Perturbation analyses of intermolecular interactions

Yohei M. Koyama,[1,*] Tetsuya J. Kobayashi,[2] and Hiroki R. Ueda[1,3,4,5,6,†]

[1]*Laboratory for Synthetic Biology, Quantitative Biology Center, RIKEN, 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan*

[2]*Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan*

[3]*Laboratory for Systems Biology, Center for Developmental Biology, RIKEN, 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan*

[4]*Functional Genomics Unit, Center for Developmental Biology, RIKEN, 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan*

[5]*Department of Bioscience, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan*

[6]*Department of Mathematics, Kyoto University, Kyoto 606-8502, Japan*

(Received 5 November 2010; revised manuscript received 19 June 2011; published 12 August 2011)

Conformational fluctuations of a protein molecule are important to its function, and it is known that environmental molecules, such as water molecules, ions, and ligand molecules, significantly affect the function by changing the conformational fluctuations. However, it is difficult to systematically understand the role of environmental molecules because intermolecular interactions related to the conformational fluctuations are complicated. To identify important intermolecular interactions with regard to the conformational fluctuations, we develop herein (i) distance-independent and (ii) distance-dependent perturbation analyses of the intermolecular interactions. We show that these perturbation analyses can be realized by performing (i) a principal component analysis using conditional expectations of truncated and shifted intermolecular potential energy terms and (ii) a functional principal component analysis using products of intermolecular forces and conditional cumulative densities. We refer to these analyses as intermolecular perturbation analysis (IPA) and distance-dependent intermolecular perturbation analysis (DIPA), respectively. For comparison of the IPA and the DIPA, we apply them to the alanine dipeptide isomerization in explicit water. Although the first IPA principal components discriminate two states (the α state and PPII (polyproline II) + β states) for larger cutoff length, the separation between the PPII state and the β state is unclear in the second IPA principal components. On the other hand, in the large cutoff value, DIPA eigenvalues converge faster than that for IPA and the top two DIPA principal components clearly identify the three states. By using the DIPA biplot, the contributions of the dipeptide-water interactions to each state are analyzed systematically. Since the DIPA improves the state identification and the convergence rate with retaining distance information, we conclude that the DIPA is a more practical method compared with the IPA. To test the feasibility of the DIPA for larger molecules, we apply the DIPA to the ten-residue chignolin folding in explicit water. The top three principal components identify the four states (native state, two misfolded states, and unfolded state) and their corresponding eigenfunctions identify important chignolin-water interactions to each state. Thus, the DIPA provides the practical method to identify conformational states and their corresponding important intermolecular interactions with distance information.

DOI: 10.1103/PhysRevE.84.026704 PACS number(s): 02.70.Rr, 36.20.Ey, 87.15.ap, 89.70.Cf

I. INTRODUCTION

Conformational fluctuations of a protein molecule are important to its function. Environmental molecules such as solvents [1,2] and ligands [3–6] significantly affect protein function by modifying its conformational fluctuations. However, it has been difficult to systematically analyze such environmental effects on the conformational fluctuation of proteins. Several methods have been developed to analyze the conformational fluctuations of a protein. For example, correlation analysis using atomic coordinates [7,8] and inter-residue interaction energy [9,10] reveal communication within the protein. The principal component analysis (PCA) [11] using atomic coordinates [12–14], atomic pair distances [15], and mapped dihedral angles [16,17] decomposes the fluctuations into large uncorrelated fluctuations. However, these methods are based only on the analysis of protein coordinates, so direct information regarding the environmental molecules is lost. Therefore, although the contributions of environmental molecules might be implicitly represented in conformational fluctuations of a protein, it is difficult to explicitly evaluate their direct contribution to the conformational fluctuations of a protein.

To understand the molecular conformational fluctuations based on their atomic interactions, we previously introduced the potential energy PCA (PEPCA) [18] in which we can identify molecular conformational fluctuations (or states) by the principal components and the important interactions by the corresponding eigenvectors. Compared to other methods, PEPCA can potentially be applied to evaluate the direct contribution of environmental molecules because PEPCA is based on the analyzing potential energies, which can represent not only intramolecular but also intermolecular atomic interactions. However, directly applying PEPCA to investigate the contribution of environmental molecules presents some difficulties. This is because PEPCA using intermolecular potential energy

*ym.koyama@gmail.com
†uedah-tky@umin.ac.jp

terms identify the intermolecular interactions that induce the largest conformational change of the "whole" system, namely, a protein and its environmental molecules (see Appendix A). Therefore, PEPCA will collect the large fluctuations of the environmental molecules that may be irrelevant to conformational fluctuations of the target protein. To understand environmental effects on conformational fluctuations of the target protein, we need to focus on the conformational change of the target protein rather than the whole system. In the present study, we generalize our previous results [18] to evaluate environmental effects on the conformational change of the target molecule.

This article is organized as follows: With respect to arbitrary intermolecular perturbations, we show that changes in the conformational distribution of the target molecule due to the perturbation can be evaluated by the variance of the conditional expectation of the perturbation potential energy. By using this result, we introduce (i) distance-independent and (ii) distance-dependent perturbations of intermolecular interactions. Then, we search the perturbations that induce the largest change in the conformational distribution of the target molecule. We show that these can be solved by (i) PCA using conditional expectations of truncated and shifted intermolecular potential energy terms and (ii) functional principal component analysis (FPCA) [11,19] using products of intermolecular forces and conditional cumulative densities. We refer to these analyses as (i) intermolecular perturbation analysis (IPA) and (ii) distance-dependent intermolecular perturbation analysis (DIPA), respectively. For comparison of the IPA and the DIPA, we applied them to the alanine dipeptide isomerization in explicit water. We see that the DIPA is more practical compared with the IPA. To test the feasibility of the DIPA for larger molecules, we apply the DIPA to the ten-residue chignolin folding in explicit water. Finally, we discuss the computational cost of the DIPA for protein molecules.

II. THEORY

A. Change in conformational distribution of target molecule due to perturbation

We consider the change in the conformational distribution of the target molecule induced by a general perturbation. The molecular system is described by the potential energy $V(\mathbf{q},\mathbf{q}')$, where $\mathbf{q}$ are the coordinates of the target molecule and $\mathbf{q}'$ are the coordinates of the environmental molecules. In this case, the canonical distribution of the whole system at inverse temperature β is represented as

$$\rho(\mathbf{q},\mathbf{q}') = \frac{1}{Z} e^{-\beta V(\mathbf{q},\mathbf{q}')}, \tag{1}$$

where Z is the partition function. Next, we perturb the system by adding a perturbation potential energy $\Delta V(\mathbf{q},\mathbf{q}')$:

$$V'(\mathbf{q},\mathbf{q}') = V(\mathbf{q},\mathbf{q}') + \Delta V(\mathbf{q},\mathbf{q}'). \tag{2}$$

The perturbed canonical distribution $\rho'(\mathbf{q},\mathbf{q}')$ is represented as

$$\rho'(\mathbf{q},\mathbf{q}') = \frac{1}{Z'} e^{-\beta V'(\mathbf{q},\mathbf{q}')} = \frac{e^{-\beta\Delta V(\mathbf{q},\mathbf{q}')}}{\langle e^{-\beta\Delta V}\rangle}\rho(\mathbf{q},\mathbf{q}'), \tag{3}$$

where we defined the expectation value as

$$\langle e^{-\beta\Delta V}\rangle \equiv \int e^{-\beta\Delta V}\rho(\mathbf{q},\mathbf{q}')d\mathbf{q}d\mathbf{q}'. \tag{4}$$

By using Eq. (3), the perturbed conformational distribution of the target molecule is

$$\rho'(\mathbf{q}) = \int \rho'(\mathbf{q},\mathbf{q}')d\mathbf{q}' = \frac{\langle e^{-\beta\Delta V}|\mathbf{q}\rangle}{\langle e^{-\beta\Delta V}\rangle}\rho(\mathbf{q}), \tag{5}$$

where the conditional expectation is defined as

$$\langle e^{-\beta\Delta V(\mathbf{q},\mathbf{q}')}|\mathbf{q}\rangle \equiv \int e^{-\beta\Delta V(\mathbf{q},\mathbf{q}')}\rho(\mathbf{q}'|\mathbf{q})d\mathbf{q}'. \tag{6}$$

Next, we measure the change in conformational distribution of the target molecule due to the perturbation. For this purpose, we use the Kullback-Leibler divergence (or relative entropy) [20–24]

$$D(\rho'(\mathbf{q})||\rho(\mathbf{q})) \equiv \int \rho'(\mathbf{q})\ln\frac{\rho'(\mathbf{q})}{\rho(\mathbf{q})}d\mathbf{q} \geqslant 0. \tag{7}$$

This can be considered to be the expectation of the log-ratio $[\ln\rho'(\mathbf{q})/\rho(\mathbf{q})]$ under the perturbed equilibrium state. By using the identity $\rho' = \rho\exp(\ln\rho'/\rho)$, Eq. (7) can be expressed by expectations under the unperturbed equilibrium state as

$$D(\rho'(\mathbf{q})||\rho(\mathbf{q})) = \sum_{k=1}^{\infty}\frac{1}{(k-1)!}\left\langle\left(\ln\frac{\rho'(\mathbf{q})}{\rho(\mathbf{q})}\right)^k\right\rangle. \tag{8}$$

By applying cumulant expansions to Eq. (5), the ratio change by the perturbation can be expanded as follows:

$$\ln\frac{\rho'(\mathbf{q})}{\rho(\mathbf{q})} = -\beta\left(\langle\Delta V|\,\mathbf{q}\rangle - \langle\Delta V\rangle\right) + \frac{1}{2}[\mathrm{var}(\beta\Delta V|\mathbf{q}) - \mathrm{var}(\beta\Delta V)] + \cdots, \tag{9}$$

where $\mathrm{var}(\beta\Delta V)$ and $\mathrm{var}(\beta\Delta V|\mathbf{q})$ are the variance and the conditional variance of $\beta\Delta V$, respectively. By using Eq. (9) and the equality [25]

$$\mathrm{var}(\beta\Delta V) = \mathrm{var}(\beta\,\langle\Delta V|\mathbf{q}\rangle) + \langle\mathrm{var}(\beta\Delta V|\mathbf{q})\rangle \tag{10}$$

(sometimes called the law of total variance), Eq. (8) can be expanded as

$$D(\rho'(\mathbf{q})||\rho(\mathbf{q})) = \tfrac{1}{2}\mathrm{var}(\beta\,\langle\Delta V|\mathbf{q}\rangle) + \cdots. \tag{11}$$

Thus, within the second-order approximation, we can evaluate the change in the conformational distribution of the target molecule by the variance of the conditional expectation of the perturbation potential energy under the unperturbed equilibrium state. In Appendix A, we summarize the relationship between the change in the conformational distribution of the whole system and the target molecule by the perturbation.

B. Intermolecular perturbation analysis

We develop a perturbation analysis of intermolecular interactions. One natural perturbation of an interaction is to multiply the original force with $1+\lambda$ by introducing a perturbation parameter λ. With this perturbation, the interaction is strengthened when $\lambda > 0$ and it is weakened when $\lambda < 0$. If a perturbation to a certain interaction significantly affects the

conformational distribution of the target molecule, we consider it to be an "important" interaction. In the current ordinary classical force fields, intermolecular interactions consist of van der Waals interactions or electrostatic interactions. Since they only depend on the atomic distance r, we denote a nonbonded potential energy between two atoms as $\phi(r)$. To avoid dealing the infinite long-range interaction, we only perturb the interactions within some cutoff length r_c. By introducing a truncated and shifted potential energy [26]

$$\phi(r;r_c) \equiv \begin{cases} \phi(r) - \phi(r_c) & \text{for} \quad r \leqslant r_c, \\ 0 & \text{for} \quad r > r_c, \end{cases} \tag{12}$$

the above-mentioned perturbation can be realized by adding potential energy $\lambda\phi(r;r_c)$ to the original potential energy. By increasing the cutoff, we check the convergence of the perturbation analysis.

Then, we perturb all intermolecular interactions by introducing perturbation parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_M)^T$. If the system includes atoms that have permutation symmetry [27], the perturbed potential energy is not invariant to the permutation of the atoms. To keep permutation invariance in the perturbed potential energy, perturbation parameters must have the identical values when the corresponding potential energy terms include atoms with permutation symmetry. With these considerations, we can perturb all intermolecular interactions with permutation invariance by adding perturbation potential energy

$$\Delta V(\mathbf{q},\mathbf{q}') = \sum_{k=1}^{M} \lambda_k V_k(\mathbf{q},\mathbf{q}'), \tag{13}$$

where $\mathbf{V}(\mathbf{q},\mathbf{q}') = [V_1(\mathbf{q},\mathbf{q}'), \ldots, V_M(\mathbf{q},\mathbf{q}')]^T$ are defined as

$$V_k(\mathbf{q},\mathbf{q}') \equiv \sum_{i\in I_k} \sum_{j\in J_k} \phi_k(r_{ij};r_c). \tag{14}$$

Labels I_k and J_k are the set of atoms with permutation symmetry in the target molecule and the environmental molecules, respectively. Although a perturbation within the environmental molecule (such as water-water interactions) can change the conformational distribution of the target molecule, for simplicity we only perturb interactions between the target molecule and environment molecules (such as protein-water interactions) in this study.

To identify important intermolecular interactions, we find a perturbation $\boldsymbol{\lambda}$ that induces a large change in the conformational distribution. For this purpose, we search for $\boldsymbol{\lambda}$ that maximizes $D(\rho_{\boldsymbol{\lambda}}(\mathbf{q})||\rho(\mathbf{q}))$ under the constraint $|\boldsymbol{\lambda}| = \delta$, where $\rho_{\boldsymbol{\lambda}}(\mathbf{q})$ is the conformational distribution of the target molecule under the perturbation described by Eq. (13). Within the second-order approximation, this is achieved by finding $\boldsymbol{\lambda}$ that maximizes the variance in Eq. (11). By using Eq. (13), the variance is represented as

$$\mathrm{var}(\beta\langle \Delta V|\mathbf{q}\rangle) = \mathrm{var}\left(\beta \sum_{k=1}^{M} \lambda_k \langle V_k|\mathbf{q}\rangle\right) \tag{15}$$

$$= \boldsymbol{\lambda}^T \,\mathrm{cov}(\beta\langle \mathbf{V}|\mathbf{q}\rangle)\boldsymbol{\lambda}, \tag{16}$$

where $\mathrm{cov}(\beta\langle \mathbf{V}|\mathbf{q}\rangle)$ is the covariance matrix whose (i,j) entry is the covariance between $\beta\langle V_i|\mathbf{q}\rangle$ and $\beta\langle V_j|\mathbf{q}\rangle$. Since the covariance matrix is a positive semidefinite matrix, it can be diagonalized with non-negative eigenvalues $\sigma_1^2, \ldots, \sigma_M^2$, which are sorted in descending order, and the corresponding orthogonal matrix $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_M)$ [11,28] as

$$\mathrm{cov}(\beta\langle \mathbf{V}|\mathbf{q}\rangle)\mathbf{u}_i = \sigma_i^2 \mathbf{u}_i. \tag{17}$$

The ith eigenvector perturbation $\boldsymbol{\lambda} = \delta\mathbf{u}_i$ maximizes the variance Eq. (15) under $|\boldsymbol{\lambda}| = \delta$ and $\boldsymbol{\lambda}\cdot\mathbf{u}_1 = 0, \ldots, \boldsymbol{\lambda}\cdot\mathbf{u}_{i-1} = 0$. By using Eqs. (11), (16), and (17), the change in the conformational distribution of the target molecule due to $\boldsymbol{\lambda} = \delta\mathbf{u}_i$ becomes

$$D(\rho_{\delta\mathbf{u}_i}(\mathbf{q})||\rho(\mathbf{q})) = \tfrac{1}{2}\delta^2\sigma_i^2 + \cdots. \tag{18}$$

Thus, we can find the important combinations of the interactions to the target molecule in the top eigenvectors. This procedure can be considered to be the PCA [11] using $\beta\langle \mathbf{V}|\mathbf{q}\rangle$ or $-\beta\langle \mathbf{V}|\mathbf{q}\rangle$.

Next, we consider the contribution of the ith eigenvector to the conformational distribution of the target molecule. The perturbation $\boldsymbol{\lambda} = \delta\mathbf{u}_i$ corresponds to the potential energy perturbation [Eq. (13)] as

$$\Delta V(\mathbf{q},\mathbf{q}') = \delta \sum_{k=1}^{M} U_{ki} V_k(\mathbf{q},\mathbf{q}'). \tag{19}$$

By using Eqs. (9) and (19), the change in the ratio (or population shift) of the conformation $\mathbf{q}$ induced by the perturbation $\boldsymbol{\lambda} = \delta\mathbf{u}_i$ is

$$\ln \frac{\rho_{\delta\mathbf{u}_i}(\mathbf{q})}{\rho(\mathbf{q})} = \delta g_i(\mathbf{q}) + \cdots, \tag{20}$$

where we have introduced

$$g_i(\mathbf{q}) \equiv -\beta\mathbf{u}_i \cdot (\langle \mathbf{V}|\mathbf{q}\rangle - \langle \mathbf{V}\rangle). \tag{21}$$

By definition, $g_i(\mathbf{q})$ is the ith principal component of the PCA using $-\beta\langle \mathbf{V}|\mathbf{q}\rangle$. Note that these principal components have the opposite sign with respect to the definition in our previous article [18], which uses $\beta\mathbf{V}(\mathbf{q})$. Performing PCA using $-\beta\langle \mathbf{V}|\mathbf{q}\rangle$ instead of $\beta\langle \mathbf{V}|\mathbf{q}\rangle$ is useful to visualize the results by the biplot [29,30], as we will see in the numerical results. By using Eqs. (19) and (20), we can analyze in detail the effects of the ith eigenvector on the target molecular conformational distribution. We consider the case $\delta > 0$. From Eq. (19), the perturbation $\delta\mathbf{u}_i$ strengthens the kth potential energy term V_k if $U_{ki} > 0$, and it weakens this term if $U_{ki} < 0$. From Eq. (20), the perturbation $\delta\mathbf{u}_i$ increases the ratio of $\mathbf{q}$ if $g_i(\mathbf{q}) > 0$, and decreases it if $g_i(\mathbf{q}) < 0$ in the first-order approximation. These results are summarized in Table I.

In summary, the PCA using $-\beta\langle \mathbf{V}|\mathbf{q}\rangle$ identifies the combinations of the intermolecular interactions that are important for the conformational distribution of the target molecule. We refer to the PCA using $-\beta\langle \mathbf{V}|\mathbf{q}\rangle$ as the IPA. The ith eigenvector $\mathbf{u}_i$ is the ith important combination of the interactions [Eq. (19)], the ith eigenvalue σ_i^2 measures the change in the conformational distribution [Eq. (18)], and the ith principal component $g_i(\mathbf{q})$ represents the change in the ratio of the target molecular conformation $\mathbf{q}$ due to the perturbation [Eq. (20)]. The detailed contributions of the interactions to the target molecular conformation are given in Table I. If we only perturb interactions within the target molecule, the equality

TABLE I. Effects of the ith eigenvector perturbation. A change in the weight of a truncated and shifted intermolecular potential energy term V_k is determined by Eq. (19). The ratio change in $\mathbf{q}$ is determined by Eq. (20). The term "increase" or "decrease" indicates the ratio change for the first-order approximation.

	The kth interaction		Ratio of $\mathbf{q}$	
	$U_{ki} > 0$	$U_{ki} < 0$	$g_i(\mathbf{q}) > 0$	$g_i(\mathbf{q}) < 0$
$\delta > 0$	Strengthen	Weaken	Increase	Decrease
$\delta < 0$	Weaken	Strengthen	Decrease	Increase

$\langle \mathbf{V}|\,\mathbf{q}\rangle = \mathbf{V}(\mathbf{q})$ holds. Therefore, IPA includes the PEPCA that corresponds to the intramolecular perturbation analysis [18]. In this sense, IPA is a natural generalization of PEPCA.

C. Distance-dependent intermolecular perturbation analysis

We develop a perturbation analysis of intermolecular interactions with distance dependence by generalizing the IPA. We consider the distance-dependent perturbation of an intermolecular interaction between the ith atom in the target molecule and the jth atom in the environmental molecules. We represent the force acting on the ith atom from the jth atom through van der Waals or electrostatic interaction as $\mathbf{F}_{ji}$, and the distance between the ith atom and the jth atom as r_{ij}. With this notation, one natural distance-dependent perturbation is represented as

$$[1+\lambda(r_{ij})]\mathbf{F}_{ji}, \tag{22}$$

where the distance-dependent perturbation function $\lambda(r_{ij})$ is introduced. To derive the potential energy that realizes the perturbation given in Eq. (22), we first note that the force $\mathbf{F}_{ji}$ is the gradient of the corresponding potential energy $\phi(r_{ij})$. Therefore, if we introduce coordinates of the ith atom as (x_i,y_i,z_i), the force becomes

$$\mathbf{F}_{ji} = -\left(\frac{\partial}{\partial x_i},\frac{\partial}{\partial y_i},\frac{\partial}{\partial z_i}\right)\phi(r_{ij}) = F(r_{ij})\,\mathbf{e}_{ji}, \tag{23}$$

where we have introduced

$$F(r) \equiv -\frac{d\phi(r)}{dr} \tag{24}$$

and $\mathbf{e}_{ji}$ is a three-dimensional unit vector directed from atom j to atom i. By using Eqs. (23) and (24), the perturbation given in Eq. (22) can be realized by adding a potential energy

$$\phi(r;r_c,\lambda) \equiv \begin{cases} \int_r^{r_c}\lambda(r')F(r')dr' & \text{for} \quad r \leqslant r_c, \\ 0 & \text{for} \quad r > r_c, \end{cases} \tag{25}$$

where we have introduced the cutoff distance r_c for convenience in the following analyses. If we ignore the dependence on distance by using $\lambda(r)=\lambda$, Eq. (25) reduces to

$$\phi(r;r_c,\lambda) = \lambda\phi(r;r_c), \tag{26}$$

where $\phi(r;r_c)$ is defined in Eq. (12). The perturbation is identical to that of the IPA. Therefore, Eq. (25) is a natural generalization of the perturbation used in the IPA. Then, we perturb all intermolecular interactions with permutation invariance by adding the potential energy $\Delta V(\mathbf{q},\mathbf{q}')$ with perturbation parameters $\boldsymbol{\lambda}(r) = [\lambda_1(r),\ldots,\lambda_M(r)]^T$ as

$$\Delta V(\mathbf{q},\mathbf{q}') = \sum_{k=1}^{M}\Delta V_k(\mathbf{q},\mathbf{q}'), \tag{27}$$

where

$$\Delta V_k(\mathbf{q},\mathbf{q}') \equiv \sum_{i\in I_k, j\in J_k}\phi_k(r_{ij};r_c,\lambda_k). \tag{28}$$

The meaning of labels I_k and J_k are identical in Sec. II B.

Next, we quantify the perturbation with Eq. (11). By introducing the conditional density $n_{iJ_k}(r|\mathbf{q})$ of J_k atoms at a distance r from the ith atom for a given $\mathbf{q}$, the conditional expectation of Eq. (28) can be expressed as

$$\langle\Delta V_k|\,\mathbf{q}\rangle = \sum_{i\in I_k}\int_0^{r_c}\left(\int_r^{r_c}\lambda_k(r')F_k(r')dr'\right)n_{iJ_k}(r|\mathbf{q})\,dr. \tag{29}$$

By defining the conditional cumulative density

$$N_{iJ_k}(r|\mathbf{q}) \equiv \int_0^r n_{iJ_k}(r'|\mathbf{q})\,dr' \tag{30}$$

and using the equality

$$\frac{dN_{iJ_k}(r|\mathbf{q})}{dr} = n_{iJ_k}(r|\mathbf{q}), \tag{31}$$

the integration by parts of Eq. (29) leads to

$$\langle\Delta V_k|\,\mathbf{q}\rangle = \int_0^{r_c}\lambda_k(r)f_k(r|\mathbf{q})\,dr, \tag{32}$$

where we have defined $\mathbf{f}(r|\mathbf{q}) = [f_1(r|\mathbf{q}),\ldots,f_M(r|\mathbf{q})]^T$ as

$$f_k(r|\mathbf{q}) \equiv F_k(r)\sum_{i\in I_k}N_{iJ_k}(r|\mathbf{q}). \tag{33}$$

As shown in Appendix B, $f_k(r|\mathbf{q})$ is also characterized by a derivative of $\langle V_k|\,\mathbf{q}\rangle$ used in the IPA. By using Eqs. (27) and (32), the variance in Eq. (11) is expressed as

$$\begin{aligned}&\mathrm{var}(\beta\,\langle\Delta V|\,\mathbf{q}\rangle)\\ &= \mathrm{var}\left(\beta\sum_{k=1}^{M}\int_0^{r_c}\lambda_k(r)f_k(r|\mathbf{q})\,dr\right) &(34)\\ &= \int_0^{r_c}\int_0^{r_c}\boldsymbol{\lambda}(r)^T\,\mathrm{cov}(\beta\mathbf{f}(r|\mathbf{q}),\beta\mathbf{f}(r'|\mathbf{q}))\boldsymbol{\lambda}(r')drdr', &(35)\end{aligned}$$

where $\mathrm{cov}(\beta\mathbf{f}(r|\mathbf{q}),\beta\mathbf{f}(r'|\mathbf{q}))$ is an $M\times M$ covariance matrix whose i,j element is the covariance between $\beta f_i(r|\mathbf{q})$ and $\beta f_j(r'|\mathbf{q})$ for given r and r'.

By introducing a covariance operator $\hat{\mathbf{C}}$ [19] that is defined as

$$\hat{\mathbf{C}}\boldsymbol{\lambda} \equiv \int_0^{r_c}\mathrm{cov}(\beta\mathbf{f}(r|\mathbf{q}),\beta\mathbf{f}(r'|\mathbf{q}))\boldsymbol{\lambda}(r')\,dr' \tag{36}$$

and an inner product

$$\langle\boldsymbol{\lambda},\boldsymbol{\lambda}'\rangle \equiv \int_0^{r_c}\boldsymbol{\lambda}(r)^T\boldsymbol{\lambda}'(r)\,dr, \tag{37}$$

Eq. (35) can be represented as

$$\mathrm{var}(\beta\langle\Delta V|\mathbf{q}\rangle) = \langle\boldsymbol{\lambda},\hat{\mathbf{C}}\boldsymbol{\lambda}\rangle \geqslant 0. \tag{38}$$

By using the equality

$$\mathrm{cov}(\beta\mathbf{f}(r|\mathbf{q}),\beta\mathbf{f}(r'|\mathbf{q}))^T = \mathrm{cov}(\beta\mathbf{f}(r'|\mathbf{q}),\beta\mathbf{f}(r|\mathbf{q})), \quad (39)$$

we can show that

$$\langle\hat{\mathbf{C}}\boldsymbol{\lambda},\boldsymbol{\lambda}'\rangle = \langle\boldsymbol{\lambda},\hat{\mathbf{C}}\boldsymbol{\lambda}'\rangle. \quad (40)$$

Therefore, $\hat{\mathbf{C}}$ is a self-adjoint operator. Furthermore, the non-negativity of the variance [Eq. (38)] indicates that $\hat{\mathbf{C}}$ is the positive-semidefinite operator. Thus, $\hat{\mathbf{C}}$ can be diagonalized by the orthonormal eigenfunctions $\mathbf{u}_i(r) = [U_{1i}(r),\ldots,U_{Mi}(r)]^T$ with the corresponding non-negative eigenvalues σ_i^2 (which are sorted in descending order) as

$$\hat{\mathbf{C}}\mathbf{u}_i = \sigma_i^2\mathbf{u}_i. \quad (41)$$

The orthonormality of the eigenfunctions is represented as

$$\langle\mathbf{u}_i,\mathbf{u}_j\rangle = \delta_{i,j}. \quad (42)$$

Generally, the eigenfunctions $\mathbf{u}_i(r)$ constitute an infinite series $i = 1,\ldots,\infty$.

To identify important intermolecular interactions with their distance information, we determine $\boldsymbol{\lambda}(r)$ that maximizes $D(\rho_{\boldsymbol{\lambda}}(\mathbf{q})||\rho(\mathbf{q}))$ under the constraint $|\boldsymbol{\lambda}| = \delta$, where the norm $|\boldsymbol{\lambda}|$ is defined as $|\boldsymbol{\lambda}| \equiv \sqrt{\langle\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle}$. Within the second-order approximation [Eq. (11)], the maximization is equivalent to maximizing $\mathrm{var}(\beta\langle\Delta V|\mathbf{q}\rangle)$ under the constraint $|\boldsymbol{\lambda}| = \delta$. Because the variance can be represented as the quadratic form of the covariance operator $\hat{\mathbf{C}}$ [Eq. (38)], the variance maximization can be solved by the diagonalization [Eq. (41)] [19]. In particular, a perturbation $\boldsymbol{\lambda}(r) = \delta\mathbf{u}_i(r)$ maximizes the variance under the constraints $|\boldsymbol{\lambda}| = \delta$ and $\langle\boldsymbol{\lambda},\mathbf{u}_1\rangle = 0,\ldots,\langle\boldsymbol{\lambda},\mathbf{u}_{i-1}\rangle = 0$. By the perturbation $\boldsymbol{\lambda}(r) = \delta\mathbf{u}_i(r)$, Eq. (11) becomes Eq. (18). Thus, eigenfunctions with larger eigenvalue represent the important intermolecular interactions with their distance information. We note that finding $\boldsymbol{\lambda}(r)$ that maximizes the functional form of Eq. (34) under the constraints $|\boldsymbol{\lambda}| = \delta$ and $\langle\boldsymbol{\lambda},\mathbf{u}_1\rangle = 0,\ldots,\langle\boldsymbol{\lambda},\mathbf{u}_{i-1}\rangle = 0$ can be considered as the FPCA [11,19] using $\beta\mathbf{f}(r|\mathbf{q})$ or $-\beta\mathbf{f}(r|\mathbf{q})$. By the perturbation $\boldsymbol{\lambda}(r) = \delta\mathbf{u}_i(r)$, Eq. (9) becomes Eq. (20), where $g_i(\mathbf{q})$ is defined as

$$g_i(\mathbf{q}) \equiv \langle\mathbf{u}_i(r),-\beta[\mathbf{f}(r|\mathbf{q}) - \langle\mathbf{f}(r|\mathbf{q})\rangle]\rangle. \quad (43)$$

By definition, $g_i(\mathbf{q})$ is the ith principal component of the FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$.

In summary, performing the FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$ identifies the important combinations of the intermolecular interactions to the target molecule with their distance information. We refer to the FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$ as the DIPA. In a manner similar to the PEPCA and the IPA, we can interpret the eigenfunctions, eigenvalues, and principal components of DIPA in terms of the perturbation as follows: The ith eigenfunction perturbation $\boldsymbol{\lambda}(r) = \delta\mathbf{u}_i(r)$ induces the largest change in the conformational distribution of the target molecule under the constraints $|\boldsymbol{\lambda}| = \delta$ and $\langle\boldsymbol{\lambda},\mathbf{u}_1\rangle = 0,\ldots,\langle\boldsymbol{\lambda},\mathbf{u}_{i-1}\rangle = 0$; the ith eigenvalue σ_i^2 measures the change in the conformational distribution of the target molecule by $\frac{1}{2}\delta^2\sigma_i^2$ [Eq. (18)]; the ith principal component $g_i(\mathbf{q})$ represents the ratio change of the conformation $\mathbf{q}$ of the target molecule by $\delta g_i(\mathbf{q})$ [Eq. (20)].

Finally, we summarize perturbation analyses of atomic interactions. The common idea is to find perturbations of atomic interactions that significantly change the conformational distribution of the target molecule [Fig. 1(a)]. The difference is the three ways in which the atomic interactions are perturbed: (i) If we consider perturbations of intramolecular interactions [Fig. 1(b)], the perturbation analysis can be solved by the PCA using $-\beta\mathbf{V}(\mathbf{q})$, where $\mathbf{V}(\mathbf{q})$ are intramolecular potential energy terms (PEPCA) [18]. (ii) If we consider perturbations of intermolecular interactions [Fig. 1(c)], the perturbation analysis can be solved by the PCA using $-\beta\langle\mathbf{V}|\mathbf{q}\rangle$, where $\langle\mathbf{V}|\mathbf{q}\rangle$ are conditional expectations of truncated and shifted intermolecular potential energy terms. (iii) If we consider distance-dependent perturbations of intermolecular interactions [Fig. 1(d)], the perturbation analysis can be solved by the FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$, where $\mathbf{f}(r|\mathbf{q})$ are products of intermolecular forces and conditional cumulative densities defined in Eq. (33) (DIPA).

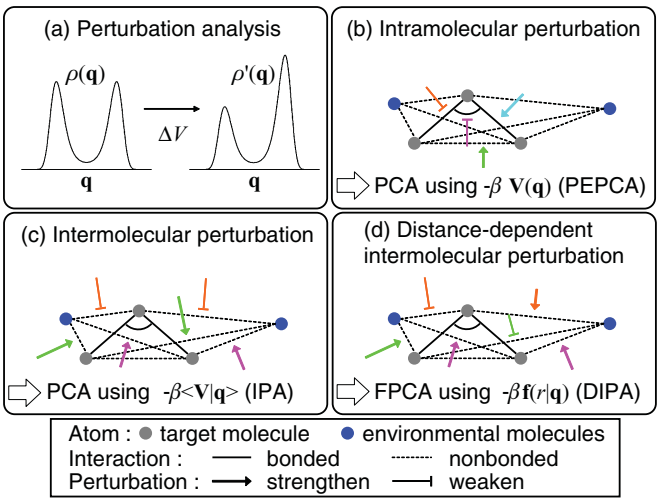

FIG. 1. (Color online) (a) Perturbation analysis of atomic interactions. Atomic interactions are perturbed by adding a perturbation potential energy ΔV. The perturbation effect is measured by the change in the conformation distribution of the target molecule $\mathbf{q}$. (b) Perturbation analysis of intramolecular interactions, which can be solved by the PCA using $-\beta\mathbf{V}(\mathbf{q})$ (PEPCA) [18], where $\mathbf{V}(\mathbf{q})$ are intramolecular potential energy terms. (c) Perturbation analysis of intermolecular interactions, which can be solved by performing the PCA using $-\beta\langle\mathbf{V}|\mathbf{q}\rangle$ (IPA), where $\langle\mathbf{V}|\mathbf{q}\rangle$ are conditional expectations of truncated and shifted intermolecular potential energy terms. (d) Perturbation analysis of intermolecular interactions with dependence on distance, which can be solved by performing the FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$ (DIPA), where $\mathbf{f}(r|\mathbf{q})$ are products of intermolecular forces and conditional cumulative densities as defined in Eq. (33).

D. IPA and DIPA procedure

For IPA and DIPA, we need to estimate the conditional expectations and the conditional cumulative densities, respectively. These require additional molecular dynamics (MD) simulations of environmental molecules $\mathbf{q}'$ with a fixed target molecule $\mathbf{q}$ to sample $\mathbf{q}'$ that follows $\rho(\mathbf{q}'|\mathbf{q})$. For this purpose, first, we generate coordinates $\mathbf{q}$ and $\mathbf{q}'$ that follow the canonical distribution $\rho(\mathbf{q},\mathbf{q}')$. This can be done using an MD simulation that can generate the canonical distribution such as a Langevin dynamics [Fig. 2(a)]. Then, we can perform PEPCA [18] by

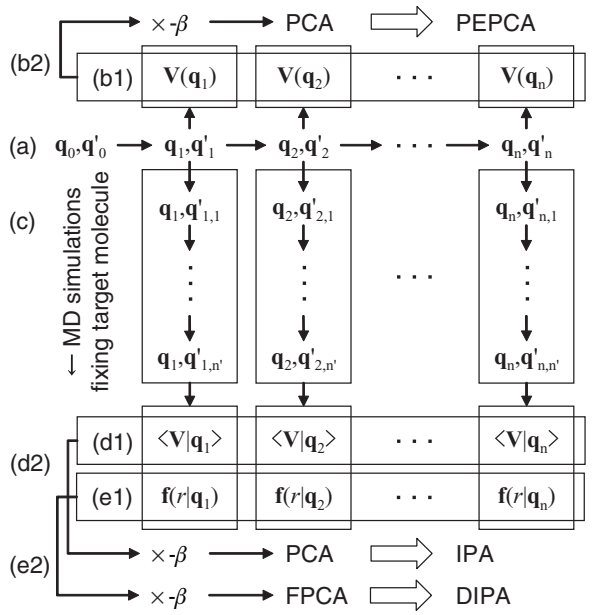

FIG. 2. Procedure to perform PEPCA, IPA, and DIPA. (a) An MD simulation of the target molecule $\mathbf{q}$ and the environmental molecules $\mathbf{q}'$. (b1) Calculation of intramolecular potential energy terms $\mathbf{V}(\mathbf{q})$. (b2) PCA using $-\beta\mathbf{V}(\mathbf{q})$ (PEPCA [18]). (c) MD simulations of the environmental molecules $\mathbf{q}'$ with the target molecule $\mathbf{q}$ fixed. (d1) Calculation of conditional expectations of truncated and shifted intermolecular potential energy terms $\langle \mathbf{V}|\mathbf{q}\rangle$. (d2) PCA using $-\beta\langle \mathbf{V}|\mathbf{q}\rangle$ (IPA). (e1) Calculation of products of intermolecular forces and conditional cumulative densities defined in Eq. (33). (e2) FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$ (DIPA).

performing PCA using $-\beta\mathbf{V}(\mathbf{q})$ [Fig. 2(b)]. To get the concrete expression for $\rho(\mathbf{q}'|\mathbf{q})$, we divide the potential energy of the whole system $V(\mathbf{q},\mathbf{q}')$ into three terms: $V(\mathbf{q},\mathbf{q}') = V_{\mathrm{T}}(\mathbf{q}) + V_{\mathrm{TE}}(\mathbf{q},\mathbf{q}') + V_{\mathrm{E}}(\mathbf{q}')$. The terms $V_{\mathrm{T}}(\mathbf{q})$, $V_{\mathrm{TE}}(\mathbf{q},\mathbf{q}')$, and $V_{\mathrm{E}}(\mathbf{q}')$ are the target molecular term, target-environmental term, and environmental term, respectively. Then the distribution $\rho(\mathbf{q}'|\mathbf{q})$ is given as

$$\rho(\mathbf{q}'|\mathbf{q}) = \frac{\rho(\mathbf{q},\mathbf{q}')}{\int \rho(\mathbf{q},\mathbf{q}')d\mathbf{q}'} = \frac{1}{Z(\mathbf{q})} e^{-\beta[V_{\mathrm{TE}}(\mathbf{q},\mathbf{q}')+V_{\mathrm{E}}(\mathbf{q}')]}. \quad (44)$$

This distribution can be considered to be a canonical distribution of $\mathbf{q}'$ for a given $\mathbf{q}$ and the potential energy $V_{\mathrm{TE}}(\mathbf{q},\mathbf{q}') + V_{\mathrm{E}}(\mathbf{q}')$ at inverse temperature β. Therefore, an MD simulation with the given $\mathbf{q}$ and the potential energy $V_{\mathrm{TE}}(\mathbf{q},\mathbf{q}') + V_{\mathrm{E}}(\mathbf{q}')$ generates $\mathbf{q}'$ following the distribution $\rho(\mathbf{q}'|\mathbf{q})$. We then iterate such MD simulations for various different $\mathbf{q}$, which can be sampled from the original MD simulation [Fig. 2(c)]. After this iteration, we can finally perform the IPA [Fig. 2(d)] and the DIPA [Fig. 2(e)].

III. NUMERICAL RESULTS

A. PEPCA of the alanine dipeptide isomerization in explicit water

For comparison of the IPA and the DIPA, we apply them to the alanine dipeptide isomerization in explicit water. In this system, the target molecule (peptide) has multiple stable states, and the environment molecules (water) have important roles regarding the states via intermolecular (peptide-water) interactions. We first apply PEPCA to understand the contribution of intrapeptide interactions to the peptide conformational distribution. Next, we apply IPA and DIPA to understand the contribution of peptide-water interactions to the peptide conformational distribution.

MD simulations were performed by MD package AMBER 10 [31] with the ff03 force field [32] and TIP3P water [33]. Water molecules were included in a cubic box whose edge length was determined to be the minimum distance between peptide atoms and the faces of the box over 15 Å. As a result, 2192 water molecules were contained within the unit cell. Bonds involving hydrogen in the peptide were constrained by SHAKE [34], and TIP3P water molecules were constrained by SETTLE [35]. A direct-space cutoff distance of 10 Å was used for the Lennard-Jones and electrostatic interactions. Long-range electrostatic interactions were calculated using particle mesh Ewald (PME) [36]. A 100 ps NpT (1 atm and 300 K) MD simulation was performed to equilibrate the water density, which was stationary around 30 ps and resulted in a $(40.6\ \text{Å})^3$ cubic periodic box at 100 ps. Then a 10 ns NVT (300 K) Langevin dynamics simulation with a collision frequency of 1.0 ps^{-1} by integration with a 2 fs time step was performed. The coordinates were saved every 1 ps, and 10 000 snapshots were used for the analysis. From the Ramachandran plot (Fig. 3), we see that the peptide has three stable states around $(\phi,\psi) = (-80°,-20°)$, $(\phi,\psi) = (-80°,150°)$, and $(\phi,\psi) = (-150°,150°)$. We call these states α, PPII, and β based on their corresponding secondary structures of α helix, polyproline II helix, and β sheet, respectively.

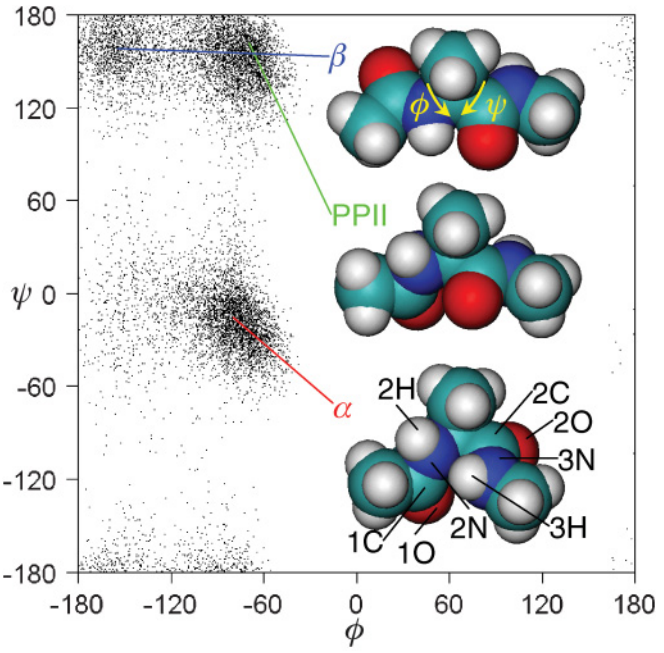

FIG. 3. (Color online) Ramachandran plot of the alanine dipeptide in explicit water. Angles ϕ and ψ are the dihedral angles indicated by the curved arrows. Dots represent the scatter plot of ϕ and ψ (Ramachandran plot). The three-dimensional structures labeled α, PPII, and β are snapshots at 120, 350, and 370 ps, respectively. The structures were rendered by a molecular visualization program (VMD) [37].

TABLE II. Number of the intrapeptide potential energy terms with permutation invariance. The permutation symmetry is relabeling hydrogens in a methyl group. The number of all potential energy terms ignoring the permutation invariance is given in parentheses.

Potential type	No.
Bond	15(21)
Angle	24(36)
Dihedral	31(45)
van der Waals	84(174)
Electrostatic	84(174)
Total	238(450)

After calculating $\mathbf{V}(\mathbf{q})$ with permutation invariance (Table II) for each of the 10 000 snapshots, PEPCA was performed [Fig. 2(b)]. The PCA was implemented by the singular value decomposition (SVD) after centering the data [11]. The top three PEPCA eigenvalues overwhelm the other eigenvalues (Fig. 4). Therefore, we expect that the three corresponding eigenvectors are sufficient to understand the intrapeptide interactions that are important to the peptide conformational distribution. In fact, we can clearly discriminate three stable states by only the first two principal components that are visualized when we plot the components of the first and the second eigenvectors (U_{k1}, U_{k2}) (squares in Fig. 5) and the first and the second principal components $[g_1(\mathbf{q}), g_2(\mathbf{q})]$ (dots in Fig. 5). This plot is known as the biplot [29,30]. As shown in Fig. 5, the first principal components discriminate between the α and the PPII + β states, and the second principal components discriminate between the PPII and the β states. Therefore, we can expect that the intrapeptide interactions that are important to these three states are identified by the first and the second eigenvectors.

The perturbation effects by the eigenvectors summarized in Table I are easily read from the biplot (Fig. 5) as follows. First, we consider the first eigenvectors. If we strengthen the interactions when $U_{k1} < 0$ (left squares in Fig. 5) and weaken when $U_{k1} > 0$ (right squares in Fig. 5), the ratio of the peptide conformation $\mathbf{q}$ will increase when $g_1(\mathbf{q}) < 0$ (left dots in Fig. 5) or decrease when $g_1(\mathbf{q}) > 0$ (right dots in Fig. 5). Similarly, if we strengthen the interactions indicated by the right squares and weaken the left squares, the peptide conformation will increase in the right dots and decrease in the left dots, respectively. We can also apply the same discussion in the second eigenvector by considering the bottom and the top direction in the biplot. Thus, the perturbation effects by the first and the second eigenvectors are systematically understood by the biplot.

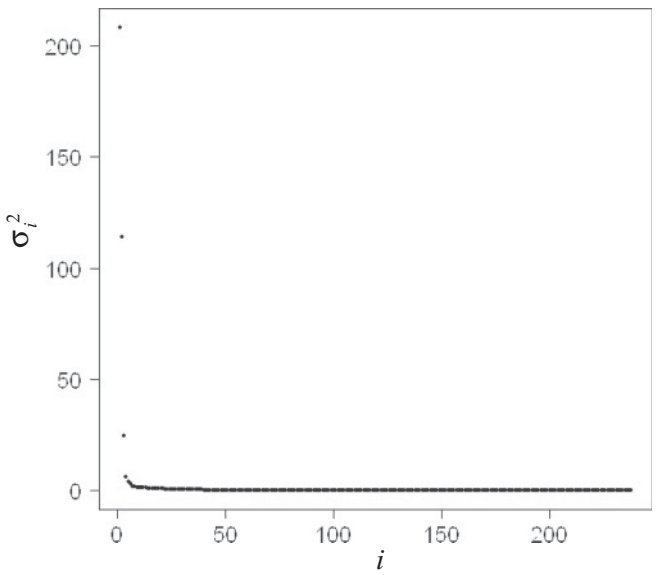

FIG. 4. PEPCA eigenvalues of the alanine dipeptide in explicit water.

By using these biplot properties, we first consider the perturbation effects of the first eigenvector, whose principal components discriminate between the α and the PPII + β states. Smaller negative components of the first eigenvector (left squares in Fig. 5) are the electrostatic interactions 2N-3H, 1C-3N, 1O-3H, and 2N-2O. Among them, 2N-3H, 1C-3N, and 1O-3H are attractive interactions and more favorable for the α state (left dots in Fig. 5). The 2N-2O interaction is repulsive, and more unfavorable to the PPII + β states (right dots in Fig. 5). This can be confirmed in the structure because 2N and 2O atoms are separate in the α state. Larger positive components of the 1st eigenvector (right squares in Fig. 5) are the electrostatic interactions 2H-2O, 1C-3H, and 1O-3N. The interaction 2H-2O is attractive and more favorable for the PPII + β states. The interactions 1C-3H and 1O-3N are repulsive and more unfavorable to the α state. Their atoms are separate in the PPII + β states.

Next, we consider the perturbation effects of the second eigenvector, whose principal components discriminate between the PPII and the β state. Smaller negative components of the second eigenvector (bottom squares in Fig. 5) are the attractive electrostatic interactions 1O-2C and 1C-2O, and more favorable for the PPII state (bottom-right dots in Fig. 5). The larger positive component of the second eigenvector (top squares in Fig. 5) is the repulsive electrostatic interaction 1O-2O. This is more unfavorable to the PPII state, and its atoms are separate in the β state (top-right dots in Fig. 5). In summary, by using the PEPCA biplot we can identify the three peptide states in explicit water by the first and the second principal components and their important intrapeptide interactions by the components of the first and the second eigenvectors.

B. IPA of the alanine dipeptide isomerization in explicit water

We apply the IPA to understand the contribution of the peptide-water interactions to the peptide conformational distribution. To perform MD simulations with fixed peptide coordinates $\mathbf{q}$ [Fig. 2(c)], we used the belly option of the sander program of AMBER 10 [31]. Although the original belly-option code turns off the forces acting on the target molecule derived from the potential energy, the random forces still act on the target molecule in the Langevin dynamics. Therefore, we slightly modified the code to turn off the random forces to fix the target molecule completely. For initial coordinates, we used 1000 conformations from the original 10 ns MD simulation [Fig. 2(a)] with 10 ps intervals. Starting from these initial conformations, 1000 *NVT* (300 K) MD simulations with fixed peptide coordinates $\mathbf{q}$ were performed [Fig. 2(c)]. After

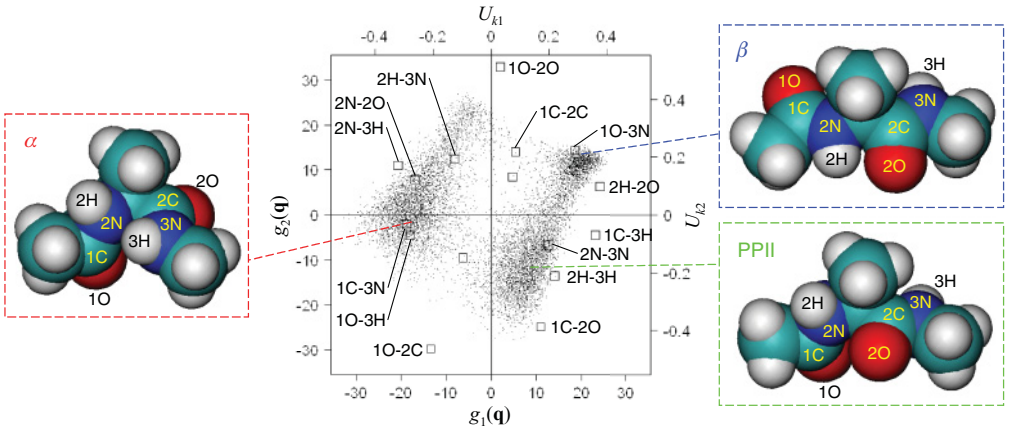

FIG. 5. (Color online) The PEPCA biplot of the alanine dipeptide in explicit water. Dots show the scatter plot of the first and the second principal components $[g_1(\mathbf{q}),g_2(\mathbf{q})]$. Left, bottom-right, and top-right structures represent the snapshot at 120 ps (in the α state), 350 ps (in the PPII state), and 370 ps (in the β state), respectively. Squares show the scatter plot of the top ten components of the first and the second eigenvectors (U_{k1},U_{k2}). The label "1O-3H" indicates the electrostatic interactions between 1O atom and 3H atom shown in structures.

calculating truncated and shifted intermolecular interactions (Table III), we performed the IPA [Fig. 2(d)].

Figure 6 shows the IPA eigenvalues with different cutoff distances and sampling times for the conditional expectations. We can see the eigenvalues decrease and converge by increasing the sampling time. Therefore, the IPA surely gives the definite result with sufficient sampling time for the conditional expectations. Figure 7 shows the IPA eigenvalues with different values for r_c. The eigenvalues converge for 11 Å $\leqslant r_c \leqslant$ 15 Å except the third eigenvalue. Figure 8 shows the IPA biplots using different values for r_c. For $r_c = 2$ to 4 Å, states can not be identified from the principal components. For $r_c = 5$ to 6 Å, the first principal components roughly discriminate the α state and PPII + β states. For $r_c \geqslant$ 7 Å, the separation of the α and PPII + β states becomes clearer. However, the separation of the PPII state and the β state is unclear. Therefore, we cannot discuss important intermolecular interactions to discriminate the PPII state and the β state.

We now identify the important intermolecular interactions to the α state and PPII + β states by using the biplot with $r_c =$ 15 Å. Smaller negative components of the first eigenvector [left squares in Fig. 8 ($r_c = 15$ Å)] are 3H-H, 3N-O, 1C-H, 1O-O, 2O-H, and 2H-H electrostatic interactions, where -H and -O indicate interactions with the hydrogen and oxygen atoms of water, respectively. Because 3H-H, 3N-O, 1C-H, 1O-O, and 2H-H electrostatic interactions are repulsive, they represent more unfavorable interactions for the PPII + β states than the α state. The attractive interaction 2O-H prefers the α state to the PPII + β states. Larger positive components [right squares in Fig. 8 ($r_c = 15$ Å)] are 3H-O, 1O-H, 3N-H, 2N-H, 1C-O, 2C-O, and 2O-O electrostatic interactions. Attractive electrostatic interactions 3H-O, 1O-H, 3N-H, 2N-H, 1C-O, and 2C-O prefer the PPII + β states to the α state. The repulsive electrostatic interaction 2O-O is a more unfavorable interaction for the α state than the PPII + β states.

TABLE III. Number of peptide-water potential energy terms with permutation invariance. The permutation symmetries consist of relabeling hydrogens in a methyl group, hydrogens in a water molecule, and water indexes. The symbols O and H in the left column indicate interactions with oxygen or hydrogen atoms of water. Van der Waals interactions about hydrogens in water are not incorporated in the TIP3P water model [33].

Potential type	No.
van der Waals (O)	16
Electrostatic (H)	16
Electrostatic (O)	16
Total	48

C. DIPA of the alanine dipeptide isomerization in explicit water

To incorporate the distance information systematically, we apply DIPA to the alanine dipeptide in explicit water. The simple method to implement FPCA using $-\beta\mathbf{f}(r|\mathbf{q})$ (DIPA) is to perform PCA with discretized functional data, as described in Appendix C. We discretize the interval $[0,r_c]$ with N_B bins as

$$r_l \equiv l\Delta r, \quad \Delta r \equiv r_c/N_B, \quad l = 0,\ldots,N_B. \tag{45}$$

To obtain $\mathbf{f}(r_l|\mathbf{q})$ [Eq. (33)], we estimate the force $F_k(r_l)$ and the conditional cumulative density $N_{iJ_k}(r_l|\mathbf{q})$. The force $F_k(r_l)$ can be calculated from the force field used. The conditional cumulative density $N_{iJ_k}(r_l|\mathbf{q})$ must be estimated from the MD simulations that fix the coordinates $\mathbf{q}$ of the target molecule [Fig. 2(c)], which are identical to the MD simulations used to perform IPA. We can estimate $N_{iJ_k}(r_l|\mathbf{q})$ by averaging the number of J_k atoms within r_l of the ith atom. The PCA was implemented by performing a SVD after centering the discretized data [11]. By using the correspondence between the PCA and the FPCA that is given in Table VII, we obtain the eigenfunctions, eigenvalues, and principal components of DIPA.

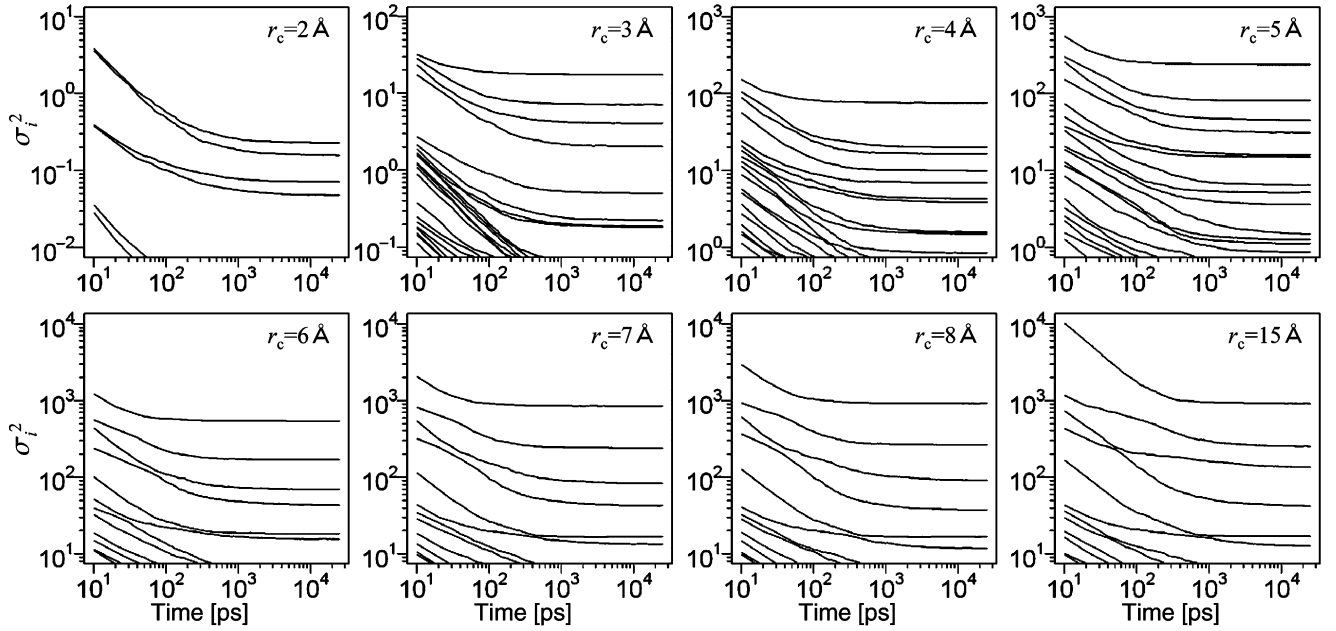

FIG. 6. Dependence of the IPA eigenvalues of alanine dipeptide in explicit water on the cutoff distance r_c and the sampling time. IPAs were performed from 10 ps (10 sample) to 25 ns (25 000 sample) averages for the conditional expectations $\langle \mathbf{V}|\,\mathbf{q}\rangle$.

We investigate the convergence of DIPA by varying the conditions for the conditional cumulative densities. First, we check the dependence of the DIPA eigenvalues on the sampling time to estimate the conditional cumulative densities (Fig. 9). Upon increasing the sampling time, the eigenvalues decrease and converge. In particular, the top three eigenvalues converge within 1 ns. This convergence rate is faster than that for the IPA using $r_c = 15$ Å (Fig. 6). Second, we investigate the dependence of the DIPA eigenvalues on the cutoff distance r_c (Fig. 10). Upon increasing r_c, the eigenvalues increase and converge for $r_c \geqslant 10$ Å in DIPA. We attribute this monotonically increasing property in DIPA to the fact that a distance-dependent perturbation with a longer cutoff distance includes shorter distances, so that a larger change in the conformational distribution can be induced. Since DIPA converges for $r_c \geqslant 10$ Å, the cutoff value $r_c = 15$ Å is sufficient for further analysis. Finally, we checked the dependence of the DIPA eigenvalues on the stride Δr of the interval $[0,r_c]$ for the conditional cumulative densities (Fig. 11). The top eigenvalues are robust against variation in Δr. In particular, for $\Delta r \leqslant 0.2$ Å, the eigenvalues are almost identical. In summary, allowing sufficient sampling time, a longer cutoff distance r_c, and a small stride Δr for conditional cumulative densities, DIPA gives definite results.

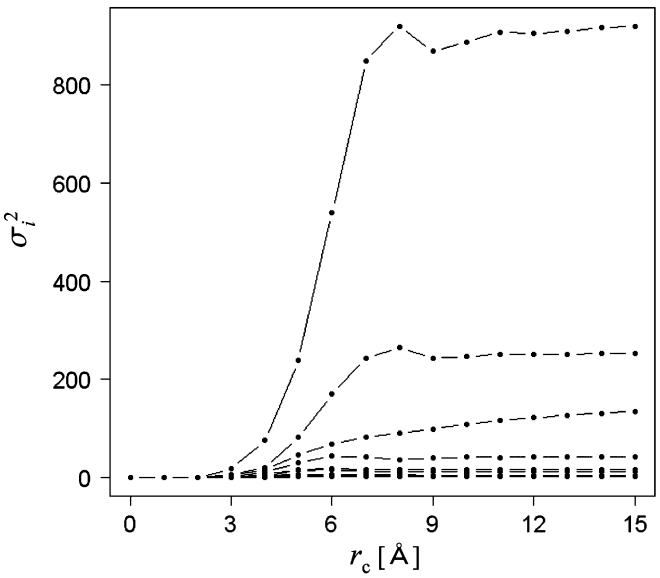

FIG. 7. Dependence of the IPA eigenvalues of alanine dipeptide in explicit water on the cutoff distance r_c. The conditional expectations were estimated from 25 ns (25 000 sample) averages.

We analyze in detail the converged result of DIPA. Because the top three eigenvalues converge within 1 ns (Fig. 9), we use the results of DIPA with 1 ns MD simulations in the following analysis. The first principal components discriminate between the α and the PPII + β states, and the second principal components discriminate between the PPII and the β states (dots in Fig. 12). Thus, the top two eigenfunctions suffice to identify important peptide-water interactions for the three states. Next, we consider the dependence on distance, which cannot be directly interpreted from the IPA biplot. By using Eqs. (37) and (42), the normality of the ith DIPA eigenfunction is represented as

$$\sum_{k=1}^{M} \int_0^{r_c} U_{ki}(r)^2 dr = 1. \tag{46}$$

By using Eq. (46), we can quantify the contribution of the distance r to the ith eigenfunction by introducing

$$S_i(r) \equiv \sum_{k=1}^{M} U_{ki}(r)^2 \geqslant 0, \tag{47}$$

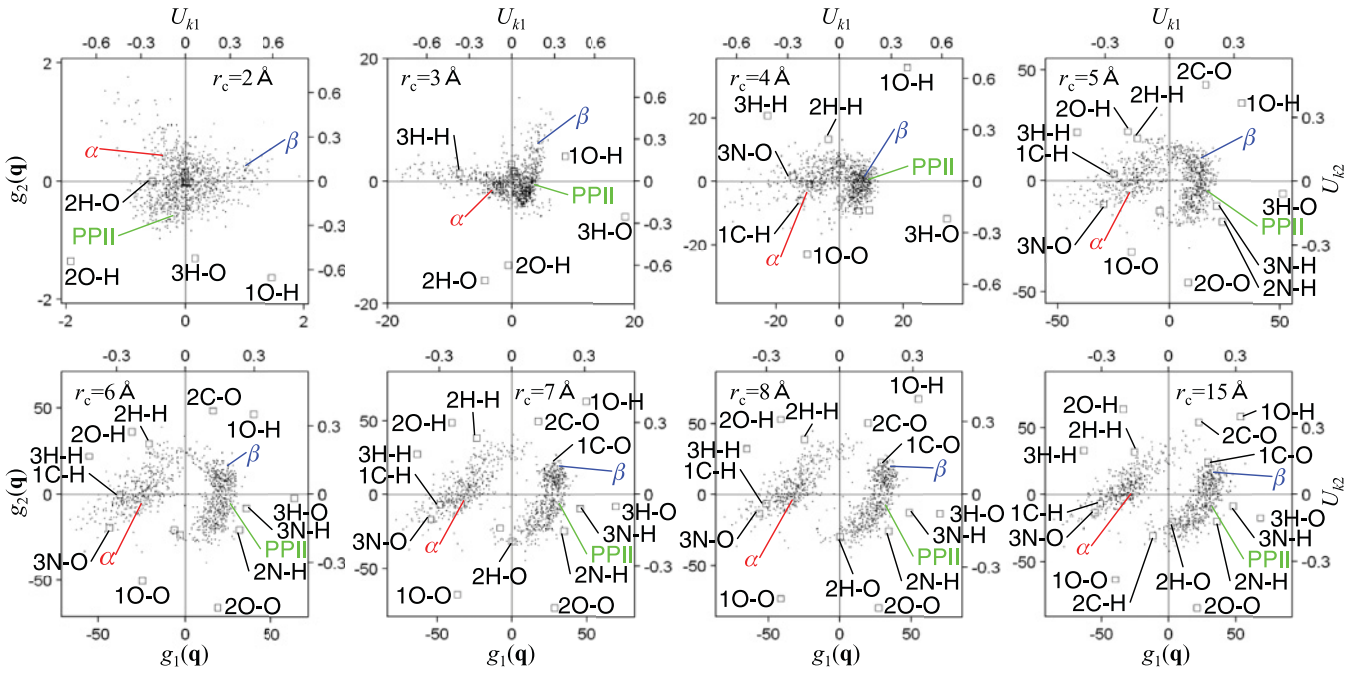

FIG. 8. (Color online) Dependence of the IPA biplot of the alanine dipeptide in explicit water on the cutoff distance r_c. The conditional expectations are estimated from 25 ns MD simulations (25 000 sample). Directions of the eigenvectors and corresponding signs of the principal components are determined to be consistent with the PEPCA biplot (Fig. 5). The meaning of biplots is same as in Fig. 5. Labels α, PPII, and β indicate principal components at identical snapshots in Fig. 5. The label "A-H" ("A-O") indicates electrostatic interactions between dipeptide atom A (shown on structures of Fig. 5) and hydrogen (oxygen) atoms of water molecules. Components of eigenvectors (squares) are selected from the union of the top ten components of the first and the second eigenvectors.

which satisfy

$$\int_0^{r_c} S_i(r)\,dr = 1. \tag{48}$$

For $r \geqslant 10$ Å, $S_1(r)$ and $S_2(r)$ are almost zero (Fig. 13). This indicates that long-range intermolecular interactions over 10 Å have no preference with respect to the three states. These results also explain the convergence of the top two eigenvalues for $r_c \geqslant 10$ Å in Fig. 10. Although

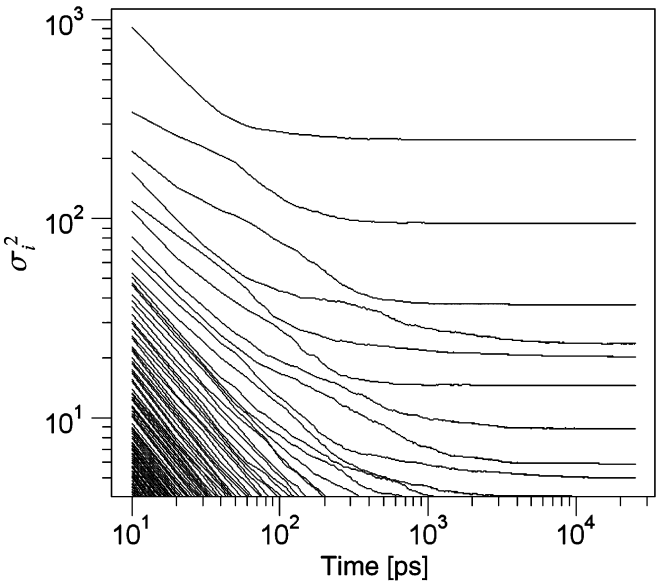

FIG. 9. Dependence of the DIPA eigenvalues of the alanine dipeptide in explicit water on the sampling time for conditional cumulative densities ($r_c = 15$ Å and $\Delta r = 0.05$ Å). DIPA was performed every 10 ps from 10 ps to 25 ns MD simulations.

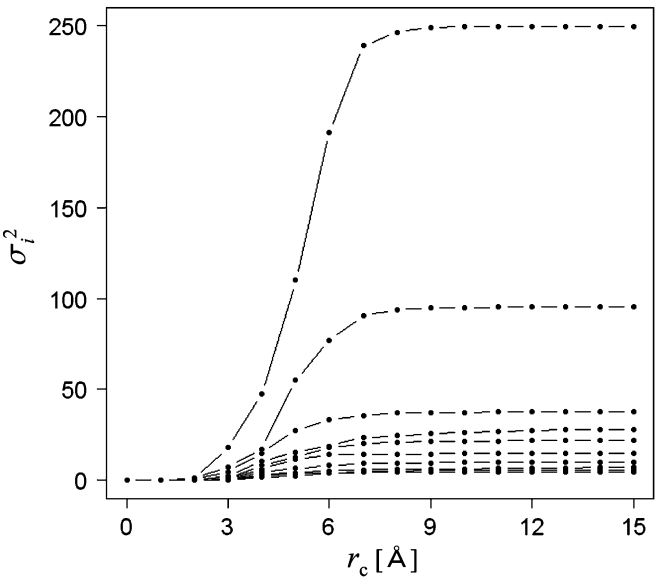

FIG. 10. Dependence of the DIPA eigenvalues of the alanine dipeptide in explicit water on the cutoff distance r_c. Conditional cumulative densities are estimated from 1 ns MD simulations ($\Delta r = 0.05$ Å).

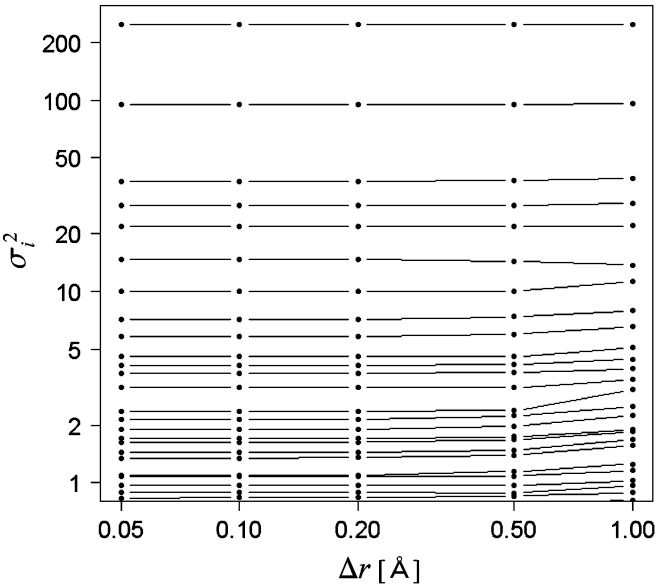

FIG. 11. Dependence of the DIPA eigenvalues of the alanine dipeptide in explicit water on the stride Δr of the interval $[0,r_c]$. Conditional cumulative densities are estimated from 1 ns MD simulations ($r_c = 15$ Å).

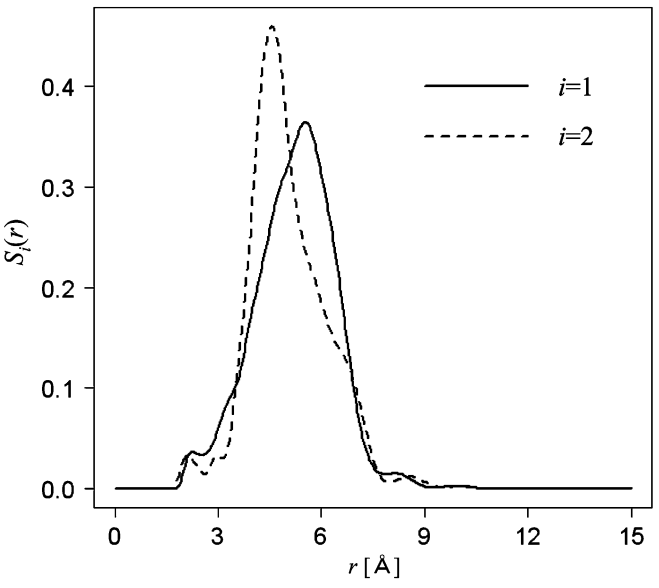

FIG. 13. Contribution of the distance r to the first and the second DIPA eigenfunctions [Eq. (47)] of the alanine dipeptide in explicit water. Conditional cumulative densities are estimated from 1 ns MD simulations ($r_c = 15$ Å and $\Delta r = 0.05$ Å).

long-range intermolecular interactions cause the slow convergence in IPA (Fig. 6), these smaller contributions of long-range interactions are considered to be responsible for the fast convergence in DIPA. The contributions $S_1(r)$ and $S_2(r)$ are maximized at $r = 5.55$ Å and $r = 4.55$ Å, respectively. Therefore, hydration shell around these distances has an important role to determine the preference among the three states.

Since components of the PEPCA or IPA eigenvectors are represented as points (squares in Figs. 5 and 8), eigenvector components are compactly plotted in a biplot. However, components of DIPA eigenfunctions are represented as curves

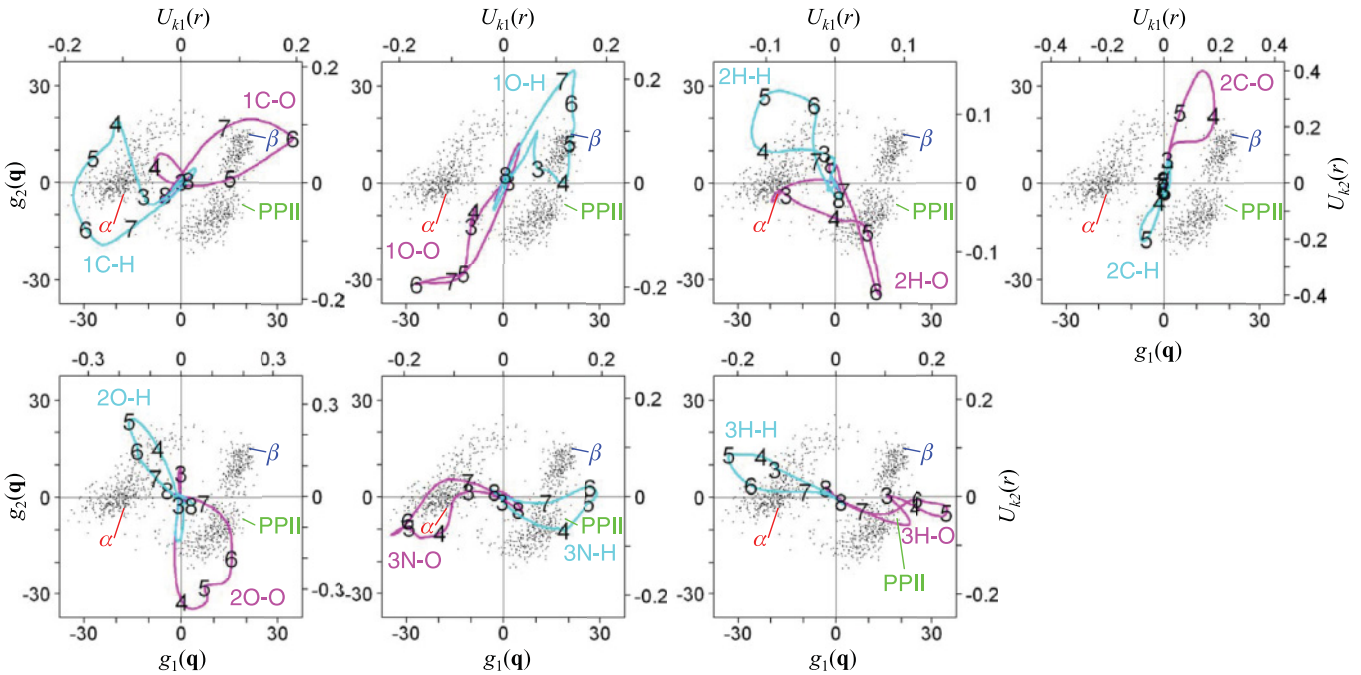

FIG. 12. (Color online) DIPA biplots of the alanine dipeptide in explicit water. Conditional cumulative densities are estimates from 1 ns MD simulations ($r_c = 15$ Å and $\Delta r = 0.05$ Å). Dots show the scatter plot of the first and the second principal components $[g_1(\mathbf{q}), g_2(\mathbf{q})]$. Labels α, PPII, and β indicate principal components at identical snapshots in Fig. 5. The curves are components of the first and the second DIPA eigenfunctions $[U_{k1}(r), U_{k2}(r)]$ (0 Å $\leqslant r \leqslant$ 15 Å). Label "A-O" ("A-H") indicates that the curve shows the electrostatic interactions between A atom of the peptide (Fig. 5) and oxygen (hydrogen) atoms of water molecules. Numbers on the curves represent r in units of Å.

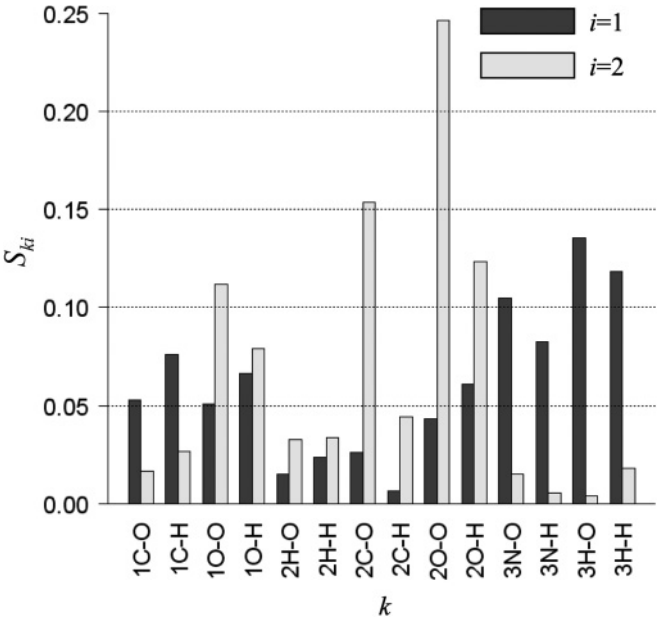

FIG. 14. Contribution of the kth component to the first and the second DIPA eigenfunctions [Eq. (49)] of the alanine dipeptide in explicit water. Conditional cumulative densities are estimated from 1 ns MD simulations ($r_c = 15$ Å and $\Delta r = 0.05$ Å). Components are selected from the union of the top nine components of the first and the second eigenfunctions. The significance of the labels (such as "1C-O") is identical to that used in Fig. 12.

because the component of DIPA eigenfunctions depends on distance. Therefore, representing all components in a biplot is complicated. To circumvent this, we select some components that are large contributors to the eigenfunction and plot them on different biplots. By using Eq. (46), we may quantify the contribution of the kth component to the ith eigenfunction by introducing

$$S_{ki} \equiv \int_0^{r_c} U_{ki}(r)^2 dr \geqslant 0, \tag{49}$$

which satisfy

$$\sum_{k=1}^{M} S_{ki} = 1. \tag{50}$$

Figure 14 shows S_{ki} with large values for the first and the second eigenfunctions.Then, we represent these top components by biplots (Fig. 12). First, the attractive interaction that prefers the α state to the PPII + β states is 2O-H electrostatic interaction. Second, the attractive interaction that prefers the PPII state to the β state is 2H-O electrostatic interaction. Finally, the attractive interactions that prefer the β state to the PPII state are 1C-O, 1O-H, 2C-O, and 2O-H electrostatic interactions. By comparing components of the IPA eigenvectors for $r_c = 15$ Å [Fig. 8 ($r_c = 15$ Å)] with components of the DIPA eigenfunctions (Fig. 12), we see that they correspond. For example, 1C-H components point to the left and 1C-O components point to the right for both IPA and DIPA. Thus, we can confirm the consistency between the IPA and the DIPA.

D. PEPCA of the chignolin folding in explicit water

To test the feasibility of the DIPA for larger molecules, we use the chignolin folding in explicit water. The chignolin is a ten-residue peptide (Gly1-Tyr2-Asp3-Pro4-Glu5-Thr6-Gly7-Thr8-Trp9-Gly10) that folds into β hairpin (Protein Data Bank (PDB) ID 1UAO) [38]. In previous studies [39–41], folding simulations of the chignolin were successfully performed. The protocol of the following MD simulation is identical to that of the alanine dipeptide (Sec. III A). The first structure in 1UAO was used for the initial coordinates of the chignolin. Water molecules were included in a cubic box whose edge length was determined to be the minimum distance between peptide atoms and the faces of the box over 20 Å. As a result, 6437 water molecules were contained within the unit cell. To neutralize the whole system, 12 Na^+ and 10 Cl^- ions were added. Małolepsza *et al*. [27] pointed out that some improper torsion potentials of AMBER force field are not invariant to exchange of symmetrical atoms and these could be avoided by reordering the atoms for the torsion angle. We applied their program [42] to the topology file. This reordered atoms of improper torsions of Tyr2, Asp3, Glu5, and C-terminal of Gly10. After energy minimization and 100 ps NpT MD simulation at 1 atm and 312 K (melting temperature) [38], the box size became $(58.4\ \text{Å})^3$. Then, a 1 μs NVT (312 K) Langevin dynamics simulation with a collision frequency 1.0 ps^{-1} by integration with a 2 fs time step was performed. The root mean square deviation (RMSD) of the chignolin (Fig. 15) shows native to non-native transitions three times.

After calculating the potential energy terms with permutation invariance (Table IV) with 1 ns interval (1000 structures), the PEPCA was performed. The PCA was implemented by the SVD after centering the potential energy terms. Two large gapped eigenvalues are observed in Fig. 16. Therefore, we can expect that the top two principal components identify conformational states and their corresponding eigenvectors can identify important intramolecular interactions to each

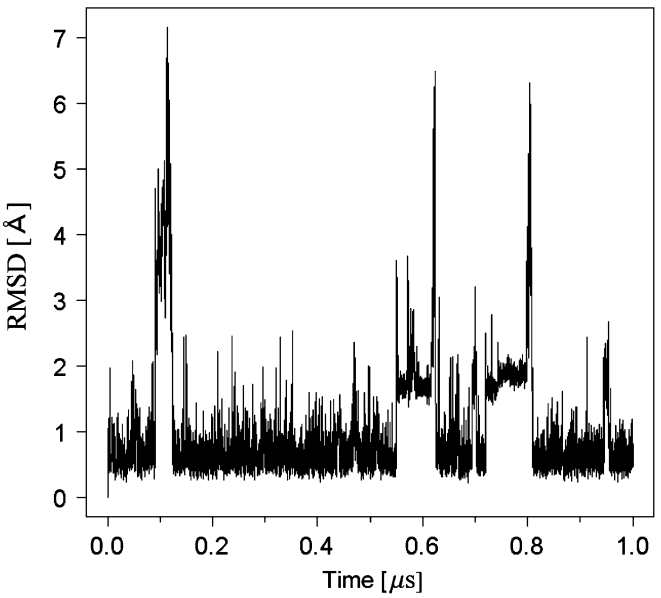

FIG. 15. RMSD of C_α atoms from the first structure of the chignolin in explicit water.

TABLE IV. Number of the chignolin potential energy terms with permutation invariance. The number of all potential energy terms ignoring the permutation invariance is given in parentheses.

Potential type	No.
Bond	116 (141)
Angle	204 (249)
Dihedral	313 (403)
van der Waals	6147 (9063)
Electrostatic	6147 (9063)
Total	12 927 (18 919)

state. Figure 17 shows the PEPCA principal components. By comparing Fig. 15, we can see that the first principal components correspond to the native and non-native transitions. The second principal components indicate that the non-native structures around 0.1 μs are different from the non-native structures around 0.6 and 0.75 μs.

We analyze the conformational states by using the PEPCA biplot (Fig. 18). In the left side of the biplot [$g_1(\mathbf{q}) < 0$], dense dots are observed. By comparing Figs. 15 and 17, these dots correspond to the native structures. A snapshot from the state also confirms the native β-hairpin structure (dashed box labeled N in Fig. 18). We refer to the state as the "N" state (native state). Dots in the right side of the biplot [$g_1(\mathbf{q}) > 0$] corresponds to the non-native state. In the bottom-right [$g_1(\mathbf{q}) > 0$ and $g_2(\mathbf{q}) < 0$] of the biplot, we can see two dense dots. Structures from these states (dashed boxes labeled M1 and M2 in Fig. 18) share the same backbone structures. Although the M1 structure has no contact between N- and C-terminal residues, the M2 structure has contact between them. By comparing the backbone of the native β hairpin, these structures can be recognized as misfolded structures. Therefore, we refer to these states as the "M1" state and the "M2" state (misfolded state), respectively. In the top-right [$g_1(\mathbf{q}) > 0$ and $g_2(\mathbf{q}) > 0$] of the biplot, we can see scattered dots. A structure from these dots (dashed box labeled U in Fig. 18) has unfolded backbone structure. We refer to this

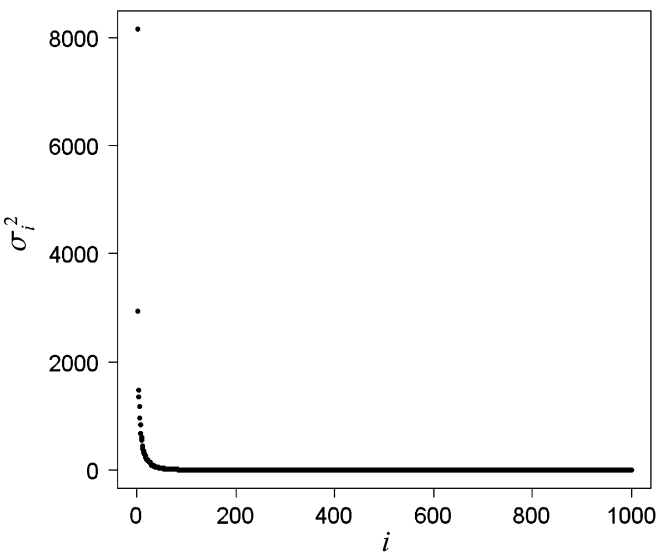

FIG. 16. The PEPCA eigenvalues of the chignolin in explicit water.

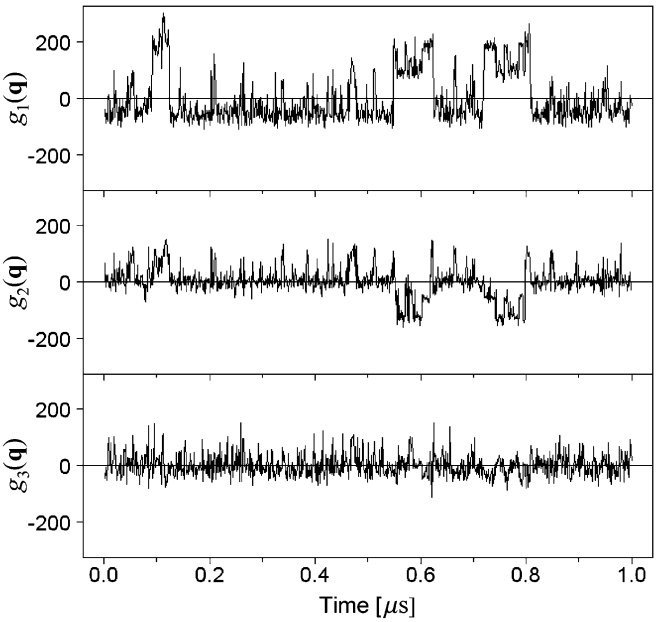

FIG. 17. The top three PEPCA principal components of the chignolin in explicit water.

state as the "U" state (unfolded state). Thus, we can identify the four states, that are the native (N) state [$g_1(\mathbf{q}) < 0$], two misfolded (M1 and M2) states [$g_1(\mathbf{q}) > 0$ and $g_2(\mathbf{q}) < 0$], and the unfolded (U) state [$g_1(\mathbf{q}) > 0$ and $g_2(\mathbf{q}) > 0$] in the PEPCA biplot.

Next we identify preferable intramolecular interactions to each state by the PEPCA biplot (Fig. 18). First, we consider the interactions that prefer the N state to the non-native state. These interactions are located on the left side of the biplot ($U_{k1} < 0$). We can see that the bottom-left ($U_{k1} < 0$ and $U_{k2} < 0$) components are attractive electrostatic interactions for the N- and the C-terminal residues except 1N-10O repulsive electrostatic interaction. We can confirm 1H-10O and 1O-10H contacts in the N structure. The top-left ($U_{k1} < 0$ and $U_{k2} > 0$) components are 3OD-6HG1, 6OG1-8HG1, 3H-8O, and 3OD-8HG1 electrostatic interactions. These are attractive electrostatic interactions among Asp3, Thr6, and Thr8 residues. We can also confirm 3OD-6HG1, 6OG1-8HG1, and 3H-8O atoms contact in the N structure.

Second, we identify important interactions to the M1 and the M2 states. The N state and the M1 + M2 states are discriminated by the bottom-right direction. As shown in Appendix D, we can understand a perturbation effect of the linear combination of the first and the second eigenvectors from the biplot. This indicates that the favorable interactions for the M1 and the M2 states can be found in the bottom-right direction. In the bottom-right direction, there are repulsive electrostatic interactions among Asp3, Thr6, and Thr8 residues that destabilize the native state. Attractive electrostatic interactions in the bottom-right directions are 3H-7O, 1O-9H, and 2C-7O. From the M1 and the M2 structures in Fig. 18, we can see 3H-7O and 1O-9H atoms contact and these contacts are not observed in the N structure. The M1 state and the M2 state are discriminated by the top-right and bottom-left directions. The attractive electrostatic interaction 1H-10O prefers the M2 state

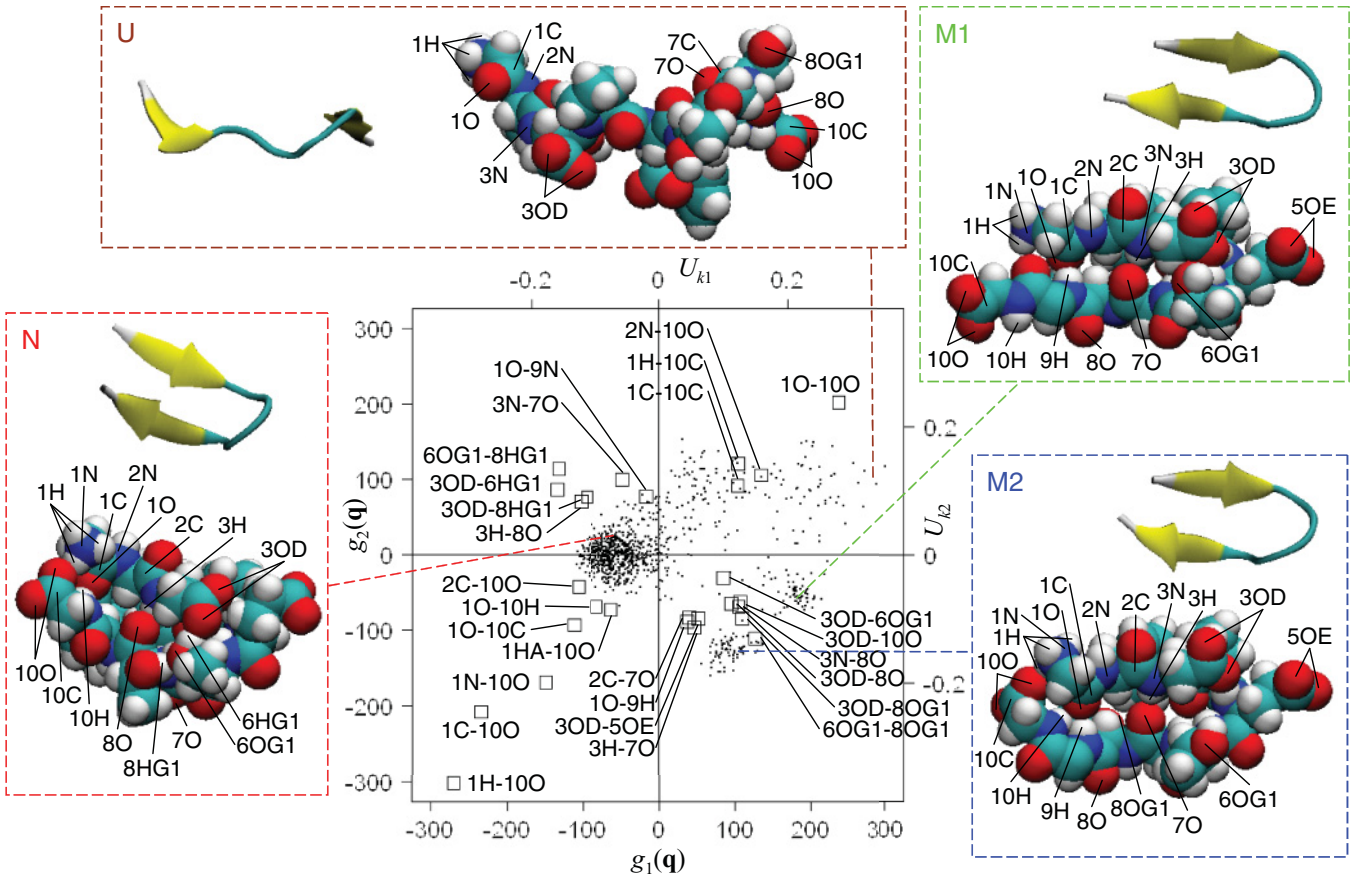

FIG. 18. (Color online) The PEPCA biplot of the chignolin in explicit water. Dots show the scatter plot of the first and the second principal components $[g_1(\mathbf{q}), g_2(\mathbf{q})]$. Squares show the scatter plot of the components of the first and the second eigenvectors (U_{k1}, U_{k2}). Top 20 components of the first and the second eigenvectors are shown. Labels such as 1H-10O indicate the electrostatic interaction between the 1H atom and the 10O atom in the dashed boxes. Structures represent the snapshot at 0.68 μs (in the N state), 0.73 μs (in the M1 state), 0.75 μs (in the M2 state), and 0.114 μs (in the U state), respectively. In M1 and M2 structures, the side-chain of Trp9 is not drawn.

to the M1 state. We can confirm that there is 1H-10O contact in the M2 structure and there is not in the M1 structure.

Finally, we identify favorable interactions for the U state. Components in the top-right direction are 1O-10O, 2N-10O, 1H-10C, and 1C-10C repulsive electrostatic interactions. Since the U state does not have stable intramolecular interactions, these results are reasonable. In summary, we can identify the four states (N, M1, M2, and U) by the top two principal components and their important intramolecular interactions by corresponding eigenvectors.

E. DIPA of the chignolin folding in explicit water

To identify important intermolecular interactions, we apply the DIPA to the chignolin folding in explicit water. The protocol to perform the DIPA is identical to that of the alanine dipeptide (Sec. III C). For initial coordinates, we used 1000 conformations from the original 1 μs MD simulation [Fig. 2(a)] with 1 ns intervals. Starting from these initial conformations, 1000 *NVT* (312 K) MD simulations with fixed chignolin coordinates $\mathbf{q}$ were performed [Fig. 2(c)]. Conditional cumulative densities were calculated using $r_c = 25$ Å and $\Delta r = 0.05$ Å. The number of the chignolin-solvent interaction with permutation invariance is $M = 798$ (Table V) and the number of bins for the conditional cumulative densities is $N_B = r_c/\Delta r = 500$. The PCA variable number for the FPCA (Appendix C) is $p = MN_B = 399\,000$ and the sample size is $n = 1000$. Since the $p \times n$ data matrix is large, the SVD implementation for the PCA is difficult to apply. Instead of the SVD, we implemented the PCA by diagonalizing an $n \times n$ centered Gram matrix of the data matrix [18,43].

There are three large DIPA eigenvalues and they are almost converged within 100 ps MD simulations (Fig. 19). Therefore, we can expect that the top three DIPA eigenfunctions identify important chignolin-solvent interactions. Interestingly, the convergence rate is faster than that of the alanine dipeptide (Fig. 9). We discuss this point in Sec. IV.

TABLE V. Number of chignolin-solvent potential energy terms with permutation invariance.

Potential type	No.
van der Waals (Na^+, Cl^-, O)	114 × 3
Electrostatic (Na^+, Cl^-, H, O)	114 × 4
Total	798

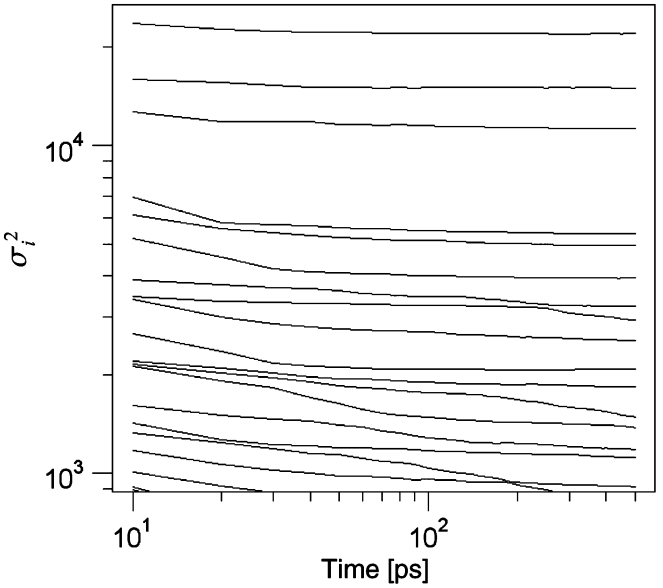

FIG. 19. Dependence of the DIPA eigenvalues of the chignolin in explicit water on the sampling time for conditional cumulative densities ($r_c = 25$ Å and $\Delta r = 0.05$ Å). DIPA was performed every 10 ps from 10 to 500 ps sampling time.

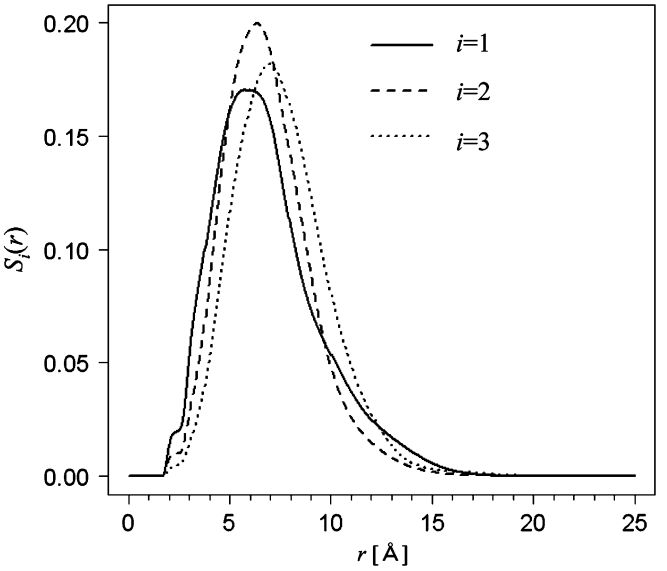

FIG. 21. Contribution of the distance r to DIPA eigenfunctions [Eq. (47)] of the chignolin in explicit water. Conditional cumulative densities are estimated from 100 ps MD simulations ($r_c = 25$ Å and $\Delta r = 0.05$ Å).

The top two principal components (Fig. 20) exhibit similar dynamics to that of the PEPCA principal components (Fig. 17). This indicates that the four states (N, M1, M2, and U) can be identified by the top two principal components and their important chignolin-solvent interactions can be identified by the corresponding eigenfunctions. The third principal components discriminate between the U state and the other states. As shown in Fig. 17, the third PEPCA principal components do not show this discrimination. Since the U state is exposed to the solvent compared with the other folded three states (N, M1, and M2), the third eigenfunction is expected to represent the chignolin-solvent interactions that stabilize these exposed atoms. We confirm this later by using biplots.

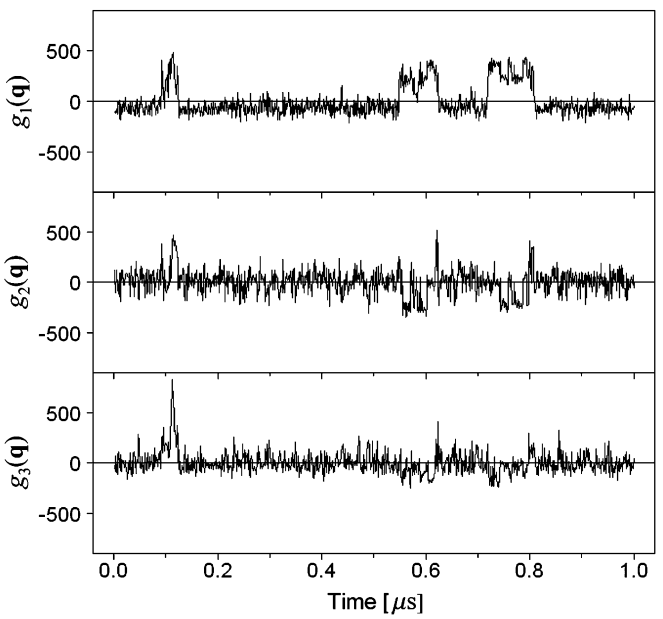

FIG. 20. Top three DIPA principal components of the chignolin in explicit water. Conditional cumulative densities are estimated from 100 ps MD simulations ($r_c = 25$ Å and $\Delta r = 0.05$ Å).

The first and the second DIPA eigenfunctions are 0 for $r \geqslant 20$ Å (Fig. 21). Therefore, the cutoff value $r_c = 25$ Å is sufficient to converge the DIPA. The eigenfunctions have large value for 5 Å $\leqslant r \leqslant 8$ Å. This suggests that the solvation shell around these distances have different contribution to each state.

To identify important intermolecular interactions to each state, we select the large contributors to each eigenfunction (Fig. 22). Based on S_{ki}, we show biplots using the first and the second or the third principal components (Fig. 23). First, the attractive interaction that prefers the N state to the M1 + M2 states is 7O-H electrostatic interaction. From structures in Fig. 18, we can see that 7O atom is exposed to the solvent in the N state and it contacts 3H atom in the M1 + M2 states. Second, the attractive interaction that prefers the M1 + M2 states to the N states is 8O-H electrostatic interaction. The 8O atom contacts 3H in the N state and it is exposed to the solvent in the M1 + M2 states. Third, the attractive interactions that prefer the M1 state to the M2 state are 10C-O and 10O-H electrostatic interactions. The 10O atom contacts 1H in the M2 state and it is exposed to the solvent in the M1 state. Finally, the attractive interaction that prefers the U state to the folded states (N, M1, and M2) is 3OD-H electrostatic interaction. The 3OD atoms contact the peptide atoms in the N, M1, and M2 states and they are exposed to the solvent in the U state.

In summary, the DIPA identify the four states (N, M1, M2, and U) by the top three principal components and their important intermolecular interactions by their corresponding eigenfunctions. Although we only use the information about intermolecular interactions, the DIPA can correctly identify the chignolin conformational states. This can be possible because some atoms in the chignolin expose to the solvent with state dependence.

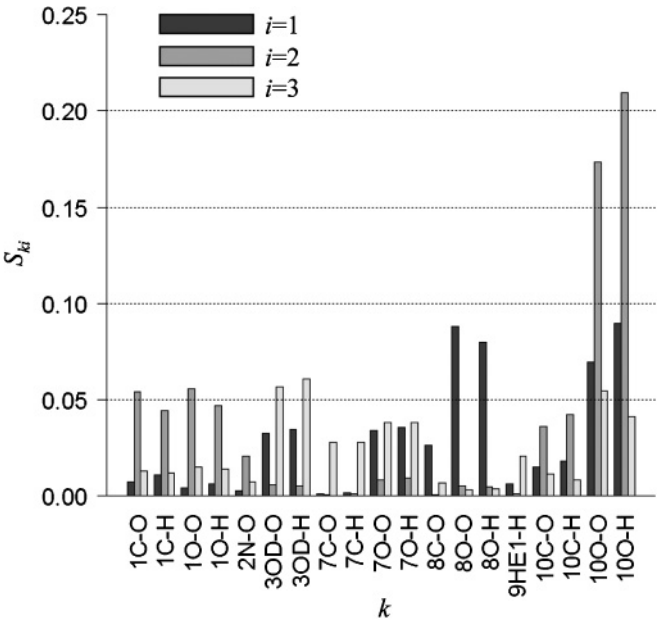

FIG. 22. Contribution of components to DIPA eigenfunctions [Eq. (49)] of the chignolin in explicit water. Conditional cumulative densities are estimated from 100 ps MD simulations ($r_c = 25$ Å and $\Delta r = 0.05$ Å). Components are selected from the union of the top nine components of the first, the second, and the third eigenfunctions. Label A-O (A-H) represents the electrostatic interactions between the atom A of the chignolin (dashed boxes in Fig. 18) and oxygen (hydrogen) atoms of water molecules.

IV. DISCUSSION AND CONCLUSIONS

To identify conformational states of the target molecule and intermolecular interactions that are important for each state, we developed perturbation analyses of intermolecular interactions. We show (i) distance-independent and (ii) distance-dependent perturbation analyses can be realized by performing (i) a PCA using conditional expectations of truncated and shifted intermolecular potential energy terms and (ii) a FPCA using products of intermolecular forces and conditional cumulative densities. We refer to these analyses as IPA and DIPA, respectively.

For comparison of the IPA and the DIPA, we applied them to the alanine dipeptide in explicit water. Although the first IPA principal components identify the α state and the PPII + β states for larger cutoff length, the separation of the PPII state and the β state is unclear in the second IPA principal components. DIPA eigenvalues converge with a sufficient sampling time, longer cutoff distance r_c, and smaller stride Δr of the interval $[0, r_c]$ for the conditional cumulative densities. With this convergence condition, DIPA identifies three peptide states from the top two eigenfunctions. The separation between the PPII and the β state is clearer than that for IPA. Although long-range peptide-water interactions cause the slow convergence in IPA, they have small contributions to the top two eigenfunctions in DIPA. This fact is considered to be responsible for the fast convergence of DIPA. Thus, DIPA improves IPA by introducing a dependence on distance.

To show the feasibility of the DIPA for larger molecules, we applied the DIPA to the ten-residue chignolin folding in explicit water. Interestingly, the DIPA converges faster

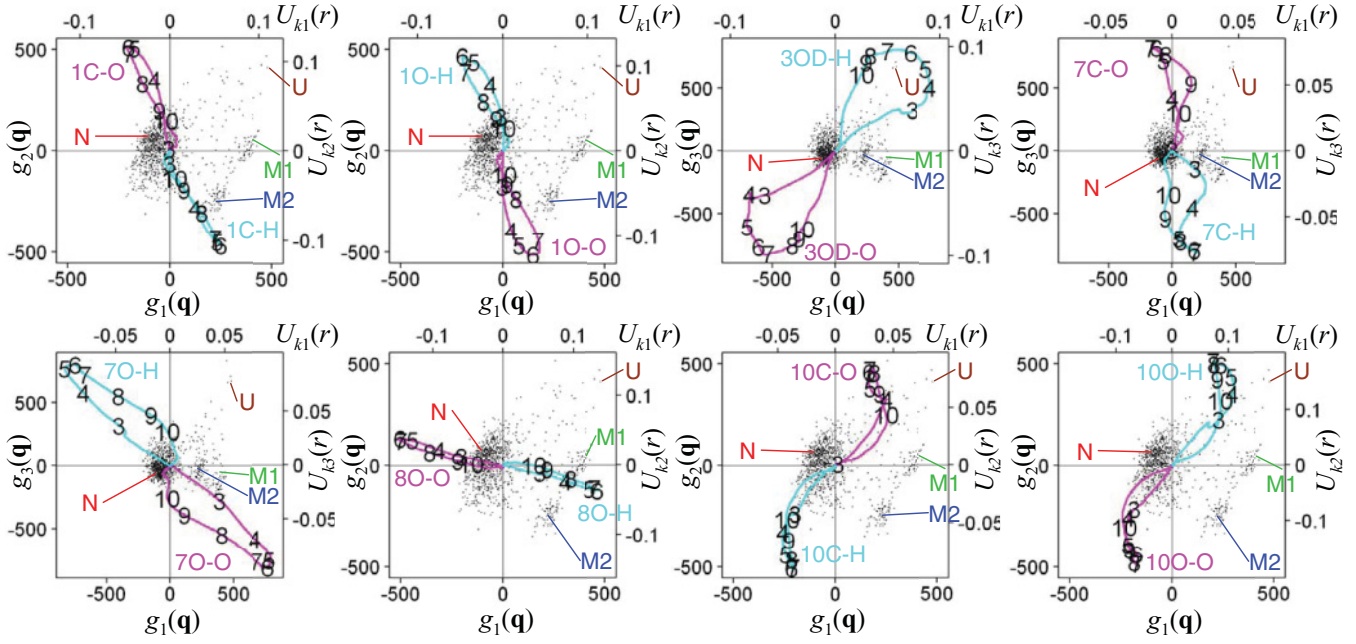

FIG. 23. (Color online) DIPA biplots of the chignolin in explicit water. Conditional cumulative densities are estimated from 100 ps MD simulations ($r_c = 25$ Å and $\Delta r = 0.05$ Å). Dots show the scatter plot of the first and the second or the third DIPA principal components. Labels N, M1, M2, and U indicate principal components at identical snapshots in Fig. 18. The curves are components of the first and the second or the third DIPA eigenfunctions. A label "A-O" ("A-H") indicates that the corresponding curve shows the electrostatic interactions between the A atom of the chignolin (dashed boxes in Fig. 18) and oxygen (hydrogen) atoms of water molecules. Numbers on the curves represent r in units of Å.

than that of the alanine dipeptide. The top three DIPA principal components identify the four states (native state, two misfolded states, and unfolded state) and the corresponding eigenfunctions identify important chignolin-water interactions to each state. Although the DIPA only use information about intermolecular interactions, it can identify the conformational states by recognizing atoms in the chignolin that expose to the solvent with state dependence.

Top two DIPA eigenvalues of the chignolin (Fig. 19) exhibit faster convergence than that of the alanine dipeptide (Fig. 9). To explain this behavior, we analyze the relative error of eigenvalues due to the finite sampling estimation. We assume that the DIPA is performed according to the procedure shown in Appendix C. For convenience, we treat $f_k(r_l|\mathbf{q})$ as a vector $[f_1(r_1|\mathbf{q}), \ldots, f_M(r_{N_B}|\mathbf{q})]^T$ in the following part. We denote the n sample estimator of $f_k(r_l|\mathbf{q})$ as $f_{k,n}(r_l|\mathbf{q})$. When n is large, the multivariate central limit theorem indicates

$$f_{k,n}(r_l|\mathbf{q}) = f_k(r_l|\mathbf{q}) + \frac{1}{\sqrt{n}}\Delta_k(r_l|\mathbf{q}), \tag{51}$$

where $\Delta_k(r_l|\mathbf{q})$ follows the multivariate normal distribution

$$\Delta_k(r_l|\mathbf{q}) \sim N(0, \mathrm{cov}(f_{k,1}(r_l|\mathbf{q})|\mathbf{q})). \tag{52}$$

For simplicity, we assume $f_k(r_l|\mathbf{q})$ and $\Delta_{k'}(r_{l'}|\mathbf{q})$ are uncorrelated and replace $\mathrm{cov}(f_{k,1}(r_l|\mathbf{q})|\mathbf{q})$ with its average $\langle \mathrm{cov}\,(f_{k,1}(r_l|\mathbf{q})|\mathbf{q})\rangle$. Under these assumptions, the covariance of Eq. (51) becomes

$$\begin{aligned}&\mathrm{cov}(-\beta f_{k,n}(r_l|\mathbf{q})\sqrt{\Delta r}) \\ &\quad = \mathrm{cov}(-\beta f_k(r_l|\mathbf{q})\sqrt{\Delta r}) + \frac{1}{n}\langle \mathrm{cov}(-\beta f_{k,1}(r_l|\mathbf{q})\sqrt{\Delta r}|\mathbf{q})\rangle.\end{aligned} \tag{53}$$

With Table VII, we note that the ith eigenvalue of the second covariance matrix in Eq. (53) is σ_i^2. We denote the ith eigenvalue of the first covariance matrix and the average of the third covariance matrix in Eq. (53) as $\sigma_{i,n}^2$ and $\Delta\sigma_i^2$, respectively. By applying Weyl's inequality [28] to Eq. (53), the relative error of the ith eigenvalue is bounded as

$$\frac{1}{n}\frac{\Delta\sigma_{MN_B}^2}{\sigma_i^2} \leqslant \frac{\sigma_{i,n}^2 - \sigma_i^2}{\sigma_i^2} \leqslant \frac{1}{n}\frac{\Delta\sigma_1^2}{\sigma_i^2}. \tag{54}$$

The first inequality in Eq. (54) indicates that finite sampling eigenvalues are overestimated. The second inequality in Eq. (54) shows the relative error is in inverse proportion to the sampling time. These properties qualitatively explain the monotonically decreasing behavior of eigenvalues with respect to the sampling time in Figs. 9 and 19. The second inequality in Eq. (54) also indicates that the sampling time required to achieve some relative error is proportional to $\Delta\sigma_1^2/\sigma_i^2$. In the third covariance matrix in Eq. (53), the variation is due to the number fluctuation of atoms within r_l because the intermolecular force is constant at r_l. Since the excluded volume of the chignolin is larger than that of the alanine dipeptide, the number of atoms within r_l is smaller in the chignolin. Therefore, the smaller number fluctuation can be expected in the chignolin. This may induce the smaller value of $\Delta\sigma_1^2$ relative to the conformational fluctuation of the target molecule (σ_i^2). Thus, one possibility of the faster convergence of the top two eigenvalues of the chignolin is due to the larger excluded volume and the larger conformational fluctuation of the chignolin.

Here we discuss the applicability of DIPA to protein molecules. To perform DIPA, we need to perform (i) a long-time MD simulation of the whole system, (ii) MD simulations that fix the target molecule, and (iii) FPCA using products of intermolecular forces and conditional cumulative densities. First, we consider the feasibility of (i) the long-time MD simulation of the whole system. Of course, to identify the conformational states by DIPA, the states have to be sampled by the MD simulation. In many protein molecules, the time scale of functional motions is submicrosecond to millisecond, so the corresponding MD simulation is required.

Second, we consider the feasibility of (ii) MD simulations that fix the target molecule. Before performing the MD simulations, we confirm that the functional states are sampled in the MD simulation. For this purpose, we can use PEPCA [18] by identifying molecular states and their important intramolecular interactions. After confirmation of the molecular states, we determine the number n of MD simulations to perform (for this study, we performed $n = 1000$ MD simulations). If the number of functional states is small, $n = 1000$ to 10 000 is sufficient. We must also determine the sampling time to estimate conditional cumulative densities. In the chignolin, the top three eigenvalues converge within 100 ps MD simulations. Because the number of water molecules around an atom of the peptide and a protein is not significantly different, we can expect that a few ns MD simulations also suffice for the protein. In the current massive computing clusters, it is feasible to perform 1000 to 10 000 independent MD simulations with a few ns.

Finally, we discuss the feasibility of (iii) the FPCA using products of intermolecular forces and conditional cumulative densities. Since we have assumed $n = 1000$ to 10 000, as per the discussion above, we can perform PCA using diagonalization of an $n \times n$ centered Gram matrix as used in Sec. III E. The centered Gram matrix can be calculated from the Gram matrix. The (i, j) component of the Gram matrix can be calculated from the ith and the jth structures. Therefore, the Gram matrix calculation can be easily decomposed and parallelized. In summary, the most time-consuming part for protein molecules is (i) the long-time MD simulation of the whole system, which is currently a common problem in the field of MD simulations. Although micro- to millisecond MD simulations are difficult, the recent breakthrough involving the special purpose machine [44] is promising. The other parts [(ii) and (iii)] are feasible with current computational power.

We can now analyze the molecular conformational fluctuations based on their intramolecular interactions (PEPCA) and intermolecular interactions (DIPA). The PEPCA is useful for understanding the molecular functional states based on their amino acid residues or bases interactions, whereas the DIPA is useful for understanding the regulation mechanism of the states by the environmental molecules. An important aspect of PEPCA and DIPA is that they can be applied to ordinary classical mechanical force fields in MD simulation. Furthermore, both results can be systematically visualized by biplots. Therefore, we believe that PEPCA and DIPA provide general analysis methods to understand the molecular conformational

fluctuations based on their intra- and intermolecular interactions.

ACKNOWLEDGMENTS

We thank Itoshi Nikaido for his valuable discussions. We are also thankful to the RIKEN Super Combined Cluster (RSCC) and the RIKEN Integrated Cluster of Clusters (RICC) for the computational resources. This research was supported by the Research Fund for Computational Science in RIKEN (H.R.U.), the Center for Developmental Biology (CDB) (H.R.U.), the Quantitative Biology Center (QBiC) (H.R.U.), a Grant-in-Aid for Scientific Research on Innovative Areas (T.J.K), and Aihara Project, the FIRST program from JSPS, initiated by CSTP (T.J.K).

APPENDIX A: CHANGES IN THE CONFORMATIONAL DISTRIBUTION OF THE SYSTEM AND TARGET MOLECULE DUE TO A PERTURBATION

Changes in the conformational distribution of the system $[D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}'))]$ and target molecule $[D(\rho'(\mathbf{q})||\rho(\mathbf{q}))]$ due to a perturbation $\Delta V(\mathbf{q},\mathbf{q}')$ can be connected by the equality [20,22]

$$D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}')) = D(\rho'(\mathbf{q})||\rho(\mathbf{q})) + \langle D(\rho'(\mathbf{q}'|\mathbf{q})||\rho(\mathbf{q}'|\mathbf{q}))\rangle_{\Delta V}, \quad \text{(A1)}$$

which is called the chain rule. The second term on the right-hand side is called the conditional Kullback-Leibler divergence (or conditional relative entropy) and is explicitly defined as

$$\langle D(\rho'(\mathbf{q}'|\mathbf{q})||\rho(\mathbf{q}'|\mathbf{q}))\rangle_{\Delta V} \equiv \int\left(\int \rho'(\mathbf{q}'|\mathbf{q})\ln\frac{\rho'(\mathbf{q}'|\mathbf{q})}{\rho(\mathbf{q}'|\mathbf{q})}\,d\mathbf{q}'\right)\rho'(\mathbf{q})\,d\mathbf{q} \geqslant 0. \quad \text{(A2)}$$

By the non-negative nature of the Kullback-Leibler divergence, Eq. (A2) is also non-negative and Eq. (A1) indicates the inequality

$$D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}')) \geqslant D(\rho'(\mathbf{q})||\rho(\mathbf{q})). \quad \text{(A3)}$$

Therefore, the change in the conformational distribution of the whole system due to the perturbation is always larger than the change in the conformational distribution of the target molecule. By using Eqs. (3) and (5), the perturbed conditional distribution $\rho'(\mathbf{q}'|\mathbf{q})$ is expressed as

$$\rho'(\mathbf{q}'|\mathbf{q}) = \frac{e^{-\beta\Delta V(\mathbf{q},\mathbf{q}')}}{\langle e^{-\beta\Delta V}|\mathbf{q}\rangle}\rho(\mathbf{q}'|\mathbf{q}). \quad \text{(A4)}$$

If we consider an intramolecular perturbation $\Delta V(\mathbf{q})$, Eq. (A4) can be expressed as $\rho'(\mathbf{q}'|\mathbf{q}) = \rho(\mathbf{q}'|\mathbf{q})$ and Eq. (A2) becomes 0. Therefore, the chain rule [Eq. (A1)] leads to the equality

$$D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}')) = D(\rho'(\mathbf{q})||\rho(\mathbf{q})). \quad \text{(A5)}$$

Thus, the changes in the conformational distribution of the system and target molecule due to the intramolecular perturbation $\Delta V(\mathbf{q})$ are equal.

TABLE VI. Second-order approximations of changes in the conformational distribution of the system $[D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}'))]$ and target molecule $[D(\rho'(\mathbf{q})||\rho(\mathbf{q}))]$ due to intra- $[\Delta V(\mathbf{q})]$ and inter- $[\Delta V(\mathbf{q},\mathbf{q}')]$ molecular perturbations.

Distribution change by the perturbation	Perturbation potential energy	
	$\Delta V(\mathbf{q})$	$\Delta V(\mathbf{q},\mathbf{q}')$
$D(\rho'(\mathbf{q},\mathbf{q}')\|\|\rho(\mathbf{q},\mathbf{q}'))$	$\frac{1}{2}\mathrm{var}(\beta\Delta V)$	$\frac{1}{2}\mathrm{var}(\beta\Delta V)$
$D(\rho'(\mathbf{q})\|\|\rho(\mathbf{q}))$	$\frac{1}{2}\mathrm{var}(\beta\Delta V)$	$\frac{1}{2}\mathrm{var}(\beta\langle\Delta V\|\mathbf{q}\rangle)$
$\langle D(\rho'(\mathbf{q}'\|\mathbf{q})\|\|\rho(\mathbf{q}'\|\mathbf{q}))\rangle_{\Delta V}$	0	$\frac{1}{2}\langle\mathrm{var}(\beta\Delta V\|\mathbf{q})\rangle$

Next we consider the second-order approximation of each term in the chain rule [Eq. (A1)]. By using Eq. (3), $D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}'))$ can be expanded as

$$D(\rho'(\mathbf{q},\mathbf{q}')||\rho(\mathbf{q},\mathbf{q}')) = \tfrac{1}{2}\mathrm{var}(\beta\Delta V) + \cdots. \quad \text{(A6)}$$

By using Eqs. (5) and (A4), Eq. (A2) can be expanded as

$$\langle D(\rho'(\mathbf{q}'|\mathbf{q})||\rho(\mathbf{q}'|\mathbf{q}))\rangle_{\Delta V} = \tfrac{1}{2}\langle\mathrm{var}(\beta\Delta V|\mathbf{q})\rangle + \cdots. \quad \text{(A7)}$$

Therefore, using Eqs. (A6), (11), and (A7), the second-order approximation of the chain rule [Eq. (A1)] gives

$$\tfrac{1}{2}\mathrm{var}(\beta\Delta V) = \tfrac{1}{2}\mathrm{var}(\beta\langle\Delta V|\mathbf{q}\rangle) + \tfrac{1}{2}\langle\mathrm{var}(\beta\Delta V|\mathbf{q})\rangle. \quad \text{(A8)}$$

We note that this equality is the law of total variance [Eq. (10)]. The results are summarized in Table VI.

APPENDIX B: RELATIONSHIP OF FUNCTIONS USED IN THE IPA AND THE DIPA

By using the conditional density $n_{iJ_k}(r|\mathbf{q})$ and Eq. (14), $\langle V_k|\mathbf{q}\rangle$ is represented as

$$\langle V_k|\mathbf{q}\rangle = \sum_{i\in I_k}\int_0^{r_c}[\phi_k(r)-\phi_k(r_c)]n_{iJ_k}(r|\mathbf{q})\,dr. \quad \text{(B1)}$$

The integration by parts of Eq. (B1) leads to

$$\langle V_k|\mathbf{q}\rangle = \int_0^{r_c} f_k(r|\mathbf{q})\,dr. \quad \text{(B2)}$$

Thus, $\langle V_k|\mathbf{q}\rangle$ (used in the IPA) is related by integration of $f_k(r|\mathbf{q})$ (used in the DIPA). This is also expressed as

$$f_k(r|\mathbf{q}) = \left.\frac{d\langle V_k|\mathbf{q}\rangle}{dr_c}\right|_{r_c=r}, \quad \text{(B3)}$$

which means $f_k(r|\mathbf{q})$ is a derivative of $\langle V_k|\mathbf{q}\rangle$.

APPENDIX C: PROCEDURE TO PERFORM FPCA

The simple method to perform FPCA is to apply PCA with discretized functional data [11,19]. We describe this method in our setting. By discretization Eq. (45), Eq. (34) can be approximated as

$$\mathrm{var}(\beta\langle\Delta V|\mathbf{q}\rangle) \approx \mathrm{var}\left(\beta\sum_{k=1}^{M}\sum_{l=1}^{N_B}\lambda_k(r_l)f_k(r_l|\mathbf{q})\Delta r\right). \quad \text{(C1)}$$

The right-hand side of Eq. (C1) can be considered the variance of the linear combination of the variable $-\beta f_k(r_l|\mathbf{q})\sqrt{\Delta r}$ with the coefficient $\lambda_k(r_l)\sqrt{\Delta r}$ for $k = 1,\ldots,M$ and

$l = 1, \ldots, N_B$. Therefore, we can consider the corresponding PCA. We denote the ith eigenvector, eigenvalue, and principal component of the PCA as $U_{ki}(r_l)\sqrt{\Delta r}$, σ_i^2, and $g_i(\mathbf{q})$, respectively. If we use n sample target molecular coordinates $\mathbf{q}$ to perform the PCA, then $i = 1, \ldots, \min(n, MN_B)$. Then we can show that the PCA estimates FPCA using $-\beta \mathbf{f}(r|\mathbf{q})$ as follows: First, the eigenvectors $U_{ki}(r_l)\sqrt{\Delta r}$ of the PCA are orthonormal:

$$\sum_{k=1}^{M}\sum_{l=1}^{N_B}(U_{ki}(r_l)\sqrt{\Delta r})(U_{kj}(r_l)\sqrt{\Delta r}) = \delta_{i,j}. \tag{C2}$$

These equalities converge to the orthonormal conditions of the eigenfunctions $U_{ki}(r)$ according to

$$\sum_{k=1}^{M}\int_0^{r_c} U_{ki}(r)U_{kj}(r)\,dr = \delta_{i,j} \tag{C3}$$

for $n, N_B \to \infty$. Second, the eigenvalue σ_i^2 of the PCA is expressed as

$$\sigma_i^2 = \mathrm{var}\left(\sum_{k=1}^{M}\sum_{l=1}^{N_B}(U_{ki}(r_l)\sqrt{\Delta r})(-\beta f_k(r_l|\mathbf{q})\sqrt{\Delta r})\right), \tag{C4}$$

which converges to the eigenvalue of the FPCA

$$\sigma_i^2 = \mathrm{var}\left(\beta\int_0^{r_c}\sum_{k=1}^{M}U_{ki}(r)f_k(r|\mathbf{q})\,dr\right) \tag{C5}$$

for $n, N_B \to \infty$. Finally, the principal component $g_i(\mathbf{q})$ of the PCA is expressed as

$$g_i(\mathbf{q}) = \sum_{k=1}^{M}\sum_{l=1}^{N_B}(U_{ki}(r_l)\sqrt{\Delta r})(-\beta(f_k(r_l|\mathbf{q}) - \langle f_k(r_l|\mathbf{q})\rangle)\sqrt{\Delta r}), \tag{C6}$$

which converges to the principal component of the FPCA

$$g_i(\mathbf{q}) = -\beta\sum_{k=1}^{M}\int_0^{r_c}U_{ki}(r)(f_k(r|\mathbf{q}) - \langle f_k(r|\mathbf{q})\rangle)\,dr \tag{C7}$$

for $n, N_B \to \infty$. Thus, we can estimate FPCA using $-\beta f_k(r|\mathbf{q})(k = 1, \ldots, M)$ by performing PCA using $-\beta f_k(r_l|\mathbf{q})\sqrt{\Delta r}(k = 1, \ldots, M$ and $l = 1, \ldots, N_B)$, which is summarized in Table VII.

TABLE VII. Estimate of FPCA using $-\beta f_k(r|\mathbf{q})(k = 1, \ldots, M)$ by performing PCA using $-\beta f_k(r_l|\mathbf{q})\sqrt{\Delta r}(k = 1, \ldots, M$ and $l = 1, \ldots, N_B)$. The ith eigenfunction $U_{ki}(r)$ is estimated from the ith eigenvector $U_{ki}(r_l)\sqrt{\Delta r}$ of PCA dividing by $\sqrt{\Delta r}$. The ith eigenvalue σ_i^2 and principal component $g_i(\mathbf{q})$ of FPCA are estimated by the ith eigenvalue σ_i^2 and principal component $g_i(\mathbf{q})$ of PCA.

	FPCA using $-\beta f_k(r\|\mathbf{q})$ $k = 1, \ldots, M$	PCA using $-\beta f_k(r_l\|\mathbf{q})\sqrt{\Delta r}$ $k = 1, \ldots, M$ $l = 1, \ldots, N_B$
ith eigenfunction (vector)	$U_{ki}(r)$	$U_{ki}(r_l)\sqrt{\Delta r}$
ith eigenvalue	σ_i^2	σ_i^2
ith principal component	$g_i(\mathbf{q})$	$g_i(\mathbf{q})$

APPENDIX D: PERTURBATION EFFECTS BY A LINEAR COMBINATION OF TWO EIGENVECTORS

We consider a unit vector $\mathbf{u} = (u_1, \ldots, u_M)^T$, which is a linear combination of the two eigenvectors $\mathbf{u}_i$ and $\mathbf{u}_j$:

$$\mathbf{u} \equiv \cos\theta\mathbf{u}_i + \sin\theta\mathbf{u}_j. \tag{D1}$$

By using Eqs. (11), (16), and (17), the change in the conformational distribution of the target molecule induced by $\boldsymbol{\lambda} = \delta\mathbf{u}$ becomes

$$D(\rho_{\delta\mathbf{u}}(\mathbf{q})||\rho(\mathbf{q})) = \tfrac{1}{2}\delta^2\left(\cos^2\theta\sigma_i^2 + \sin^2\theta\sigma_j^2\right) + \cdots. \tag{D2}$$

By using Eqs. (9), (13), and (21), the ratio change of the target molecular conformation $\mathbf{q}$ by the perturbation can be expanded as

$$\ln\frac{\rho_{\delta\mathbf{u}}(\mathbf{q})}{\rho(\mathbf{q})} = \delta g(\mathbf{q}) + \cdots, \tag{D3}$$

where we define

$$g(\mathbf{q}) \equiv \cos\theta g_i(\mathbf{q}) + \sin\theta g_j(\mathbf{q}). \tag{D4}$$

By introducing $\mathbf{n} \equiv (\cos\theta, \sin\theta)$, Eqs. (D1) and (D4) are represented by the projection to $\mathbf{n}$ as

$$u_k = \mathbf{n}\cdot(U_{ki}, U_{kj}) \tag{D5}$$

$$g(\mathbf{q}) = \mathbf{n}\cdot[g_i(\mathbf{q}), g_j(\mathbf{q})]. \tag{D6}$$

Thus, u_k and $g(\mathbf{q})$ are obtained on the biplot by projection onto $\mathbf{n}$. In summary, the perturbation effects of the linear combination of eigenvectors can be understood from the biplot as well as each eigenvector.

[1] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and F. G. Parak, Proc. Natl. Acad. Sci. USA **99**, 16047 (2002).

[2] P. W. Fenimore, H. Frauenfelder, B. H. McMahon, and R. D. Young, Proc. Natl. Acad. Sci. USA **101**, 14408 (2004).

[3] K. Gunasekaran, B. Ma, and R. Nussinov, Proteins **57**, 433 (2004).

[4] K. Henzler-Wildman and D. Kern, Nature (London) **450**, 964 (2007).

[5] R. G. Smock and L. M. Gierasch, Science **324**, 198 (2009).

[6] D. D. Boehr, R. Nussinov, and P. E. Wright, Nat. Chem. Biol. **5**, 789 (2009).

[7] T. Ichiye and M. Karplus, Proteins **11**, 205 (1991).

[8] P. H. Hünenberger, A. E. Mark, and W. F. van Gunsteren, J. Mol. Biol. **252**, 492 (1995).

[9] Y. Kong and M. Karplus, Structure **15**, 611 (2007).

[10] Y. Kong and M. Karplus, Proteins **74**, 145 (2009).
[11] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. (Springer, New York, 2002).
[12] A. Kitao, F. Hirata, and N. Go, Chem. Phys. **158**, 447 (1991).
[13] A. E. García, Phys. Rev. Lett. **68**, 2696 (1992).
[14] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).
[15] R. Abseher and M. Nilges, J. Mol. Biol. **279**, 911 (1998).
[16] Y. Mu, P. H. Nguyen, and G. Stock, Proteins **58**, 45 (2005).
[17] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. **126**, 244111 (2007).
[18] Y. M. Koyama, T. J. Kobayashi, S. Tomoda, and H. R. Ueda, Phys. Rev. E **78**, 046702 (2008).
[19] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis*, 2nd ed. (Springer, New York, 2006).
[20] S. Kullback, *Information Theory and Statistics* (Wiley & Sons, New York, 1959).
[21] S. Amari and H. Nagaoka, *Methods of Information Geometry* (AMS and Oxford University Press, New York, 2000).
[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley Interscience, New York, 2006).
[23] D. Ming and M. E. Wall, Proteins **59**, 697 (2005).
[24] D. Ming and M. E. Wall, J. Mol. Biol. **358**, 213 (2006).
[25] D. R. Brillinger, Ann. Inst. Stat. Math. **21**, 215 (1969).
[26] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, 2nd ed. (Academic Press, San Diego, 2002), Chap. 3.2.2.
[27] E. Małolepsza, B. Strodel, M. Khalili, S. Trygubenko, S. N. Fejer, and D. J. Wales, J. Comput. Chem. **31**, 1402 (2010).
[28] R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, 1985).
[29] K. R. Gabriel, Biometrika **58**, 453 (1971).
[30] J. C. Gower and D. J. Hand, *Biplots* (Chapman & Hall, London, 1996).
[31] D. A. Case *et al.*, AMBER10 (University of California, San Francisco, 2008).
[32] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, J. Comput. Chem. **24**, 1999 (2003).
[33] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[34] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, J. Comput. Phys. **23**, 327 (1977).
[35] S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).
[36] T. Darden, D. York, and L. Pedersen, J. Chem. Phys. **98**, 10089 (1993).
[37] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[38] S. Honda, K. Yamasaki, Y. Sawada, and H. Morii, Structure **12**, 1507 (2004).
[39] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel, J. Mol. Biol. **354**, 173 (2005).
[40] D. Satoh, K. Shimizu, S. Nakamura, and T. Terada, FEBS Lett. **580**, 3422 (2006).
[41] A. Suenaga, T. Narumi, N. Futatsugi, R. Yanai, Y. Ohno, N. Okimoto, and M. Taiji, Chem. Asian J. **2**, 591 (2007).
[42] [http://www-wales.ch.cam.ac.uk/∼wales/perm-prmtop.ff03.py].
[43] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006), Chap. 12.1.4.
[44] D. E. Shaw *et al.*, Commun. ACM **51**, 91 (2008).