Introduction

The solvation effect of organic solvent is important in the fields of chemistry and physics. In organic synthesis, it affects the reaction rate and reaction pathway: For example, stabilization of the leaving group by solvation increases the reaction rate, and the kinetic stability and thermodynamic stability is changed by the solvation situation. In pharmacy, as many drugs are weakly acidic or basic, they are sensitive to equilibrium migration based on solvent species. Therefore, because the effects of solvation appear as changes in solubility and absorption rate, in vivo, studies of the solvation are essential for drug design [1].

Methanol is an amphiphilic protic molecule that resembles a water molecule at arbitrary concentration it has a hydrophobic methyl group. Thus, methanol is used as a solvent for molecules with both hydrophobicity and polarity. As is similarly the case with aqueous solutions, as mentioned in the previous chapter, it is difficult to estimate the free energy of excess protons in methanol. Earlier reports have shown that it varies greatly, depending on experimental or computational conditions. For example, Hwang et al. suggest that methanol in solution forms over pentamer and $G(\mathrm{H}^{+})$ is -263.4 $\mathrm{kcal\,mol^{-1}}$ under that situation [2]. However, Fifen et al. suggest values of -258 to -226 $\mathrm{kcal\,mol^{-1}}$ in some methanol n-mers. [3]

As described in the previous chapter, we have developed a theory that avoids calculation of the free energy of excess protons, which then allows us to obtain quantitative $\mathrm{p}K_{\mathrm{a}}$ prediction in aqueous solution. [4][9][10][11] Here, the parameters corresponding to the proton free energy are determined by data learning on experimental data. Therefore, this method is expected to be easily extended to methanol systems by using the training data based on the experimental data in a methanol system. In this chapter, we examine the LFC/3D-RISM-SCF scheme for methanol system. First, the new LFC parameters for a methanol are determined. Thereafter, the $\mathrm{p}K_\mathrm{a}$ values obtained from the data sets are compared with experimental data. The suitability of the LFC/3D-RISM-SCF scheme for a methanol system is then discussed.

3.2 Computational Details

In this study, the parameters for three chemical groups were determined, namely, amine, phenol, and carboxyl. Table 3.1.1 ~ 3.1.3 summarize the training data sets for parameter fitting [19]. The determined parameters are examined to test data sets in Table 3.1.4. The rational formulas of training data sets and test data sets are also in Supporting information (Table S1).

Prior to the Gibbs energy calculation, the structure optimization of protonated (HA) and deprotonated (A^-) states was performed at the B3LYP/6-31++G(d,p) level, in methanol, with the PCM, for all the training set molecules [5]. For the Gibbs energy calculation, two different sizes of basis sets were employed, 6-31++G(d,p) to examine the basis set dependency of the parameter fitting.

The parameters used in the 3D-RISM calculation [6-8] were the following: temperature 298.15 K, density of solvent methanol 0.79 $\mathrm{g\,cm^{-3}}$. The Lennard-Jones parameters for solute molecules were taken from the general Amber force field (GAFF) parameter set with antechamber software [12]. The extended simple point charge model (SPC/E) parameter set for the geometrical and potential parameters for the solvent methanol was employed with modified hydrogen parameters (σ = 1.0 Å, ε = 0.056 kcal $\mathrm{mol^{-1}}$) [13][14]. The grid spacing for the 3D grid was 0.5 Å and the number of grid points on each axis was 128.

All calculations were performed with a modified version of the GAMESS program package, for which the 3D-RISM-SCF program was implemented [15-18].

Table 3.1.1. Training data set of carboxyl group

Carboxyl	pK_a		pK_a
2,3-dichloropropanoic acid	7.50	2-phenylacetic acid	9.43
2,4-dichlorobenzoic acid	7.80	3-chlorobenzoic acid	8.83
2,4-dinitrobenzoic acid	6.45	3-cyanobenzoic acid	8.53
2,6-dinitrobenzoic acid	6.30	3-nitrobenzoic acid	8.32
2-bromoacetic acid	8.06	3-trifluoromethylbenzoic acid	8.69
2-bromobenzoic acid	8.19	4-bromobenzoic acid	8.93
2-chloroacetic acid	7.88	4-chlorobenzoic acid	9.09
2-chlorobenzoic acid	8.31	4-cyanobenzoic acid	8.42
2-cyanoacetic acid	7.50	4-fluorobenzoic acid	9.23
2-fluoroacetic acid	7.99	4-methylbenzoic acid	9.51
2-fluorobenzoic acid	8.41	4-nitrobenzoic acid	8.34
2-nitrobenzoic acid	7.64	propanoic acid	9.71

Table 3.1.2. Training data set of the amine group

Amine	pK_a		pK_a
2,4,6-trimethylpyridine	7.72	aniline	6.05
2-bromoaniline	3.46	hydroxylamine	6.29
2-chloroaniline	3.71	N-ethylamine	11.00
2-methylaniline	5.95	N-methylamine	11.00
2-nitroaniline	0.20	N,N-dimethylamine	11.20
4-benzylaniline	5.98	N,N,N-triethylamine	10.78
4-chloro-2-nitroaniline	-0.67	N,N,N-trimethylamine	9.80
4-chloroaniline	4.95	o-methylhydroxylamine	5.13
4-hydroxyaniline	7.41	piperidine	11.07
4-methoxyaniline	6.89	pyridine	5.44
4-methylaniline	6.57	quinoline	5.16
4-nitroaniline	1.55	ammonia	10.78

Table 3.1.3. Training data set of the phenol group

Phenol	pK_a		pK_a
1-naphtol	13.91	2-nitrophenol	11.53
2,4,6-trimethylphenol	15.53	2-*tert*-buthylphenol	16.50
2,4,6-trinitrophenol	3.55	3-bromophenol	13.30
2,4-dimethylphenol	15.04	3-chlorophenol	13.10
2,4-dinitrophenol	7.83	3-methylphenol	14.43
2,5-dinitrophenol	8.94	3-nitrophenol	12.41
2,6-dinitrophenol	7.64	4-bromophenol	13.63
2-chloro-4-phenylphenol	12.70	4-chlorophenol	13.59
2-chlorophenol	12.97	4-methylphenol	14.54
2-fluorophenol	12.94	4-nitrophenol	11.30
2-methoxyphenol	14.48	4-*tert*-buthylphenol	14.52
2-methylphenol	14.86	salicylaldehyde	12.82

Table 3.1.4. Test data set of three groups [20]

Phenol	pK_a	Carboxyl	pK_a
2_3_dimethylphenol	15.08	2_2_dichloroaceticacid	6.38
2_4_6_tribromophenol	10.1	2_cyanoaceticacid	7.5
2_5_dimethylphenol	14.91	2_sulfanylaceticacid	8.52
2_6_dimethylphenol	15.26	3_4_dinitrobenzoicacid	7.44
2_chloro_4_bromophenol	12.7	aceticacid	9.63
3_4_dimethylphenol	14.63	aspartame	7.68
3_5_dichlorophenol	12.11	benzoicacid	9.3
3_5_dimethylphenol	14.57	maronicacid	7.66
3_5_dinitrophenol	10.29		
4_hydroxybenzaldehyde	12.01		
phenol	14.1		
Amine	pK_a		
1_methylpiperidine	10.88		
2_amino_1_ethanol	6.06		
2_methylquinoline	4.42		
3_bromoaniline	5.99		
3_hydroxyaniline	6.04		
3_methoxyaniline	6.92		
4_ethoxyaniline	6.05		
4_methylpyridine	5.82		
n_ethyl_n_phenylamine	5.45		
n_methyl_n_phenylamine	10.88		

3.3 Results and Discussion

3.3.1 Parameter determination

Table 3.2 shows the parameters determined by the least squares fitting for methanol. The parameters of water solution taken from a previous study are shown in Table 3.3 for comparison. The results show a good correlation for the phenol and carboxyl groups. (r = 0.9334, RMSE = 0.6926 and r = 0.8052, RMSE = 0.3670) This suggests that our correction scheme is effective in nonaqueous solvent. The estimated free energy of protons of the two group are -267 and -262, which are similar to the values suggested in other studies (except for amines). The values in methanol are lower than in water for the phenol and carboxyl groups. For the amine group, although the RMSE value is slightly higher than values for carboxyl and phenol groups, the correlation shows good results. (r = 0.8022, RMSE = 1.5025) This suggests that high accuracy is achieved by assigning different parameters to each functional group in this scheme.

Table 3.2. Determined parameters of each chemical group and RMSE in methanol

	k (mol kcal^{-1})	C_0 (kcal mol^{-1})	RMSE	r	$G(\mathrm{H}^+)$ (kcal mol^{-1})
Phenol	0.383	−102.5	0.6926	0.9334	−267.4
Amine	0.333	−82.2	1.5025	0.8022	−246.6
Carboxyl	0.260	−68.4	0.3670	0.8052	−262.4

Table 3.3. Determined parameters of each chemical group and RMSE in water [4]

	k (mol kcal^{-1})	C_0 (kcal mol^{-1})	RMSE	r	$G(\mathrm{H}^+)$ (kcal mol^{-1})
Phenol	0.317	−78.7	0.423	0.931	−248.5
Amine	0.396	−98.6	0.469	0.973	−248.9
Carboxyl	0.319	−81.5	0.661	0.832	−255.3

3.3.2 Comparison with experimental $\mathrm{p}K_\mathrm{a}$

Figure 3.1 shows the curves of the correlation between the experimental and the computational $\mathrm{p}K_\mathrm{a}$ values obtained by the direct scheme, panel (a), and the LFC scheme, panel (b), respectively. In the direct scheme, the deprotonation reaction in methanol was considered as below.

$$\mathrm{HA} + \mathrm{CH_3OH} \rightleftharpoons \mathrm{A^-} + \mathrm{CH_3OH_2}^+$$

Further details of the direct scheme are explained in the previous chapter.

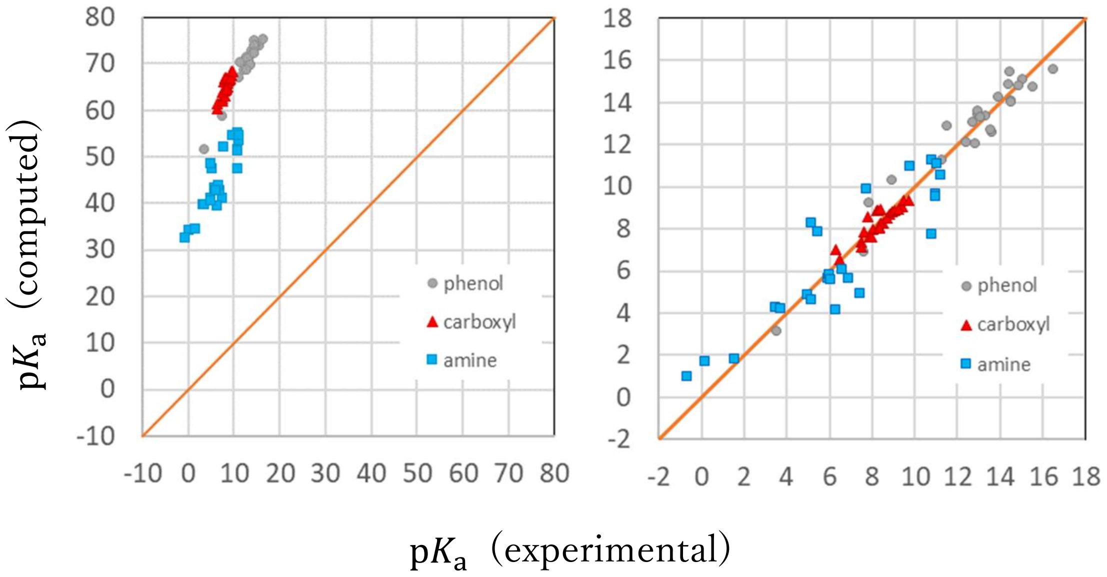

Figure 3.1. Comparison of the experimental $\mathrm{p}K_\mathrm{a}$ values with the computed values, determined by the direct scheme and the LFC/3D-RISM-SCF scheme. References for the experimental values are given in Table 3.1. Values for each of the chemical groups are presented in separate panels: (a) the direct scheme, (b) LFC/3D-RISM-SCF.

The result of each scheme seems to have good linear correlation; however, the LFC scheme has smaller errors than the direct scheme. This clearly indicates that correction of the LFC scheme is also effective for the prediction of $\mathrm{p}K_\mathrm{a}$ in methanol. As seen in Figure 3.1(a), the $\mathrm{p}K_\mathrm{a}$ values of the phenol and carboxyl groups are overestimated by about 55 $\mathrm{p}K_\mathrm{a}$ unit and those of the amine group by about 35 $\mathrm{p}K_\mathrm{a}$ unit. These differences may be attributed to differences in the charges of the molecules participating in the reaction. Namely, the molecules in the amine group have a positive charge in the protonated state whereas they are charge neutral in the deprotonated state. On the other hand, the molecules in phenol and carboxyl groups are charge neutral in the protonated state and have a negative charge in the deprotonated state. In the state with a net charge, strong hydrogen bonds form between the solute and solvents. In the case of the amine group, the oxygen of the hydroxyl group of the solvent forms a hydrogen bond with the excess proton of solute amine, whereas in the case of phenol and carboxyl groups, the hydrogen of the hydroxyl group of the solvent coordinates with the oxygen of the solute. The difference in the hydrogen bond form is thought to be reflected in the difference in the degree of overestimation. The LFC/3D-RISM-SCF method was shown to be able to handle such differences because of the molecular nature, as the parameters are determined for each functional group.

3.3.3 Application of determined parameters

For the evaluation of LFC/3D-RISM-SCF method in methanol, the relationship between calculated and experimental $\mathrm{p}K_\mathrm{a}$ values are examined. Figure 3.2 shows the relationship of test data set. Phenol, carboxyl, and amine are included. The errors of each group are corrected by fitting parameters and the correlation in each group is confirmed. The correlation of entire group is very small (R^2:0.288) however the correlations of each group show better values. (Phenol R^2:0.941, RMSE:2.324, Carboxyl R^2:0.310, RMSE:2.217, Amine R^2:0.421, RMSE:4.177.) The phenol group showed good agreement with experimental data. On the other hand, Carboxyl and Amine group showed relatively low correlation compared with the Phenol group.

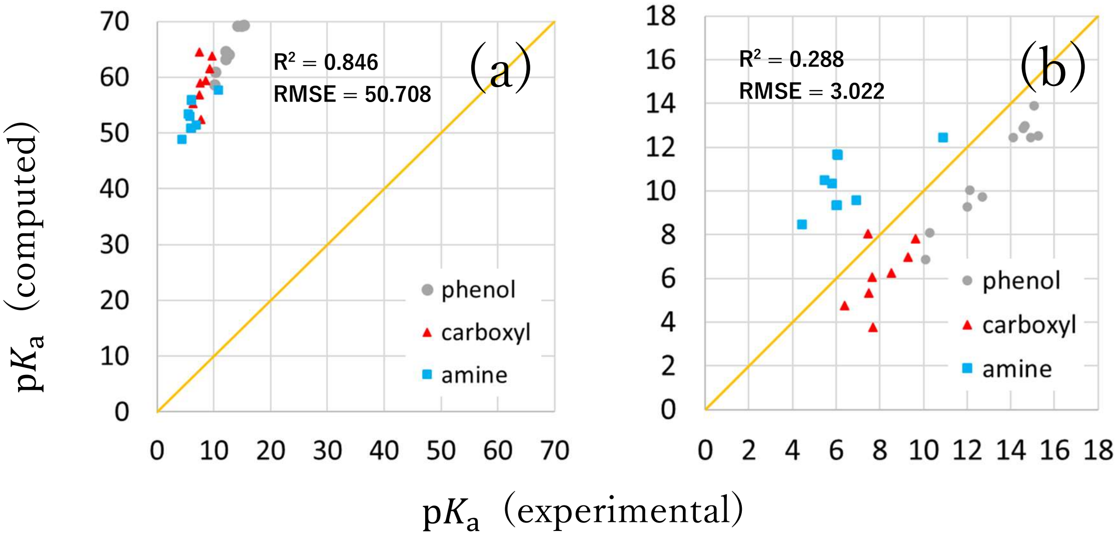

Figure 3.2. Comparison of the experimental $\mathrm{p}K_\mathrm{a}$ values with the computed values, determined by the direct scheme and the LFC/3D-RISM-SCF scheme. Test data sets are given in Table S2. Values for each of the chemical groups are presented in separate panels: (a) the direct scheme, (b) LFC/3D-RISM-SCF.

Especially, the Amine group shows very low correlation and not enough to correct only by fitting parameters. In the determination of fitting parameter of Amine group, as the relationship is already worse than other groups, this result of test data set indicates the probability of additional correction parameter or more detailed separation of amine group than current reference like aniline, quinoline, and pyridine.

3.4 Summary

In this chapter, the application of the LFC/3D-RISM-SCF method in methanol solution was examined. The three groups were considered: phenol, amine and carboxyl. The free energy in the solvent required for the determination of parameters in the LFC scheme was calculated by the 3D-RISM-SCF method. Each parameter for the prediction of $\mathrm{p}K_\mathrm{a}$ was determined by least squares fitting for the experimental values of the training set for each chemical group based on the LFC scheme and applied to some molecules. Calculations results showed good correlation for all groups considered and enough qualitative agreement with experiment. This suggests that LFC/3D-RISM-SCF is also useful in the prediction of $\mathrm{p}K_\mathrm{a}$ in methanol.

On the other hand, the results for the Amine group show larger error compared with the phenol and carboxyl groups. Future research is expected to identify the causes of the errors and improve the accuracy.

The results of this study indicate extension to other organic solvents and mixed solvents, by the LFC/3D-RISM-SCF method. 3D-RISM-SCF can easily handle multiple-component solvent systems, which are difficult to be handled by the continuum models. The method proposed as described in this thesis may be an effective $\mathrm{p}K_\mathrm{a}$ prediction tool in complex systems.

References

[1] T. N. Brown, N. M. Diez, *J. Phys. Chem. B*, 2006, **110**, 9270-9279.

[2] G. J. Tawa, I. A. Topol, S. K. Burt, R. A. Caldwell, A. A. Rashin, *J. Chem. Phys.* 1998, **109**, 4852-4863.

[3] J. J. Fifen, M. Nsangou, Z. Dhaouadi, O. Motapon, N. E. Jaidane, *J. Chem. Theory Comput.* 2013, **9**, 1173-1181

[4] R. Fujiki, Y. Kasai, Y. Seno, T. Matsui, Y. Shigeta, N. Yoshida, H. Nakano, *Phys. Chem. Chem. Phys.*, 2018, **20**, 27272.

[5] J. Tomasi, B. Mennucci and R. Cammi, *Chem. Rev.*, 2005, **105**, 2999-3093.

[6] D. Beglov, B. Roux, *J. Phys. Chem. B,* 1997, **101**, 7821-7826.

[7] D. Beglov, B. Roux, *J. Chem. Phys.,* 1996, **104**, 8678-8689.

[8] A. Kovalenko, F. Hirata, *Chem. Phys., Lett.* 1998, **290**, 237-244.

[9] S. Ten-No, F. Hirata, S. Kato, *Chem. Phys. Lett.,* 1993, **214**, 391-396.

[10] A. Kovalenko, F. Hirata, *J. Chem. Phys.,* 1999, **110**, 10095-10112.

[11] H. Sato, A. Kovalenko, F. Hirata, *J. Chem. Phys.,* 2000, **112**, 9463-9468.

[12] J. Wang, W. Wang, P. A. Kollman and D. A. Case, *J. Mol. Graph. Model.*, 2006, **25**, 247-260.

[13] H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269-6271.

[14] S. Ten-No, F. Hirata and S. Kato, *J. Chem. Phys.*, 1994, **100**, 7443-7453.

[15] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. J. Su, T. L. Windus, M. Dupuis and J. A. Montgomery, *J. Comput. Chem.*, 1993, **14**, 1347-1363.

[16] N. Yoshida, F. Hirata, *J. Comput. Chem.*, 2006, **27**, 453-462.

[17] N. Yoshida, Y. Kiyota, F. Hirata, *J. Mol. Liq.*, 2011, **159**, 83-92.

[18] N. Yoshida, *J. Chem. Phys.*, 2014, **140**, 214118.

[19] E. L. M. Miguel, P. L. Silva, J. R. Pliego, *J. Phys. Chem. B*, 2014, **118**, 5730-5739

[20] F. Rived, M. Rosés, E. Bosch, *ANAL. CHIM. ACTA.,* 1998, **374**, 309-324

4 Development of sampling method of protonation states of polymer at constant pH based on the molecular dynamics simulation and integral equation theory of liquids

4.1 Introduction

The protonation state of dissociative functional group of biopolymers changes depending on the pH value of the solvent environment [1]. The protonation states of amino acid residues are related to ligand binding affinity and its binding conformations. The change of protonation state also affects the 3D conformation of biopolymers. All these processes attract special attention in the field of biochemistry, biophysics, and pharmacology because they are related to the function and structure of biopolymers [2][3].

The equilibrium of the deprotonation reaction of dissociative residues in a protein shifts is depending on the conformation of the surrounding amino acids, hydration structure, ion distribution, and pH. The thermodynamic quantities in biological processes involving proteins should be evaluated by an ensemble according to the probability of finding the protonated and deprotonated states of the contained amino acids corresponding to the shifted equilibrium of the deprotonation reaction. Therefore, for structural sampling of proteins containing dissociative amino acids, it is not sufficient to use a fixed protonation state estimated from K_a: rather, it is necessary to take into account the frequency of appearance of the protonation state according to the shifted K_a.

To realize such sampling of protonation states, Mongan et al. proposed the simulation method referred to as the constant-pH molecular dynamics (CpHMD) method [12]. In this method, the protonation states of target dissociable residues are sampled based on Monte Carlo (MC) trials with pH conditions. Several CpHMD related techniques have subsequently been proposed [27][28][33-35]. In the original CpHMD method, to take solvent effects into account, the generalized Born (GB) model was employed. The method has been successfully applied to various systems. However, it is known that the GB model is unable to reproduce the solvent environment in the clefts or pores of proteins where become candidates of ligand binding targets because of high anisotropy of those parts. For example, while Hen-Egg White Lysozyme (HEWL) has various dissociative Asp residues, inner one, Asp66, cannot be reproduced by GB model. Therefore, the development of a method with the solvation model, which can handle the solvation thermodynamics for fully anisotropic systems, is required.

In this study, we propose a combined method of the CpHMD with the three-dimensional reference site model (3D-RISM) theory. The 3D-RISM theory allows us to evaluate the solvation structure of even complex biomolecular systems [13][15][16]. In this chapter, a sampling method of protonation states of dissociable residues in proteins is proposed; it is referred to as CpHMD/3D-RISM.

Section 4.2 provides an introduction to various methods. The 3D-RISM theory is described in some detail in Section 2.2.2. Section 4.3 gives computational details. Results are presented and discussed in Section 4.4. Section 4.4.1 reports the relationship between sampling time and $\mathrm{p}K_\mathrm{a}$. Section 4.4.2 addresses the convergence of $\mathrm{p}K_\mathrm{a}$ value. Section 4.4.3 describes the application of the method to selected tripeptides, in order to assess the transferability. Section 4.4.4 addresses the distribution of solvent around an Asp-containing peptide as the advantage of 3D-RISM. A summary of this chapter is presented in Section 4.5.

4.2 Methods

4.2.1 Constant pH Molecular Dynamics / 3D-RISM method

The original CpHMD algorithm was proposed by Mongan et al. [12] It is the combination method of MD simulation with periodic or MC sampling of protonation states. In their method, the GB implicit solvent model is employed as a solvation model for both the structure sampling and Metropolis MC for the sake of computational efficiency. In the present method, the 3D-RISM theory is employed for the Metropolis MC criteria, whereas the GB model is used for the structure sampling. Figure 4.1 shows a flowchart of the computational scheme. First, the structure of the target solute molecule is sampled with standard MD simulations, with the GB model. Change in the protonation state of dissociable residues is examined at regular intervals, by a stochastic Metropolis MC algorithm with the 3D-RISM theory [17]. For applying the Metropolis criterion to a given pH value, the transition free energy for the deprotonation reaction (ΔG) is defined by Eq. (1)

$$\Delta G = \mathrm{k_B} T(\mathrm{p}K_{\mathrm{a,ref}} - \mathrm{pH})\ln 10 + \Delta G_{\mathrm{elec}} - \Delta G_{\mathrm{elec,ref}} \tag{1}$$

where, k_{B} is the Boltzmann constant, T is a temperature, and $\mathrm{p}K_{\mathrm{a,ref}}$ is the experimental $\mathrm{p}K_{\mathrm{a}}$ value of a model compound in aqueous solution. The applied $\mathrm{p}K_{\mathrm{a,ref}}$ values are given in Table 4.1. ΔG_{elec} and $\Delta G_{\mathrm{elec,ref}}$ are the electrostatic components of the free energy differences between the protonated and deprotonated states of the dissociable residue in the protein and of the model compound in aqueous solution, respectively. Usually a simple monomeric amino acid with capped N-terminal acetyl and C-terminal amide groups is selected as a model compound. Eq. (1) includes of electrostatic and nonelectrostatic components [12]. According to Mongan et al. [12], the nonelectrostatic components, including the free energy of the bond between the proton and the dissociable residue, and the proton solvation free energy are assumed to be canceled out between the dissociable residue in the protein and in aqueous solution. The electrostatic

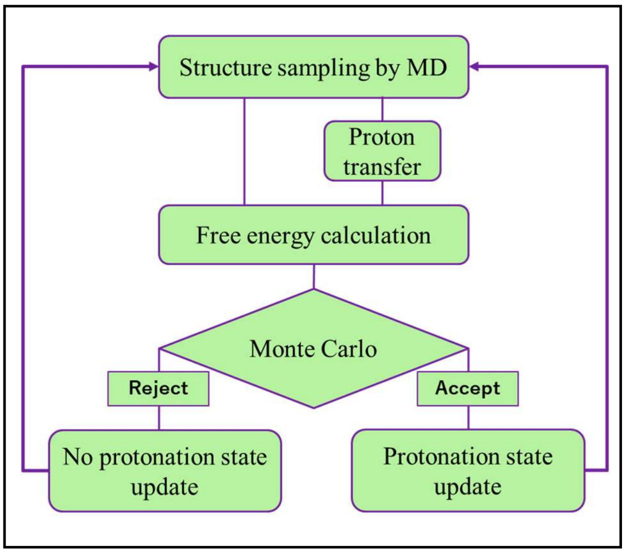

Figure 4.1 Flow chart of CpHMD/3D-RISM scheme.

components of the free energy ΔG_{elec} is the difference in the electrostatic energy calculated with the charges of the current state and the charges of the proposed state, which is given by

$$\Delta G_{\mathrm{elec}} = \Delta E_{\mathrm{Ptot}} + \Delta G_{\mathrm{sol}} \tag{2}$$

where ΔE_{Ptot} is the potential energy of the structure and ΔG_{sol} is the solvation free energy calculated by the 3D-RISM theory.

During updating of the charge set, van der Waals radii are assumed to be unchanged.

Table 4.1. The $\mathrm{p}K_{\mathrm{a,ref}}$ and $\Delta G_{\mathrm{elec,ref}}$ values of dissociable residues

Dissociable residue	$\mathrm{p}K_{\mathrm{a,ref}}$[18][19]	$\Delta G_{\mathrm{elec,ref}}$ (kcal mol^{-1})
Asp	4.0	47.9717
His(δ)	6.5	−22.4113
His(ε)	7.1	−19.1692
Lys	10.2	−2.0680

The ratio of the protonated and deprotonated states of a dissociable residue at a given pH is $\exp(-\beta\Delta G)$ in the grand canonical ensemble. Here, $\beta = 1/k_{\mathrm{B}}T$, where k_{B} and T are the Boltzmann constant and absolute temperature, respectively. The Metropolis scheme for updating of the protonation state is carried out based on this ratio. The transition probability from the deprotonated state to the protonated state, $w(\mathrm{d} \rightarrow \mathrm{p})$, is defined by Eq.(3).

$$w(\mathrm{d} \rightarrow \mathrm{p}) = \begin{cases} 1 & \text{if } (\Delta G \leq 0) \\ \exp(-\beta\Delta G) & \text{if } (\Delta G > 0) \end{cases} \tag{3}$$

If the transition is accepted, then the MD simulations are continued with the new protonation state, and if it is rejected, MD simulations are continued without changing the protonation state.

4.3 Computational Details

MD simulation was performed using AMBER 20 [32]. The ff14SB force field was employed [20]. 3D-RISM theory was applied only for the solvation free energy calculation for MC trials, whereas the GB solvation model was employed for MD structure sampling [14] [21] [22]. Salt concentration was set at 0.1 M. Solute temperature was coupled to a Berendsen thermostat at 300 K (time constant of 2 ps) [23]. SHAKE method was used to constrain the bond length, including hydrogen. The time step was 2 fs.

The following parameters were used in the 3D-RISM calculation: temperature 298.15 K, density of solvent water 1.0 $\mathrm{g\,cm^{-3}}$. The Lennard-Jones parameters for solute molecules were taken from the general Amber force field parameter set assigned by antechamber software [24]. The extended simple point charge model parameter set for the geometrical and potential parameters for the solvent water was employed with modified hydrogen parameters (σ = 1.0 Å, and ε = 0.056 $\mathrm{kcal\,mol^{-1}}$) [25][26]. The grid spacing for the 3D grid was 0.5 Å and the number of grid points on each axis was 64.

For constant pH MD simulations, the free energy differences $\Delta G_{\mathrm{elec,ref}}$ in Eq. (1) of reference compounds were determined (see Table 4.1). As the reference compounds, the following amino acids, with their N-termini and C-termini capped by acetyl groups and N-methyl groups, respectively, were used: Asp, Glu, His, and Lys. The free energy differences $\Delta G_{\mathrm{elec,ref}}$ were determined based on the procedure of Mongan et al. [12] In an iterative fashion, $\Delta G_{\mathrm{elec,ref}}$ was adjusted until the populations of the protonated and deprotonated states obtained at pH = $\mathrm{p}K_{\mathrm{a,ref}}$ were equal.

To determine the $\Delta G_{\mathrm{elec,ref}}$ for each group, molecule with only one titrative functional group were considered for the calculations. Thus, the values of $\Delta G_{\mathrm{elec,ref}}$ obtained through the constant pH simulations should coincide with the results obtained from thermodynamic integration, as in Eq. (4)

$$\Delta G_{\mathrm{TI}} = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{4}$$

where λ is the coupling parameter between the protonated and deprotonated states of the dissociable residue, and V is the electrostatic potential energy.

4.4 Results and Discussion

4.4.1 Assessment of reference system

For the assessment of a reference system for the CpHMD/3D-RISM method, we examined the titration of the Asp reference system as an example. For this system, the deprotonation fraction was calculated by varying the pH in the range 2.0-6.0. The titration curve was obtained. The simulation was performed at 300 K, and carried out 1.0 ns after an equilibration of 2.0 ns. The $\mathrm{p}K_\mathrm{a}$ values were determined based on the scheme of Mongan et al. [12] The relationship of $\mathrm{p}K_\mathrm{a}$ and pH is given by

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} - n\log_{10}\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \tag{5}$$

where $[\mathrm{A}^-]$ and $[\mathrm{HA}]$ are the numbers of the deprotonated and protonated states respectively, and n is the Hill coefficient. From Eq. (5), it is rewritten about the ratio of the deprotonated fraction f_d and pH,

$$f_\mathrm{d} = \frac{1}{1+10^{n_H(\mathrm{p}K_\mathrm{a}-\mathrm{pH})}} = \frac{[\mathrm{A}^-]}{[\mathrm{HA}]+[\mathrm{A}^-]} \tag{6}$$

where $n_H = 1/n$.

Figure 4.1 shows the titration curve of the Asp reference system. Based on the data sets of the pH condition and f_d, n_H and $\mathrm{p}K_\mathrm{a}$ is determined by fitting to Eq. (6). The result is in good agreement with the titration curve reported by Mongan et al. and Itoh et al. [12] [28] The determined titration curve indicated that the CpHMD/3D-RISM method can reproduce the K_a with quantitative accuracy.

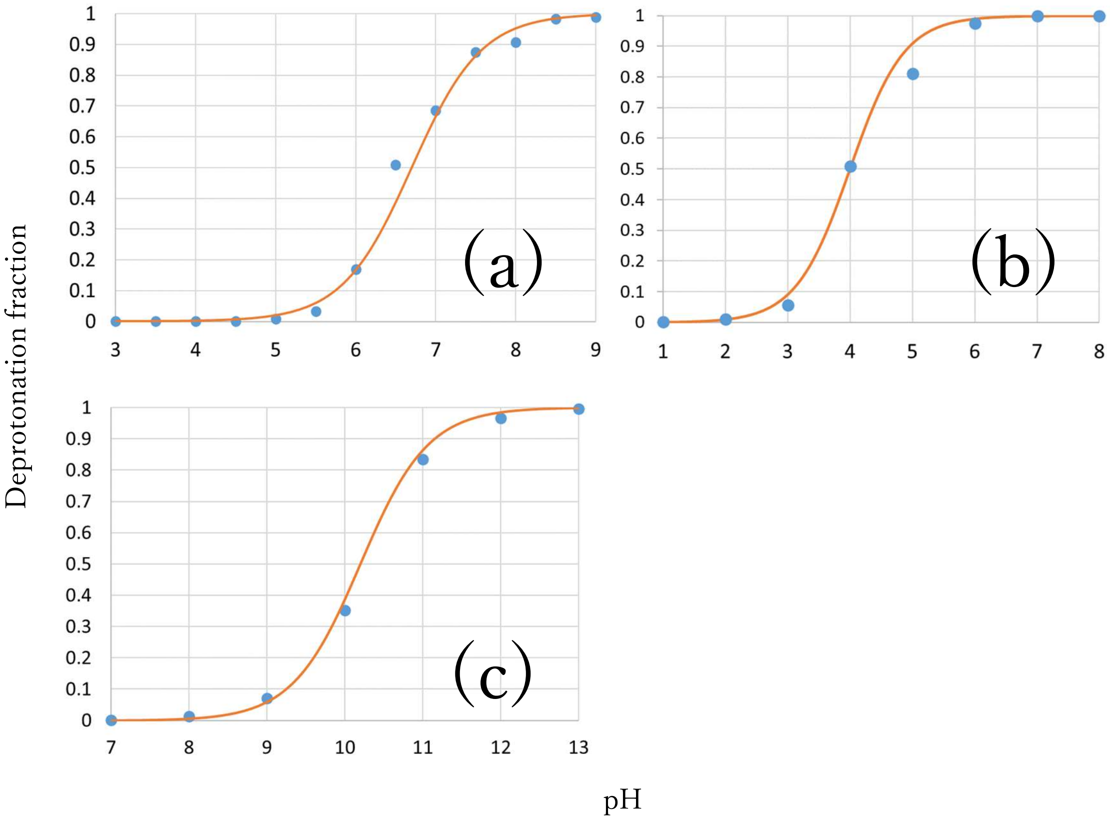

Figure 4.1 Calculated titration curve of (a) His, (b) Asp, (c) Lys reference system obtained with 1-ns simulation. The solid line is based on the Henderson-Hasselbalch equation with each $\mathrm{p}K_\mathrm{a}$ values.

4.4.2 Convergence of deprotonation fraction

The convergence of the deprotonation ratio is depicted in Figure 4.2. The vertical axis is the calculated $\mathrm{p}K_\mathrm{a}$ and the horizontal axis is the simulation time. The $\mathrm{p}K_\mathrm{a}$ value is calculated from Eq. (7),

$$\mathrm{p}K_\mathrm{a} = \mathrm{pH} - \log_{10} \frac{f_\mathrm{d}(t)}{1 - f_\mathrm{d}(t)} \tag{7}$$

where $f_\mathrm{d}(t)$ is deprotonation fraction at simulation step t.

As the pH value approaches 4.0 ($\mathrm{p}K_\mathrm{a}$), the fraction is converged faster. Although the total simulation time was 1 ns, it took about 0.4 ns. On the other hand, it may be necessary to apply a longer simulation for high pH conditions because the variation of the calculated $\mathrm{p}K_\mathrm{a}$ in that situation becomes large. In this calculation, the solute molecule is small, hence structural change is also small, and the conversion is almost the same for each pH condition. The tendency of pH = 4.0 was shifted to lower $\mathrm{p}K_\mathrm{a}$. This may be caused by trapped in unstable state. Itoh et al. reported that the *anti* conformation of a carboxyl proton is more unstable than the *syn* conformation and showed the same tendency of $\mathrm{p}K_\mathrm{a}$ shift.

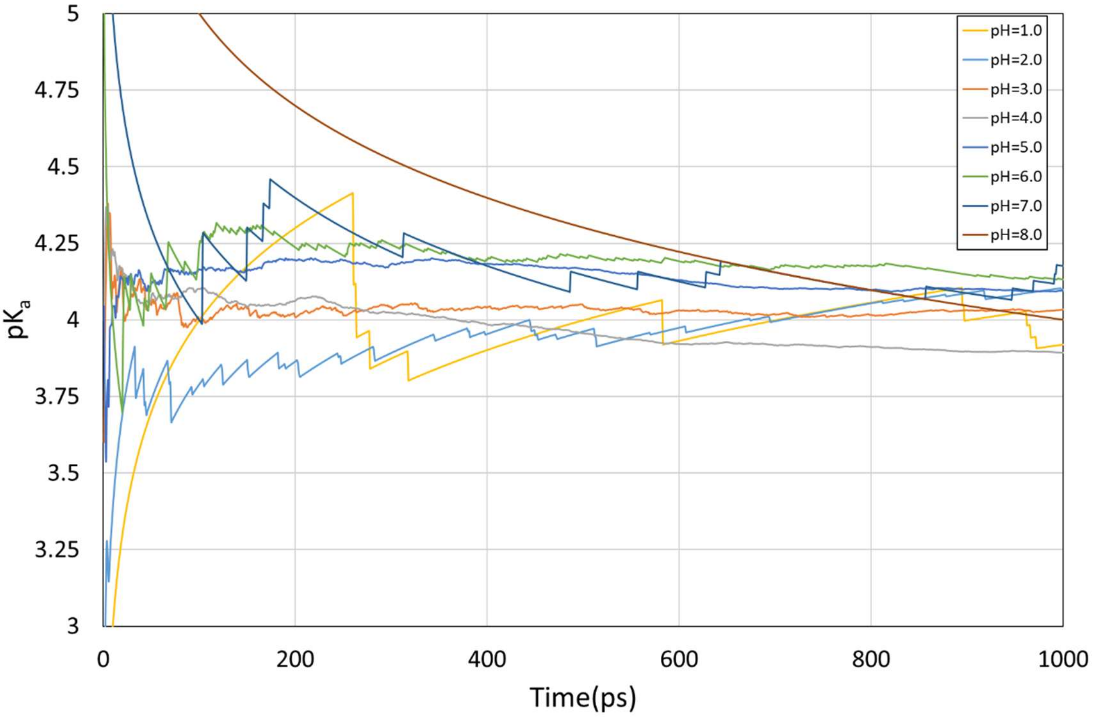

Figure 4.2 The pK_a convergence of Asp reference molecule. The pK_a values were calculated by Eq. (7)

4.4.3 Application of CpHMD/3D-RISM method to polypeptide

As an application of the CpHMD/3D-RISM method, simulations of polypeptides, including three types of dissociative residues (Asp, His and Lys) were carried out. The titration curves of each polypeptide obtained by the CpHMD/3D-RISM method are depicted in Figure 4.3. Table 4.2 gives the computed and experimental $\mathrm{p}K_\mathrm{a}$ and n_H values and comparison with the CpHMD of Mongan et al. [12] The computed values were determined by least squares fitting.

Table 4.2 List of polypeptides with $\mathrm{p}K_\mathrm{a}$, n_H, and RMSE values [29][30][31]

Model peptides	n_H	$\mathrm{p}K_\mathrm{a}$(CpHMD /3D-RISM)	n_H	$\mathrm{p}K_\mathrm{a}$(GB)	$\mathrm{p}K_\mathrm{a}$(exp.)
Ala-His-Lys	0.618	6.66(+0.56)	0.561	6.24(−0.6)	6.1
Ac-Gly-Asp-Gly-Gly-Me	0.88	4.20(+0.14)	1.011	3.80(−0.3)	4.06
Gly-Gly-Lys-Ala	0.831	10.27(−0.9)	1.03	10.54(−0.6)	11.1

The computed values by the CpHMD/3D-RISM method showed good agreement with the results of GB and experimental data. n_H values are various to each peptide. If $n_\mathrm{H} = 1$, the titration curve is the same as each reference compound and, as long as the titrating residues are considered to have little interaction with the surrounding environment, the titration curves of the system, including the Asp residue, follow this curve. The deprotonated fraction values are slightly different from the titration curve, as caused by the parameter $\Delta G_\mathrm{elec,ref}$. In the His-containing peptide, n_H is the smallest (of the three) the three and it is insensitive to pH change. This indicates that deprotonation of the His-containing peptide is slower than that of the other two peptides. In addition, the deprotonation tendency of the Asp-containing peptide and the Lys-containing peptide is similar to that of each reference compound.

Although the His-containing peptide also contains Lys residues, the result of titration shows complete protonation. For appropriate titration of N-termini or C-termini, a special $\Delta G_{\mathrm{elec,ref}}$ is necessary because the distribution of partial charges differs from that of inner residues [28].

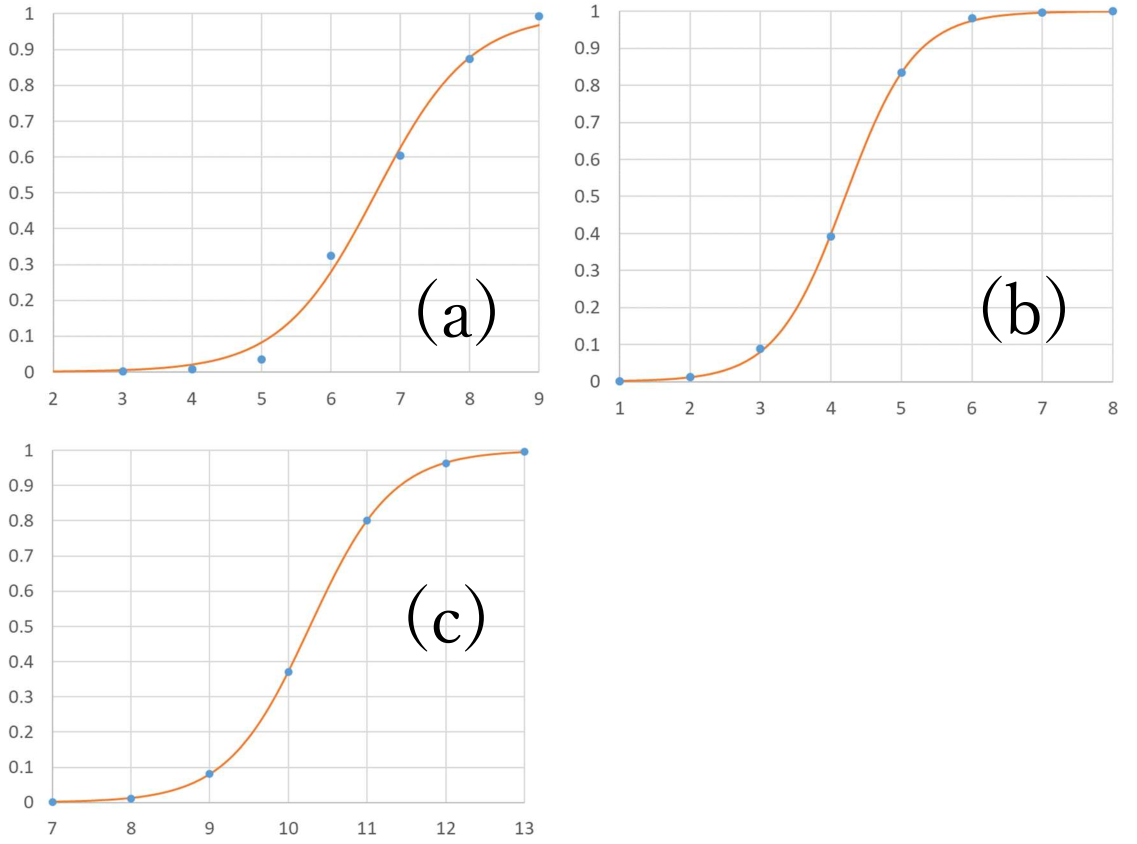

Figure 4.3 Titration curves based on CpHMD/3D-RISM calculations: (a) His peptide, (b) Asp peptide, (c) Lys peptide.

4.4.4 Solvation structure of Lys-containing peptide

The radial distribution of solvent molecules around a solute was obtained by the 3D-RISM theory. The radial distribution is evaluated by the orientational averaging respect to the specific solute site or atom of the three-dimensional spatial distribution of water obtained by the 3D-RISM theory. The positions of hydrogen and nitrogen of dissociative amine residue are chosen as averaging centers in Figure 4.4(a)(c) and Figure 4.4(b)(d), respectively. The Figures 4.4(a)(b) and (c)(d) are for deprotonation and protonation states, respectively. It is noted that Figure 4.4.(a) is centered at the same position as the hydrogen in 4.4.(b). The solvation structure changed, from Figure 4.4(c) to Figure 4.4(a) with the dissociation of protons. Figure 4.4(a)(b) show the conspicuous peaks of hydrogen at about 1 Å and 1.5 Å. Because the proton of Figure 4.4(a) is already dissociated, the value is existed in 0 Å and high peak of hydrogen is showed. This peak is caused by the hydrogen bond between nitrogen of amine and solvent hydrogen. In Figure 4.4(b), as the distribution of hydrogen atoms is dense around 1.5 Å, it is considered that hydrogen bonds are formed between the nitrogen of amine and the hydrogen of solvent molecules. On the other hand, as seen in Figure 4.4(c) and (d), because the dissociative residue protons remain, the local distribution of oxygen could be confirmed in each figure.

Figure 4.4(c) and 4.4(d) show the distribution of water molecules in the deprotonated condition. Regarding the solvent distribution around the dissociative residue, because the distribution of hydrogen atoms becomes strong (in Figure 4.4(d)), it is considered that hydrogen bonds are formed between nitrogen of amine and the oxygen of solvent molecules. However, the solvent distribution around hydrogen such as in Figures 4.4(c) did not show sharp peak. A possible reason is that the distribution of oxygen is decentral because amine has three equivalent hydrogens.

When focusing on the solvent distribution of the entire molecule, the distribution of solvent is almost same with that of Lys reference molecule even though the scale of molecule is larger than reference molecule. This indicates that the environment around the dissociative residue in Lys-containing peptide is similar to that of reference molecule and the difference of each pK_{a} value may be mainly contributed by the factor except solute-solvent interaction like structural fluctuation. The result of reference molecule is showed in supporting information (Figure S1).

The 3D-RISM theory can provide the information on partial solvent structure, which is otherwise difficult in continuum models. Investigations into solvation structure are considered important, particularly for predicting the reaction of proteins in solvents.

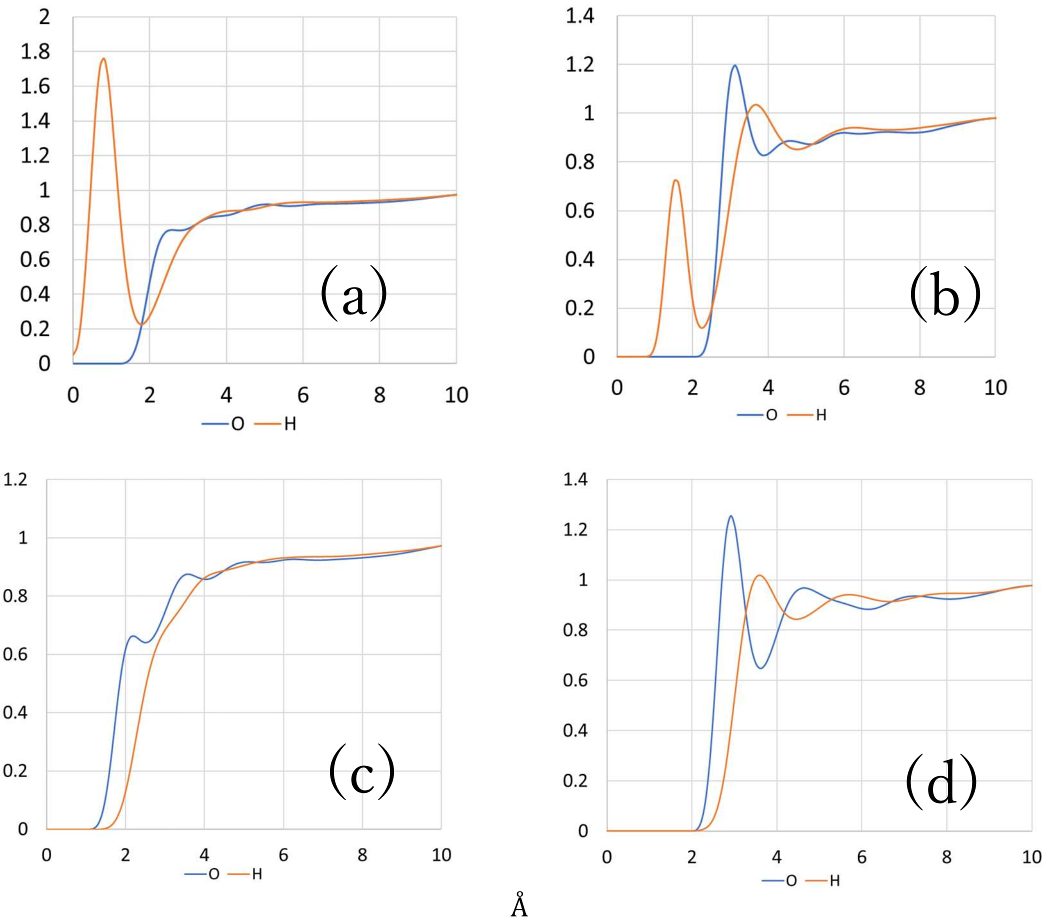

Figure 4.4 Radial distribution of oxygen and hydrogen of water molecule around (a), (c) dissociative proton and (b), (d) nitrogen of amine in Lys-containing peptide determined by 3D-RISM calculation. Figure (a), (b) are deprotonated and (c), (d) are protonated condition.

4.5 Summary

A combination method of CpHMD and 3D-RISM, referred to as CpHMD/3D-RISM, is proposed. It is expected that, therewith, a description of the solvation effect will become more detailed than with the GB solvent model, by employing the 3D-RISM theory. The method was applied to an Asp reference molecule and three polypeptides, including Asp, His and Lys. The Hill coefficient (n_{H}) and $\mathrm{p}K_{\mathrm{a}}$ value were determined for investigation of the behavior of CpHMD/3D-RISM simulation. The results were in good agreement with experimental results, thus indicating its applicability. In each reference molecule, quantitative results and an appropriate depiction of the protonation state were obtained. Application to three types of polypeptides also showed good agreement with experimental results and other methods. The computed $\mathrm{p}K_{\mathrm{a}}$ values were in reasonable agreement with experimental results, and thus the usefulness of CpHMD/3D-RISM was confirmed. Furthermore, adoption of the 3D-RISM method provided information on the solvation structure that is difficult to treat with continuum models. As the method can take mixed solvent systems into account, it is expected that it would be applicable to complex systems, such as chemical reactions in vivo. For example, Asp66 in HEWL which is buried inside protein has unexpectedly low $\mathrm{p}K_{\mathrm{a}}$ value compared with the isolated Asp. The original CpHMD simulation employing the continuum solvent model may not be appropriate to be applied to such systems.

In future work, systems containing polymer-like proteins or DNA that require long simulation times to obtain appropriate statistical averages because of structural fluctuation, should be given attention. The replica exchange method (REM) is an effective method for efficient sampling [28]. The combinational method of CpHMD/3D-RISM with REM may be a possible approach for further improvement in investigating protonation states.

References

[1] B. G. Moreno., *J. Biol.*, 2009, **8**, 98.

[2] D. A. Karp, A. G. Gittis, R. Gittis, E. E. Lattman, B. G. Moreno, *Biophys. J.,* 2004, **86**, 86–87.

[3] A. Damjanović, X. Wu, B. G. Moreno, B. R. Brooks, *Biophys. J.,* 2008, **95**, 4091–4101.

[4] C. Tanford, J. G. Kirkwood, *J. Am. Chem. Soc.*, 1957, **79**, 5333–5339.

[5] D. Bashford, M. Karplus, *Biochemistry*, 1990, **29**, 10219–10225.

[6] A. Nicholls, B. Honig, *J. Comput. Chem.* 1991, **12**, 435–445.

[7] J. Antosiewicz, J. A. McCammon,M. K. Gilson, *J. Mol. Biol.,* 1994, **238**, 415–436.

[8] J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosiewicz, M. K. Gilson, B. Bagheri, L. R. Scott, J. A. McCammon, *Comput. Phys. Commun.,* 1995, **91**, 57–95.

[9] T. Simonson, J. Carlsson, D. A. Case, J. Am. Chem. Soc., 2004, **126**, 4167–4180.

[10] N. Ghosh, Q. Cui, *J. Phys. Chem. B,* 2008, **112**, 8387–8397.

[11] L. Zheng, M. Y. Chen, W. Yang, *Proc. Natl. Acad. Sci. USA,* 2008, **105**, 20227–20232.

[12] J. Mongan, D. A. Case, J. A. McCammon, *J. Comput. Chem.*, 2004, **25**, 2038– 2048.

[13] N. Yoshida, S. Phongphanphanee, Y. Maruyama, T. Imai, F. Hirata, *J. Am. Chem. Soc.*, 2006, **128**, 12042-12043

[14] A. Onufriev, D. A. Case, D. J. Bashford, *J. Comput. Chem.,* 2002, **23**, 1297.

[15] A. Kovalenko, F. Hirata, *J. Chem. Phys.*, 1999, **110**, 10095-10112.

[16] H. Sato, A. Kovalenko, F. Hirata, *J. Chem. Phys.*, 2000, **112**, 9463-9468.

[17] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.,* 1953, **21**, 1087–1092.

[18] Y. Nozaki, C. Tanford, *Enzymol.*, 1967, **11**, 715–734.

[19] J. Kyte, Structure in protein chemistry. New York: Garland Publishing, Inc., 1995.

[20] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C., *J. Chem. Theory Comput.,* 2015, **11**, 3696–3713.

[21] A. Onufriev, D. Bashford,D. A. Case, *J. Phys. Chem. B*, 2000, **104**, 3712.

[22] A. Onufriev, D. Bashford,D. A. Case, *Proteins,* 2004, **55**, 383.

[23] H. J. C. Berendsen, J. P. M. Postma, W. F. V. Gunsteren, A. DiNola, J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684.

[24] J. Wang, W. Wang, P. A. Kollman, D. A. Case, *J. Mol. Graph. Model.*, 2006, **25**, 247-260.

[25] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *J. Phys. Chem.*, 1987, **91**, 6269-6271.

[26] S. Ten-No, F. Hirata, S. Kato, *J. Chem. Phys.*, 1994, **100**, 7443-7453.

[27] A. M. Baptista, V. H. Teixeira, C. M. Soares, *J. Chem. Phys.* 2002, **117** 4184–4200.

[28] S. G. Itoh, A. Damjanović, B. R. Brooks, *Proteins*, 2011, **79**, 3420-3436.

[29] M. Ohe, A. Kajita, *Biochemictry*, 1980, **19**, 4443-4450

[30] A. Bundi, K. Wüthrich, *Biochemistry*, 1979, **18**, 285-297

[31] S. Capasso, *Thermochimica acta*, 1996, **286**, 41-50

[32] D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz,R Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F. Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, San Francisco.

[33] M. Dlgosz, J. M. Antosiewicz, *Chem. Phys.*, 2004, **302**, 161-170

[34] J. E. Mertz, B. M. Pettitt, *Int. J. Supercomput. Appl. High Perform. Comput.,* 1994, **8**, 47-53

[35] U. Börjesson, P. H. Hünenberger., *J. Chem. Phys.,* 2001, **114,** 9706–9719

5 General Conclusions

In this thesis, theoretical methods have been proposed to investigate the protonation state change of molecules in a solution based on the statistical mechanics theory of a molecular liquids combined with quantum chemical theory, molecular dynamics simulation method, and informatics techniques.

In Chapter 2, the $\mathrm{p}K_\mathrm{a}$ prediction method, LFC/3D-RISM-SCF, was proposed and applied to dissociative amino acids, with fitted parameters determined by linear correlation for the training sets of molecules. The prediction of $\mathrm{p}K_\mathrm{a}$ was successfully achieved at quantitative level and the transferability of the scheme was demonstrated. The basis set dependency was also investigated. Results showed that quantitative predictions are possible even using smaller basis functions. These advantages demonstrate the applicability of this method to macromolecular systems such as biomolecules. The LFC/3D-RISM-SCF showed better accuracy than a previously proposed LFC/PCM scheme.

In Chapter 3, the LFC/3D-RISM-SCF scheme was applied to methanol solution systems. This scheme also showed good accuracy at a quantitative level. Although the results obtained for the Lys group exhibited a slightly larger error than for the other functional groups (Asp and His), the quantitative accuracy was adequate. These results indicate consideration of the extension of the scheme to other organic solvents and mixed solvent.

In Chapter 4, the simulation method, CpHMD/3D-RISM was proposed for the sampling of protonation/deprotonation equilibrium at given pH. This method was successfully applied to the Asp reference molecule; results indicated the applicability of the method. This method was also applied to three types of polypeptides containing Asp, His and Lys, and gave comparable results to the experimental observations. The solvation structure change of the peptide depending on the protonation states is discussed.

In contrast to conventional methods, the methods proposed in this thesis for investigating protonation states employ the statistical mechanics theory of molecular liquids, namely the 3D-RISM theory, for describing solvent effects. Compared to the conventional methods based on a continuum model, it is expected that higher accuracy can be achieved with the present method by employing 3D-RISM. Furthermore, by utilizing the advantages of 3D-RISM, such as the ability to handle mixed solvents and to describe solvation structures in highly anisotropic regions, it is expected that analyses that have been challenging when using conventional methods will now become possible in the future.

It is expected that the exploration and accumulation of knowledge based on molecular theory will not only promote developments in the fields of material development and chemical technology, but also lead to innovative discoveries in pharmacy and medical fields.

Supporting information

Table S1 Rational formula and experimental values of data set for LFC/3D-RISM-SCF in methanol

Phenol	Rational formula	pK_a
1-naphtol	$C_{10}H_7OH$	13.91
2,4,6-trimethylphenol	$C_6H_2(CH_3)_3OH$	15.53
2,4,6-trinitrophenol	$C_6H_2(NO_2)_3OH$	3.55
2,4-dimethylphenol	$C_6H_3(CH_3)_2OH$	15.04
2,4-dinitrophenol	$C_6H_3(NO_2)_2OH$	7.83
2,5-dinitrophenol	$C_6H_3(NO_2)_2OH$	8.94
2,6-dinitrophenol	$C_6H_3(NO_2)_2OH$	7.64
2-chloro-4-phenylphenol	$C_6H_3ClNO_2OH$	12.7
2-chlorophenol	C_6H_4ClOH	12.97
2-fluorophenol	C_6H_4FOH	12.94
2-methoxyphenol	$C_6H_4OCH_3OH$	14.48
2-methylphenol	$C_6H_4CH_3OH$	14.86

Phenol	Rational formula	pK_a
2-nitrophenol	$C_6H_4NO_2OH$	11.53
2-*tart*-buthylphenol	$C_6H_4C(CH_3)_3OH$	16.5
3-bromophenol	C_6H_4BrOH	13.3
3-chlorophenol	C_6H_4ClOH	13.1
3-methylphenol	$C_6H_4CH_3OH$	14.43
3-nitrophenol	$C_6H_4NO_2OH$	12.41
4-bromophenol	C_6H_4BrOH	13.63
4-chlorophenol	C_6H_4ClOH	13.59
4-methylphenol	$C_6H_4CH_3OH$	14.54
4-nitrophenol	$C_6H_4NO_2OH$	11.3
4-*tart*-buthylphenol	$C_6H_4C(CH_3)_3OH$	14.52
salicylaldehyde	C_6H_5CHOOH	12.82

Carboxyl	Rational formula	pK_a
2,3-dichloropropanoic acid	$CH_2Cl\text{-}CHCl\text{-}COOH$	7.5
2,4-dichlorobenzoic acid	$C_6H_3Cl_2COOH$	7.8
2,4-dinitrobenzoic acid	$C_6H_3(NO_2)_2COOH$	6.45
2,6-dinitrobenzoic acid	$C_6H_3(NO_2)_2COOH$	6.3
2-bromoacetic acid	$CH_2BrCOOH$	8.06
2-bromobenzoic acid	$C_6H_4BrCOOH$	8.19
2-chloroacetic acid	$CH_2ClCOOH$	7.88
2-chlorobenzoic acid	$C_6H_4ClCOOH$	8.31
2-cyanoacetic acid	$CH_2CNCOOH$	7.5
2-fluoroacetic acid	CH_2FCOOH	7.99
2-fluorobenzoic acid	C_6H_4FCOOH	8.41
2-nitrobenzoic acid	$C_6H_4NO_2COOH$	7.64

Carboxyl	Rational formula	pK_a
2-phenylacetic acid	$C_6H_5CH_2COOH$	9.43
3-chlorobenzoic acid	$C_6H_4ClCOOH$	8.83
3-cyanobenzoic acid	$C_6H_4CNCOOH$	8.53
3-nitrobenzoic acid	$C_6H_4NO_2COOH$	8.32
3-trifluoromethylbenzoic acid	$C_6H_5CF_3COOH$	8.69
4-bromobenzoic acid	$C_6H_4BrCOOH$	8.93
4-chlorobenzoic acid	$C_6H_4ClCOOH$	9.09
4-cyanobenzoic acid	$C_6H_4CNCOOH$	8.42
4-fluorobenzoic acid	C_6H_4FCOOH	9.23
4-methylbenzoic acid	$C_6H_4CH_3COOH$	9.51
4-nitrobenzoic acid	$C_6H_4NO_2COOH$	8.34
propanoic acid	CH_3CH_2COOH	9.71

Amine	Rational formula	pK_a
2,4,6-trimethylpyridine	$(CH_3)_3C_5H_2N$	7.72
2-bromoaniline	$C_6H_4BrNH_2$	3.46
2-chloroaniline	$C_6H_4ClNH_2$	3.71
2-methylaniline	$C_6H_4CH_3NH_2$	5.95
2-nitroaniline	$C_6H_4NO_2NH_2$	0.2
4-benzylaniline	$C_6H_5C_6H_4NH_2$	5.98
4-chloro-2-nitroaniline	$C_6H_3ClNO_2NH_2$	-0.67
4-chloroaniline	$C_6H_4ClNH_2$	4.95
4-hydroxyaniline	$C_6H_4OHNH_2$	7.41
4-methoxyaniline	$C_6H_4OCH_3NH_2$	6.89
4-methylaniline	$C_6H_4CH_3NH_2$	6.57
4-nitroaniline	$C_6H_4NO_2NH_2$	1.55

Amine	Rational formula	pK_a
aniline	$C_6H_5NH_2$	6.05
hydroxylamine	$HO\text{-}NH_2$	6.29
N-ethylamine	$CH_3CH_2NH_2$	11.00
N-methylamine	CH_3NH_2	11.00
N,N-dimethylamine	$(CH_3)_2NH$	11.2
N,N,N-triethylamine	$(CH_3CH_2)_3N$	10.78
N,N,N-trimethylamine	$(CH_3)_3N$	9.8
o-methylhydroxylamine	CH_3ONH_2	5.13
piperidine	$C_5H_{11}N$	11.07
pyridine	C_5H_5N	5.44
quinoline	C_9H_7N	5.16
ammonia	NH_3	10.78

Table S2 Rational formula and experimental values of test data set for LFC/3D-RISM-SCF in methanol

Phenol	Rational formula	pK_a
2,3-dimethylphenol	$C_6H_3(CH_3)_2OH$	15.08
2,4,6-tribromophenol	$C_6H_2Br_3OH$	10.1
2,5-dimethylphenol	$C_6H_3(CH_3)_2OH$	14.91
2,6-dimethylphenol	$C_6H_3(CH_3)_2OH$	15.26
2-chloro-4-bromophenol	$C_6H_3ClBrOH$	12.7
3,4-dimethylphenol	$C_6H_3(CH_3)_2OH$	14.63
3,5-dichlorophenol	$C_6H_3Cl_2OH$	12.11
3,5-dimethylphenol	$C_6H_3(CH_3)_2OH$	14.57
3,5-dinitrophenol	$C_6H_3(NO_2)_2OH$	10.29
4-hydroxybenzaldehyde	C_6H_4CHOOH	12.01
phenol	C_6H_5OH	14.1

Carboxyl	Rational formula	pK_a
2,2-dichloroaceticacid	$CHCl_2COOH$	6.38
2-cyanoaceticacid	$CH_2CNCOOH$	7.5
2-sulfanylaceticacid	$CH_2SHCOOH$	8.52
3,4-dinitrobenzoicacid	$C_6H_3(NO_2)_2COOH$	7.44
aceticacid	CH_3COOH	9.63
Aspartame	$HOOCCH(CH_2COOH)NH\text{-}COCH(CH_2CH_2C_6H_5)NHCOCH_3$	7.68
benzoicacid	C_6H_5COOH	9.3
marinicacid	$CH_2(COOH)_2$	7.66

Amine	Rational formula	pK_a
1-methylpiperidine	$C_5H_{10}CH_3N$	10.88
2-amino-1-ethanol	$CH_2NH_2CH_2OH$	6.06
2-methylquinoline	$C_9H_6CH_3N$	4.42
3-bromoaniline	$C_6H_4BrNH_2$	5.99
3-hydroxyaniline	$C_6H_4OHNH_2$	6.04
3-methoxyaniline	$C_6H_4OCH_3NH_2$	6.92
4-ethoxyaniline	$C_6H_4OCH_2CH_3NH_2$	6.05
4-methylpyridine	$C_5H_4CH_3N$	5.82
N-ethyl-N-phenylamine	$CH_3CH_2(C_6H_5)NH$	5.45
N-methyl-N-phenylamine	$CH_3(C_6H_5)NH$	10.88

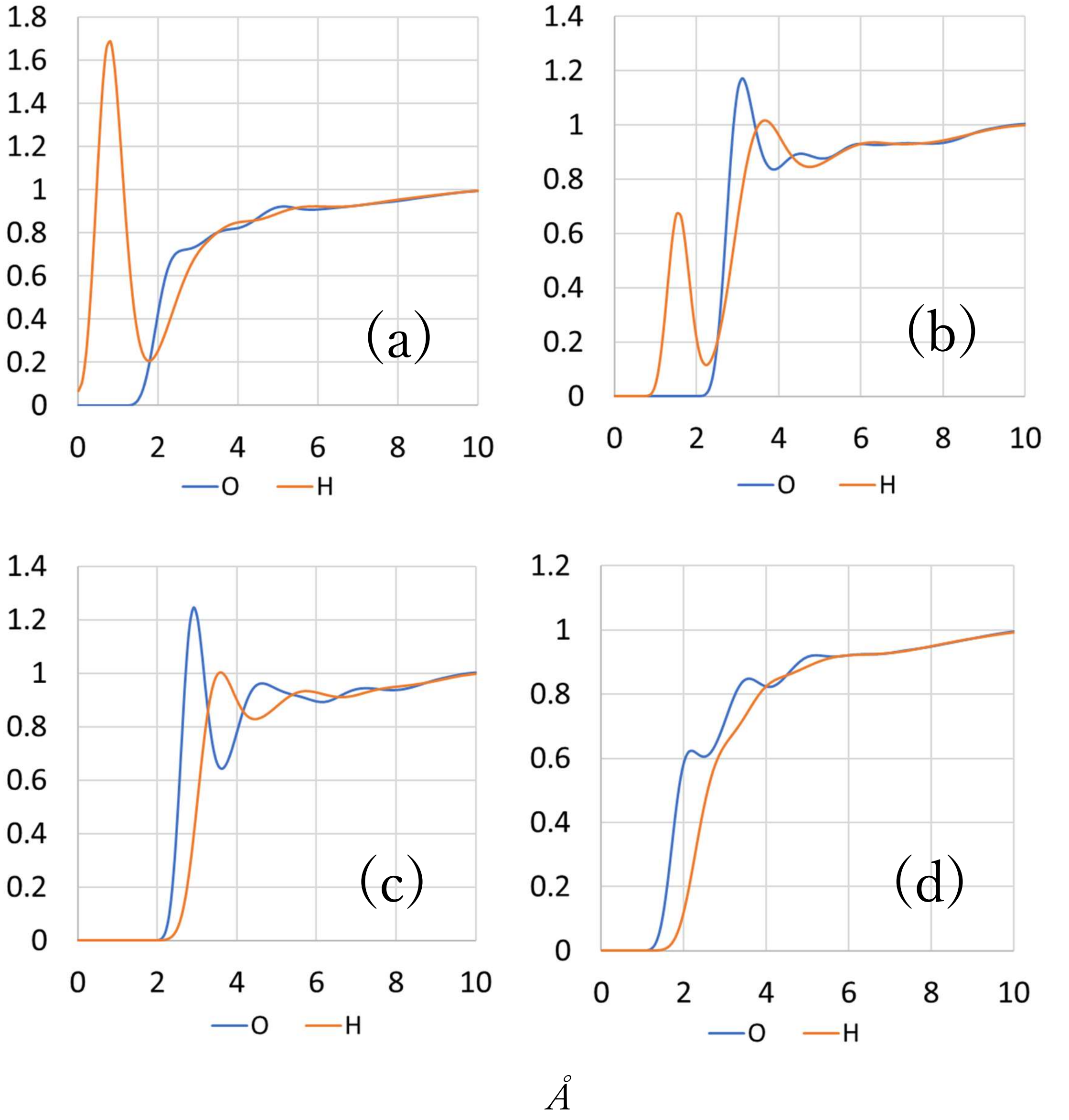

Figure S1. Distribution of oxygen and hydrogen of water molecule around (a), (c) dissociative proton and (b), (d) nitrogen of amine in Lys reference molecule determined by 3D-RISM calculation. Figure (a), (b) are deprotonated and (c), (d) are protonated condition.

List of Publications

[1] "A computational scheme of pK_a values based on the three-dimensional reference interaction site model self-consistent field theory coupled with the linear fitting correction scheme"

R. Fujiki, Y. Kasai, Y. Seno, T. Matsui, Y. Shigeta, N. Yoshida, H. Nakano, *Phys. Chem. Chem. Phys.*, 2018, 20, 27272.

[2] "Application of quantitative prediction method of protonation state in methanol based on quantum chemical calculation and integral equation theory of liquids"

R. Fujiki, T. Matsui, Y. Shigeta, N. Yoshida, H. Nakano, in preparation.

[3] "Development of sampling method of protonation state of polymer based on Constant pH Molecular Dynamics and integral equation theory of liquids"

R. Fujiki, SG. Itoh, H. Okumura, N. Yoshida, H. Nakano, in preparation.

[4] "Recent Developments of Computational Methods for pK_a Prediction Based on Electronic Structure Theory with Solvation Models"

R. Fujiki, T. Matsui, Y. Shigeta, H. Nakano, N. Yoshida, *J*, 2021, 4, 849.

Acknowledgements

I would like to express deepest appreciation to Professor Haruyuki Nakano and Professor Norio Yoshida for their grateful discussions and advice. I would also like to thank Dr. Yoshihiro Watanabe and Dr. Satoshi Suzuki for their helpful comments. I thank all the members in Nakano group in Kyushu University for their useful comments. I am grateful to all collaborators, Profs. Shigeta, Dr. Matsui, Dr. Itoh and Dr. Okumura, for their helpful comments.