Special Issue
"Progress of Theoretical and Computational Biophysics"

Vol. 16, pp. 407–429 (2019)
doi: 10.2142/biophysico.16.0_407

Biophysics and Physicobiology
https://www.jstage.jst.go.jp/browse/biophysico/

Review Article

Analysis of molecular dynamics simulations of 10-residue peptide, chignolin, using statistical mechanics: Relaxation mode analysis and three-dimensional reference interaction site model theory

Yutaka Maruyama[1], Hiroshi Takano[2] and Ayori Mitsutake[3]

[1] *Architecture Development Team, FLAGSHIP 2020 Project, RIKEN Center for Computational Science, Kobe, Hyogo 650-0047, Japan*
[2] *Department of Physics, Faculty of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan*
[3] *Department of Physics, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan*

Received July 2, 2019; accepted August 29, 2019

Molecular dynamics simulation is a fruitful tool for investigating the structural stability, dynamics, and functions of biopolymers at an atomic level. In recent years, simulations can be performed on time scales of the order of milliseconds using specialpurpose systems. Since the most stable structure, as well as meta-stable structures and intermediate structures, is included in trajectories in long simulations, it is necessary to develop analysis methods for extracting them from trajectories of simulations. For these structures, methods for evaluating the stabilities, including the solvent effect, are also needed. We have developed relaxation mode analysis to investigate dynamics and kinetics of simulations based on statistical mechanics. We have also applied the three-dimensional reference interaction site model theory to investigate stabilities with solvent effects. In this paper, we review the results for designing amino-acid substitution of the 10-residue peptide, chignolin, to stabilize the misfolded structure using these developed analysis methods.

Key words: molecular simulation, protein, analysis method, dynamics, stability

Professor Nobuhiro Go is one of the pioneers who introduced the concepts of statistical mechanics to protein research using computers. Protein folding mechanisms were examined using lattice models [1]. The algorithms to determine protein structures based on experimental NMR data were also developed using computers [2]. In addition, he and his coworkers introduced normal mode analysis (NMA) to protein simulation systems [3], in which the idea of mode decomposition was introduced in protein systems. Langevin mode analysis (LMA) was introduced to investigate modes around a minimum-energy state, including the water effect [4–6]. Principal component analysis (PCA) was also introduced [6,7] into protein systems. The free energy profiles along the normal modes of a protein with the solvent effect were calculated by using the extended reference interaction site model (XRISM) theory [8]. Complex protein motion

Corresponding author: Ayori Mitsutake, Department of Physics, School of Science and Technology, Meiji University, 1-1-1 Higashi-Mita, Tama-ku, Kawasaki, Kanagawa 214-8571, Japan.
e-mail: ayori@meiji.ac.jp

◀ ***Significance*** ▶

It is important to develop analysis methods for molecular simulations of protein systems. We have developed relaxation mode analysis to investigate dynamics and kinetics of simulations. We also have applied the three-dimensional reference interaction site model theory to investigate stability with solvent effects. Here, we review the results for designing amino-acid substitution of 10-residue peptide, chignolin, to stabilize the misfolded structure using these analysis methods based on statistical mechanics.

was decomposed to PC modes with large structural fluctuations; PC modes can be also used as the free energy axis to calculate the free energy landscape. In PCA, the PC modes with large structural fluctuations correspond to the main motions of a protein with a harmonic free energy surface. However, for an anharmonic energy surface with multiple local energy minima, they do not correspond to transitions between the minimum-energy states. Thus, the jumping-among-minima (JAM) model was introduced to treat multiple-hierarchy free-energy landscape [9]. Professor Nobuhiro Go contributed to the development of analysis methods for mode decomposition to investigate protein stability, dynamics, and functions.

Molecular dynamics (MD) simulation is a powerful tool to investigate protein stability, dynamics, and functions because simulations can be performed on time scales of the order of milliseconds using special hardware [10–12]. Methods related to molecular simulations based on statistical mechanics must be developed. From a molecular perspective, amino acid mutations can alter the stabilities or functions of proteins. Many researchers have altered the structural stabilities or enhanced the functions of proteins by artificially inducing mutations using computational approaches. We were interested in designing artificial proteins based on amino acid substitution. Therefore, we needed analysis methods to extract not only the most-stable structure but also other characteristic structures like meta-stable and intermediate structures from the simulations. We have developed relaxation mode analysis (RMA) [13–15] to investigate the dynamics and kinetics of the simulations. After obtaining these structures, we also need to estimate their stabilities in amino acid level. We have applied the three-dimensional reference interaction site model (3D-RISM) theory to investigate these stabilities with solvent effects [16]. These above-described methods are based on statistical mechanics. Especially, by comparing the methods and results of RMA and PCA, we improved RMA such that it can be applied to heteropolymer, (protein,) systems, and confirmed the inferences from the RMA results for protein simulations [17–19]. Here, we introduce methods based on statistical mechanics, RMA, and 3D-RISM approaches, and review the results of designing amino-acid substitution of the 10-residue peptide, chignolin [20], to stabilize the misfolded structure rather than the native structure (See Refs. 21 and 22 for the detailed results.)

In this review, we describe a background of RMA and a simple RMA for protein systems in the section titled "Relaxation Mode Analysis" (See Ref. 23 for more details), 3D-RISM in the section titled "Reference Interaction Site Models" and report the results of designing amino-acid substitution of chignolin using these observations in the section titled "Results".

Relaxation Mode Analysis (RMA)

Background of mode decomposition methods for protein simulations

As longer and larger MD simulations are performed, it has become increasingly important to develop methods to extract the "essential" information from the trajectory. The reduction in the large number of degrees of freedom of coordinates to a few collective ones is an active field of theoretical research. In NMA, the normal modes near the minimum potential energy of the protein molecule are obtained [3,24,25]. LMA investigates modes around a minimum-energy state, including the water effect [4,6,26,27]. An elastic network model and a Gaussian network model approximately calculate normal modes with large amplitudes by using the harmonic potential of coarse-grained models [28–32]. This method extracts collective modes with large amplitudes for large protein systems like viruses, because such proteins have rigid-like motions [33]. PCA, also called quasiharmonic analysis or the essential dynamics method [6,7,34–38], is one of the most popular methods for analyzing the structural fluctuations around the average structure. The modes with large structure fluctuations are extracted and are considered cooperative movement, and the relation of these fluctuations with function has been widely examined. The obtained modes are also used as the axis of the free-energy surface. The JAM model was introduced for treating multiple-hierarchy free-energy landscapes [9]. They divided protein motions into intra-substate and inter-substate motions. Various other analysis methods have also been proposed, such as full correlation analysis [39], subspace joint approximate diagonalization of eigenmatrices [40], wavelet analysis [41], and others [42,43] (See also Ref. 44). Manifold learning techniques for analyzing nonlinear data have been also applied to protein systems; for example, isomap method [45] and diffusion map method [46].

Analysis methods have also been developed depending on the increase in simulation time. In recent years, prolonged simulations can be performed; thus, dynamic analysis methods are required to identify local-minimum energy states and analyze the transitions between them. (Here, dynamic analysis methods mean that the variations of physical quantities with time are directly used in analysis.) Accordingly, many methods have been developed to analyze the dynamics and kinetics of protein simulations [47–55]. The Markov state model has been reported and applied to many protein systems [49,52,53,56–68]; it can analyze transitions between local minimum-energy states, which are identified by using clustering analysis methods. Core-set milestoning analysis has also been applied [69]. Isomap and diffusion map methods were used to cluster states for dynamical analysis methods [45,46].

RMA was developed to investigate the "dynamic" properties of spin systems [13] and homopolymer systems for Monte Carlo (MC) [14] and MD [15] simulations and has been applied to various homopolymer systems [70–73] to investigate their slow relaxation dynamics [74,75]. RMA approximately estimates slow relaxation modes and rates by solving the generalized eigenvalue problem, $\sum_{j=1}^{3N} C_{i,j}(t_0+\tau) f_{p,j} = \exp(-\lambda_p \tau) \sum_{j=1}^{3N} C_{i,j}(t_0) f_{p,j}$, where $C(t)$ is the time-displaced correlation matrix of relevant physical quantities calculated from trajectories. Here, t_0 and τ are parameters of RMA called evolution time and lag time, respectively. The estimated relaxation modes and rates are given by $f_{p,j}$ and λ_p, respectively. Recently, RMA has also been applied to biomolecular systems [17–19,21,31,76–78] and its effectiveness has been demonstrated. RMA is effective in folding simulations because sufficiently long trajectories and only C_α coordinates are used for RMA.

However, for the trajectory of short simulations and with many degrees of freedom, it is difficult to solve the generalized eigenvalue problem, especially with t_0 because of relatively low statistical precision of the matrices. Recently, we have performed several μs simulations using graphics processing units. However, even these simulations do not provide enough sampling. In addition, for μs simulations, we must treat heavy atom coordinates for RMA because the side chains are rearranged in an approximately μs time scale. In this case, we must treat large degrees of freedom, i.e., coordinates of the heavy atoms for the side chains. In our experience, RMA can automatically extract rare events during short simulations [79]. Even the rare rotational movements of the side chains for μs simulations are extracted from RMA, though the numerical solution of generalized eigenvalue problems with t_0 is difficult to obtain (see Ref. 80 for more details). To solve the problem and improve the relaxation modes and times, we developed improved versions of RMA, including RMA with multiple evolution times, principal component RMA [31], two-step RMA [73,79], multistep RMA [81], and positive definite RMA (PDRMA) [80] to treat the trajectory of short simulations and with many degrees of freedom. Especially, PDRMA is a convenient method to perform RMA with t_0. (Notably, when t_0=0, because $C(0)$ has the positive definiteness, we can solve the generalized eigenvalue problems with t_0=0. However, we obtained many modes with negative eigenvalues when τ increases. Since these modes arise from the noise of the system, we may ignore these modes with negative eigenvalues and focus on a few modes with slower relaxations. Slow modes correspond to transitions of rare events during the simulation. To solve the negative value problems and improve the relaxation times and modes, we can also use improved RMA methods.) We also developed Markov state RMA [21] to introduce t_0 to the Markov state model. (See a review of Ref. 23 for biomolecular systems.)

Conventional RMA approximately estimates slow relaxation modes by solving the generalized eigenvalue problem of the time correlation matrices of coordinates for two different times, $C(\tau+t_0)$ and $C(t_0)$, which are calculated from the trajectory. Recently, dynamical analysis methods have been developed for molecular simulations of biopolymer systems. In these techniques, such as time structure-based independent component analysis (tICA) [50,51], time-lagged independent component analysis (TICA) [52,53], and dynamic component analysis (DCA) [54,55], time correlation matrices of certain physical quantities or states are used. Notably, tICA is a special case of RMA with t_0=0, and gave rise to a different concept, the independent component analysis, to RMA with t_0=0. (See Refs. 50 and 19.) (See Refs. 80 and 23 for the relation between RMA and other dynamical analysis methods of biomolecular systems.) In tICA, TICA, and DCA, the time correlation functions C(τ) and C(0) are used, whereas C($\tau+t_0$) and C(t_0) are used in RMA. Because the relaxation modes and rates are given as the left eigenfunctions and the eigenvalues of the time-evolution operator of the master equation of the system, respectively, RMA is related to the Markov state models. (The relationship among the Markov state model, tICA, and TICA is explained in Refs. 21,52,53.) A Markov state model was constructed from clustering in the subspace determined by tICA. (A combination of tICA and a Markov state model was also proposed in Refs. 52,53.) Here, we describe the statistical mechanical background of RMA and the original RMA used for protein simulations (See Ref. 23 for more details).

Introduction of RMA

The relaxation modes $\{X_p\}$ satisfy

$$\langle X_p(t) X_q(0) \rangle = \delta_{p,q} e^{-\lambda_p t}. \tag{1}$$

Here, $\langle A(t)B(0) \rangle$ denotes the equilibrium correlation of A at time t and B at time 0:

$$\langle A(t)B(0) \rangle = \sum_{Q,Q'} A(Q) T_t(Q|Q') B(Q') P_{\mathrm{eq}}(Q'), \tag{2}$$

where $T_t(Q|Q')$ is the conditional probability that the system is in state Q at time t given that it is in state Q'at time t=0. Further, $P_{\mathrm{eq}}(Q')$ denotes the probability that the system is in state Q' at equilibrium. The relaxation rate of X_p is denoted by λ_p. The relaxation time is calculated by $1/\lambda_p$. In RMA, we consider the variational problem, which is equivalent to the eigenvalue problem of the time evolution operator, and choose an appropriate trial function to estimate the slow relaxation modes and rates in the system. From these processes, we derive the generalized eigenvalue problem of the time correlation matrices for two different times. From the eigenvectors and eigenvalues, we approximately estimate slow relaxation modes and rates.

Relaxation times and modes

In this section, we provide the definition of relaxation modes and rates from the viewpoint of the statistical mechanics [82,83]. The relaxation modes $\{X_p\}$ satisfy Eq. (1). The relaxation modes and rates are given as left eigenfunctions and eigenvalues of the time evolution operator of the master equation of the system, respectively. We first explain the relation in three types of simulations satisfying the detailed balance condition.

In a MC simulation satisfying the detailed balance condition, the time evolution of the probability $P(Q;\, t)$ that the biomolecule is in a state $Q=(\boldsymbol{r}_1^{\mathrm{T}}, \boldsymbol{r}_2^{\mathrm{T}}, ..., \boldsymbol{r}_N^{\mathrm{T}})^{\mathrm{T}}$ at time t is described by a master equation:

$$\frac{\partial}{\partial t}P(Q;\, t) = -\sum_{Q'}\Gamma(Q|Q')P(Q';\, t). \tag{3}$$

Here, $\Gamma(Q|Q')$ denotes the $(Q,\, Q')$-component of the time evolution matrix Γ, and $\sum_{Q'}$ denotes the summation over all possible states. $\Gamma(Q|Q')$ is chosen so that the detailed balance for the equilibrium distribution function $P_{\mathrm{eq}}(Q)$ is satisfied:

$$\Gamma(Q|Q')P_{\mathrm{eq}}(Q') = \Gamma(Q'|Q)P_{\mathrm{eq}}(Q). \tag{4}$$

In the Brownian dynamics simulation, the time evolution of coordinates $\boldsymbol{r}_i$, $(i=1, ..., N)$ is given by the Langevin equation for a biomolecule with N atoms:

$$\frac{d\boldsymbol{r}_i}{dt} = -\frac{1}{\zeta}\left[-\frac{\partial}{\partial \boldsymbol{r}_i}U(\{\boldsymbol{r}_j\}) + \boldsymbol{w}_i\right]. \tag{5}$$

Here, $\boldsymbol{r}_i(t)$ denotes the position of the ith atom at time t, and ζ is the friction constant. The interaction between atoms is described by the potential $U(\{\boldsymbol{r}_i\})=U(\boldsymbol{r}_1, ..., \boldsymbol{r}_N)$. The random force $\boldsymbol{w}_i(t)$ acting on the ith atom is a Gaussian white stochastic process, and satisfies

$$\langle w_{i,\alpha}(t)w_{j,\beta}(t)\rangle = 2\zeta k_B T\delta_{\alpha,\beta}\delta_{i,j}\delta(t-t'), \tag{6}$$

where $w_{i,\alpha}$, k_B, and T denote the α-component of $\boldsymbol{w}_i$ ($\alpha=x$, y, or z), the Boltzmann constant, and the temperature of the system, respectively. The Smoluchowski equation equivalent to Eq. (5) can be written as

$$\frac{\partial}{\partial t}P(Q,\, t) = -\Gamma(Q)P(Q,\, t)$$

$$= \sum_{i=1}\frac{\partial}{\partial \boldsymbol{r}_i}\cdot\frac{1}{\zeta}\left\{k_B T\frac{\partial}{\partial \boldsymbol{r}_i} + \frac{\partial U}{\partial \boldsymbol{r}_i}\right\}P. \tag{7}$$

Here, $Q=\{\boldsymbol{r}_1, ..., \boldsymbol{r}_N\}$ denotes a point in the phase space of the system, and $P(Q,\, t)dQ$ denotes the probability that the system is found at time t in an infinitesimal volume dQ at point Q in the phase space. The time evolution operator Γ satisfies the detailed balance condition [82]:

$$P_{\mathrm{eq}}(Q')\Gamma(Q)\delta(Q-Q') = P_{\mathrm{eq}}(Q)\Gamma^{\dagger}(Q)\delta(Q-Q'), \tag{8}$$

where $P_{\mathrm{eq}}(Q) \propto \exp\left[-\frac{U(\{\boldsymbol{r}_j\})}{k_B T}\right]$. Here, $\Gamma(Q)$ and the adjoint operator $\Gamma^{\dagger}(Q)$ act only on Q in $\delta(Q-Q')$. In the matrix representation, so that $\Gamma(Q)\delta(Q-Q')=\Gamma(Q|Q')$ and $\Gamma^{\dagger}(Q)\delta(Q-Q')=\Gamma(Q'|Q)$, the detailed balance condition is the same as that in Eq. (4).

In a MD simulation with the Langevin thermostat, the time evolution of coordinates $\boldsymbol{r}_i$, $(i=1, ..., N)$ is given by the Langevin equation for a biomolecule with N atoms:

$$m_i\frac{d\boldsymbol{v}_i}{dt} = -\zeta\boldsymbol{v}_i - \frac{\partial}{\partial \boldsymbol{r}_i}U(\{\boldsymbol{r}_j\}) + \boldsymbol{w}_i, \tag{9}$$

with

$$\frac{d\boldsymbol{r}_i}{dt} = \boldsymbol{v}_i. \tag{10}$$

Here, $\boldsymbol{r}_i(t)$ and $\boldsymbol{v}_i(t)$ denote the position and the velocity of the ith atom at time t, respectively. The mass of the ith atom is denoted by m_i is and ζ is the friction constant.

The Kramers equation, equivalent to Eqs. (9) and (10), can be written as

$$\frac{\partial}{\partial t}P(Q,t) = -\Gamma(Q)P(Q,t)$$

$$= \sum_{i=1}^{N}\left\{\frac{\partial}{\partial \boldsymbol{r}_i}\cdot\boldsymbol{v}_i - \frac{1}{m_i}\frac{\partial}{\partial \boldsymbol{v}_i}\cdot\frac{\partial U}{\partial \boldsymbol{r}_i} - \frac{\zeta}{m_i}\frac{\partial}{\partial \boldsymbol{v}_i}\cdot\left(\boldsymbol{v}_i + \frac{K_B T}{m_i}\frac{\partial}{\partial \boldsymbol{v}_i}\right)\right\}P. \tag{11}$$

Here, $Q=\{\boldsymbol{r}_1, ..., \boldsymbol{r}_N, \boldsymbol{v}_1, ..., \boldsymbol{v}_N\}$ denotes a point in the phase space of the system. The time evolution operator Γ satisfies the detailed balance condition:

$$P_{\mathrm{eq}}(Q')\Gamma(Q)\delta(Q-Q') = P_{\mathrm{eq}}(\epsilon Q)\Gamma^{\dagger}(\epsilon Q)\delta(\epsilon Q-\epsilon Q'), \tag{12}$$

where $P_{\mathrm{eq}}(Q) \propto \exp\left(-\frac{1}{k_B T}\left[\frac{1}{2}\sum_i m_i\boldsymbol{v}_i^2 + U(\{\boldsymbol{r}_j\})\right]\right)$ and $P_{\mathrm{eq}}(Q)=P_{\mathrm{eq}}(\epsilon Q)$. Here, ϵQ denotes the time-reversed state of the state Q, namely,
$\epsilon Q = \{\epsilon_1\boldsymbol{r}_1, ..., \epsilon_N\boldsymbol{r}_N, \epsilon_{N+1}\boldsymbol{v}_1, ..., \epsilon_{2N}\boldsymbol{v}_N\}$ with

$$\epsilon_i = \begin{cases} 1 & \text{for } i = 1, ..., N, \\ -1 & \text{for } i = N+1, ..., 2N. \end{cases} \tag{13}$$

In the matrix representation, the detailed balance condition is written as:

$$\Gamma(Q|Q')P_{\mathrm{eq}}(Q') = \Gamma(\epsilon Q'|\epsilon Q)P_{\mathrm{eq}}(\epsilon Q). \tag{14}$$

The time evolution equation of $P(Q;\, t)$ of Eqs. (7) and (11) corresponds to Eq. (3) in the matrix representation. In MC and Brownian dynamics, because only coordinates are the degrees of freedom in the system, $\epsilon Q=Q$, the detailed balance condition in all three cases is given by Eq. (14).

We now consider the eigenvalue problem of the time evolution operator $\Gamma(Q|Q')$ of the master equation:

$$\sum_{Q}\phi_n(Q)\Gamma(Q|Q') = \lambda_n\phi_n(Q'). \tag{15}$$

$$\sum_{Q'}\Gamma(Q|Q')\psi_n(Q') = \lambda_n\psi_n(Q). \tag{16}$$

Here, $\phi_n(Q)$ and $\psi_n(Q)$ are the left and right eigenfunctions of the time-evolution operator Γ with eigenvalue λ_n, respectively. When we define a quantity $\hat{\phi}_n(Q)$ through

$$\psi_n(Q) = \hat{\phi}_n(Q)P_{\mathrm{eq}}(Q), \tag{17}$$

then $\hat{\phi}_n(Q)=\phi_n(\epsilon Q)$. The eigenfunctions are chosen to satisfy the orthonormal condition:

$$\begin{aligned}\sum_Q \phi_m(Q)\psi_n(Q) &= \sum \phi_m(Q)\hat{\phi}_n(Q)P_{\mathrm{eq}}(Q) \\ &= \langle \phi_m \hat{\phi}_n \rangle = \delta_{m,n}.\end{aligned} \tag{18}$$

The equilibrium time-displaced correlation function of $\phi_n(Q)$ and $\hat{\phi}_m(Q)$ is given by:

$$\begin{aligned}\langle \phi_m(t)\hat{\phi}_n(0)\rangle &= \sum_Q \sum_{Q'} \phi_m(Q)T_t(Q|Q')\hat{\phi}_n(Q')P_{\mathrm{eq}}(Q') \\ &= \sum_Q \sum_{Q'} \phi_m(Q)e^{-\Gamma t}(Q|Q')\hat{\phi}_n(Q')P_{\mathrm{eq}}(Q') \\ &= \sum_Q \sum_{Q'} \phi_m(Q)e^{-\Gamma t}(Q|Q')\psi_n(Q') \\ &= \sum_Q \phi_m(Q)e^{-\lambda_n t}\psi_n(Q) \\ &= \delta_{m,n}e^{-\lambda_n t}.\end{aligned} \tag{19}$$

where $T_t(Q|Q')=e^{-\Gamma t}(Q|Q')$ is the conditional probability that the system is found at time t at Q given that the system is at Q' at time 0.

If two quantities $A(Q)$ and $B(Q)$ are expanded as

$$A(Q) = \sum_n a_n \phi_n(Q) \quad \text{and} \quad B(Q) = \sum_n \hat{b}_n \hat{\phi}_n(Q), \tag{20}$$

then the time correlation function of A and B in the equilibrium state is given by

$$\langle A(t)B(0)\rangle = \sum_n a_n \hat{b}_n \exp(-\lambda_n t). \tag{21}$$

Thus, in terms of $\phi_n(Q)$ and $\hat{\phi}_n(Q)$, the correlation function $\langle A(t)B(0)\rangle$ is decomposed into a sum of exponentially relaxing contributions. Therefore, we use two sets of functions, $\{\phi_n(Q)\}$ and $\{\hat{\phi}_n(Q)\}$, as relaxation modes, and refer to $\{\lambda_n\}$ as their relaxation rates. The relaxation modes and rates are given as left eigenfunctions and eigenvalues of the time evolution operator of the master equation of the system, respectively.

RMA with a single evolution time, t_0

Here, we explain the manner in which to estimate the slow relaxation modes and rates by RMA. We consider the variational problem, which is equivalent to the eigenvalue problem of the time evolution operator, and choose an appropriate trial function in order to estimate the slow relaxation modes and rates. We consider the following equations for the conditional probability:

$$\sum_Q \phi_n(Q)T_\tau(Q|Q') = e^{-\lambda_n \tau}\phi_n(Q'), \tag{22}$$

$$\sum_{Q'} T_\tau(Q|Q')\psi_n(Q') = e^{-\lambda_n \tau}\psi_n(Q). \tag{23}$$

The eigenvalue problem in Eqs. (22) and (23) is equivalent to the variational problem

$$\delta R = 0 \tag{24}$$

with

$$R[\phi_n] = \frac{\langle \phi_n(\tau)\hat{\phi}_n(0)\rangle}{\langle \phi_n(0)\hat{\phi}_n(0)\rangle}, \tag{25}$$

and the stationary value of R gives the eigenvalue $\exp(-\lambda_n\tau)$. RMA treats the variational problem of Eqs. (24) and (25) using trial functions instead of the eigenvalue problem of Eqs. (22) and (23). To choose the trial function provided by a linear combination of important relevant quantities, we can evaluate the relaxation modes and rates from simulation data. Herein, we consider a biopolymer composed of N atoms and only treat the coordinates because the velocities have faster relaxations (~picosecond order) than coordinates in protein systems. We assume that $\boldsymbol{R}$ is a $3N$-dimensional column vector that comprises a set of atomic coordinates relative to their average coordinates

$$\boldsymbol{R}^{\mathrm{T}} = (\boldsymbol{r}_1'^{\mathrm{T}}, \boldsymbol{r}_2'^{\mathrm{T}}, \ldots, \boldsymbol{r}_N'^{\mathrm{T}}) = (x_1', y_1', z_1', \ldots, x_N', y_N', z_N') \tag{26}$$

with

$$\boldsymbol{r}_i' = \boldsymbol{r}_i - \langle \boldsymbol{r}_i \rangle, \tag{27}$$

where $\boldsymbol{r}_i$ is the coordinate of the ith atom of the biopolymer in the center-of-mass coordinate system, and $\langle \boldsymbol{r}_i \rangle$ is its average after removing the rotational degrees of freedom. Because we consider the coordinates only, $\hat{\phi}_n(Q)=\phi_n(\epsilon Q)=\phi_n(Q)$ holds. In RMA, we use the following function as an approximate relaxation mode:

$$X_p(Q) = \sum_{i=1}^{3N} f_{p,i}R_i(t_0/2;\, Q), \tag{28}$$

with

$$R_i(t;\, Q) = \sum_{Q'} R_i(Q')T_t(Q'|Q). \tag{29}$$

Here, $R_i(Q)$ is the ith component of $\boldsymbol{R}$. The quantity $R_i(t;\, Q)$ is the expectation value of R_i after a period t starting from a state Q and satisfies $R_i(t;\, Q)|_{t=0}=R_i(Q)$. The parameter t_0 is introduced to reduce the relative weight of the faster modes contained in $\boldsymbol{R}$, and it is expected that Eq. (28) becomes a better approximation as t_0 becomes larger.

For the trial function (28), R defined by Eq. (25) is given by

$$R[X_p] = \frac{\sum_{i=1}^{3N}\sum_{j=1}^{3N} f_{p,i} C_{i,j}(t_0+\tau) f_{p,j}}{\sum_{i=1}^{3N}\sum_{j=1}^{3N} f_{p,i} C_{i,j}(t_0) f_{p,j}}, \tag{30}$$

where $C_{i,j}(t)$ is a component of a $3N\times 3N$ symmetric matrix $C(t)$ defined by

$$C_{i,j}(t) = \langle R_i(t) R_j(0)\rangle. \tag{31}$$

Then, the variational problem of Eq. (25) becomes a generalized eigenvalue problem

$$\sum_{j=1}^{3N} C_{i,j}(t_0+\tau) f_{p,j} = \exp(-\lambda_p \tau) \sum_{j=1}^{3N} C_{i,j}(t_0) f_{p,j}. \tag{32}$$

The orthonormal condition of Eq. (18) for X_p is written as

$$\sum_{i=1}^{3N}\sum_{j=1}^{3N} f_{p,i} C_{i,j}(t_0) f_{p,j} = \delta_{p,q}. \tag{33}$$

Equations (32) and (33) determine the relaxation rates λ_p and the corresponding relaxation modes $f_{p,i}$. We chose the indices of λ_p such that $0<\lambda_1\leq\lambda_2\leq$... holds. Here, the relation

$$T_t(Q|Q')P_{\text{eq}}(Q') = T_t(Q'|Q)P_{\text{eq}}(Q), \tag{34}$$

which is equivalent to the detailed balance condition of Eq. (14) with $\epsilon Q=Q$, and the Markovian property

$$\sum_{Q'} T_{t_1}(Q|Q') T_{t_2}(Q'|Q'') = T_{t_1+t_2}(Q|Q'') \tag{35}$$

are used. The inverse transformation of Eq. (28) is given by

$$R_i(t_0/2;\, Q) = \sum_{p=1}^{3N-6} g_{i,p} X_p(Q) \tag{36}$$

with

$$g_{i,p} = \sum_{j=1}^{3N} C_{i,j}(t_0) f_{p,j}. \tag{37}$$

The upper limit of the summation in Eq. (36) is $3N-6$. This is because the translational and rotational degrees of freedom are removed from $\boldsymbol{R}$.

The time correlation functions of R_i are reproduced by

$$\begin{aligned}\langle R_i(t) R_j(0)\rangle &= \sum_p \sum_q g_{i,p} g_{j,q} \langle X_p(t-t_0) X_q(0)\rangle \\ &\simeq \sum_p g_{i,p} g_{j,p} \exp[-\lambda_p(t-t_0)] \\ &= \sum_p \tilde{g}_{i,p} \tilde{g}_{j,p} \exp(-\lambda_p t)\end{aligned} \tag{38}$$

for $t\geq t_0$. Here,

$$\tilde{g}_{i,p} = g_{i,p} \exp(\lambda_p t_0/2). \tag{39}$$

Relaxation mode expansion of R_i is given by

$$R_i \simeq \sum_{p=1}^{3N-6} \tilde{g}_{i,p} X_p. \tag{40}$$

As we are considering position coordinates only, the detailed balance condition yields the following consequences: $C(t)$ is a symmetric matrix, $C_{i,j}(t)=C_{j,i}(t)$; $\{\lambda_p\}$ are real and positive, which corresponds to pure relaxation. We refer to this method as the "RMA method with a single evolution time," which is $t_0/2$.

RMA for protein simulations

Takano and coworkers developed RMA to investigate the "dynamic" properties of spin systems [13] and homopolymer systems for MC [14] and MD [15]; subsequently, RMA has been applied to various homopolymer systems [70–73] to investigate their slow relaxation dynamics [74,75]. RMA corresponds to one of the mode decomposition methods. When authors started applying RMA to heteropolymer, (protein,) systems, several works related to PCA from the works of Professor Nobuhiro Go such as NMA [3], LMA [4], PCA [6,7], and JAM model [9] were first studied. This is because that the equations used in PCA and RMA are similar while their theoretical backgrounds are different. PCA extracts modes with large structural fluctuation by diagonalizing the covariance matrix of coordinates. RMA extracts modes with slow motions by solving the generalized eigenvalue problem of the time correlation matrices of coordinates for two different times. When we applied PCA to a replica-exchange MC simulation with one-dimensional reference interaction site model (1D-RISM) theory [84], we knew that the most of the difference between homopolymer and heteropolymer systems lay with the treatment of the translational and rotational degrees of freedom. In homopolymer systems, relaxation of the positions of a polymer relative to the center of the mass is examined. This means that the translational degrees of freedom are removed from the coordinates of the polymer. Because the rotational degrees of freedom remain, the rotational relaxation of the polymer is observed as slow relaxations. Treatment for removing the translational degrees of freedom in the homopolymer system was introduced [14,15]. In contrast, in PCA of protein systems, evaluating fluctuations of the conformations of a biomolecule around its average conformation is of interest. The translational and rotational degrees of freedom are removed from the sampled conformations. Thus, we proposed how to treat the generalized eigenvalue problem for removing the translational and rotational degrees of freedom for RMA [19,23]. We tested RMA to a long MC simulation of five-residue peptide, Enkephalin [17–19], because one of the authors previously performed multicanonical simulations of the peptide and knew that the system has several local minimum states. A MC simulation of the system at 298.15 K was performed 100 times more than the multicanonical simulation [85] because we needed high statistical precision for the time correlation matrices. In our previous study [17–19], we also confirmed how to calculate the free energy surface by using two characteristic vectors, $\{g_{i,n}\}$ and $\{f_{i,n}\}$ of RMA by comparing the results of PCA and RMA. After these processes, the results obtained by RMA had similar but different ones from PCA. The slowest relaxation mode corresponded to the first PC mode. However, we obtained the second slowest relaxation mode, which did not correspond to PC modes with large fluctuation. It took considerable time to understand the meanings of the second slowest relaxation modes. At the beginning, we only focus on the C_α atoms because we used C_α atoms for PCA and RMA. As we investigated the conformations with side chains after clustering, we realized that the second slowest mode corresponded to the transition of a side chain, which had a slow motion but small fluctuation. The side-chain motions affect the main chains. By comparing with PCA and RMA, we can introduce RMA to protein systems and examine the meaning of RMA. After RMA was applied to folding simulations [21,78], we confirmed that RMA is suitable for analyzing simulations with large conformational changes. RMA can also automatically extract rare events during short simulations [79].

In this section, we explain how to treat the generalized eigenvalue problem for removing translational and rotational degrees of freedom when using the coordinates for the trial function [19]. In this process, the generalized eigenvalue problem for real symmetric matrices can be easily solved numerically if the matrices are positive definite. Therefore, we shift the zero eigenvalues to finite positive values without changing the other eigenvalues and the corresponding eigenvectors. The process for RMA using coordinates as the trial function is as follows (see Fig. 1 of Ref. 23 for a schematic illustration of the process). First, we remove the translational and rotational degrees of freedom in the same manner as when conducting PCA [86,87]. After the average structure converges, the origin of the coordinate system is chosen to be the center of the mass of the average positions, $\langle \boldsymbol{r}_i \rangle$ with i=1, ..., N. In addition, the axes of the coordinate system are chosen to be the principal axes of the moment of the inertia tensor of the average positions. After the coordinates are root mean-square deviation (RMSD) fit to the obtained average structure, we calculate $C_{i,j}(t)=\dfrac{C_{i,j}(t)+C_{j,i}(t)}{2}$ and $C'(t)$:

$$C'(t) = C(t) + \sum_{\alpha=x,y,z} \exp(-\lambda_\alpha^{\mathrm{tr}}(t-t_0))\boldsymbol{d}_\alpha^{\mathrm{tr}}\boldsymbol{d}_\alpha^{\mathrm{tr}\mathrm{T}} + \sum_{\alpha=x,y,z} \exp(-\lambda_\alpha^{\mathrm{rot}}(t-t_0))\boldsymbol{d}_\alpha^{\mathrm{rot}}\boldsymbol{d}_\alpha^{\mathrm{rot}\mathrm{T}}, \tag{41}$$

where $\boldsymbol{d}_x^{\mathrm{tr}}$, $\boldsymbol{d}_y^{\mathrm{tr}}$, and $\boldsymbol{d}_z^{\mathrm{tr}}$ are unit vectors given by

$$\begin{aligned}\boldsymbol{d}_x^{\mathrm{tr}} &= \frac{1}{\sqrt{N}}(1,0,0,1,0,0,...,1,0,0)^{\mathrm{T}},\\ \boldsymbol{d}_y^{\mathrm{tr}} &= \frac{1}{\sqrt{N}}(0,1,0,0,1,0,...,0,1,0)^{\mathrm{T}},\\ \boldsymbol{d}_z^{\mathrm{tr}} &= \frac{1}{\sqrt{N}}(0,0,1,0,0,1,...,0,0,1)^{\mathrm{T}},\end{aligned} \tag{42}$$

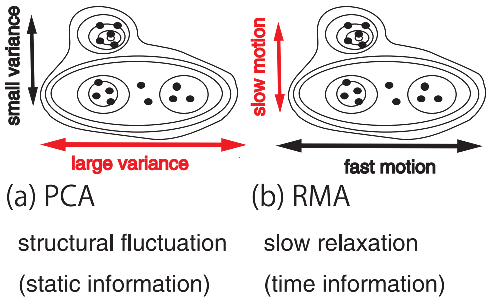

Figure 1 Schematic illustration of difference between PCA and RMA.

and $\boldsymbol{d}_x^{\mathrm{rot}}$, $\boldsymbol{d}_y^{\mathrm{rot}}$, and $\boldsymbol{d}_z^{\mathrm{rot}}$ are unit vectors given by

$$\begin{aligned}\boldsymbol{d}_x^{\mathrm{rot}} &= \frac{1}{\sqrt{\sum_{i=1}^{N}(\langle z_i\rangle^2+\langle y_i\rangle^2)}}\\ &\quad\times(0,-\langle z_1\rangle,\langle y_1\rangle,0,-\langle z_2\rangle,\langle y_2\rangle,...,0,-\langle z_N\rangle,\langle y_N\rangle)^{\mathrm{T}},\\ \boldsymbol{d}_y^{\mathrm{rot}} &= \frac{1}{\sqrt{\sum_{i=1}^{N}(\langle z_i\rangle^2+\langle x_i\rangle^2)}}\\ &\quad\times(\langle z_1\rangle,0,-\langle x_1\rangle,\langle z_2\rangle,0,-\langle x_2\rangle,...,\langle z_N\rangle,0,-\langle x_N\rangle)^{\mathrm{T}},\\ \boldsymbol{d}_z^{\mathrm{rot}} &= \frac{1}{\sqrt{\sum_{i=1}^{N}(\langle y_i\rangle^2+\langle x_i\rangle^2)}}\\ &\quad\times(-\langle y_1\rangle,\langle x_1\rangle,0,-\langle y_2\rangle,\langle x_2\rangle,0,...-\langle y_N\rangle,\langle x_N\rangle,0)^{\mathrm{T}}.\end{aligned} \tag{43}$$

The values of $\lambda_\alpha^{\mathrm{tr}}$ and $\lambda_\alpha^{\mathrm{rot}}$ are usually set to zero. These unit vectors satisfy the following relations:

$$\boldsymbol{d}_\alpha^a \cdot \boldsymbol{d}_\beta^b = \boldsymbol{d}_\alpha^{a\mathrm{T}}\boldsymbol{d}_\beta^b = \delta_{\alpha,\beta}\delta_{a,b} \tag{44}$$

and

$$C(t)\boldsymbol{d}_\alpha^a = 0, \tag{45}$$

where α, β=x, y, z and a, b=tr, rot. Then, we solve the generalized eigenvalue problem for $C'(t_0+\tau)$ and $C'(t_0)$, $C'(t_0+\tau)\boldsymbol{v}'_p=\exp(-\lambda'_p\tau)C'(t_0)\boldsymbol{v}'_p$, with the orthonormal condition $\boldsymbol{v}'^{\mathrm{T}}_p C'(t_0)\boldsymbol{v}'_q=\delta_{p,q}$. The unit vectors $\boldsymbol{d}_\alpha^a$ are eigenvectors of this generalized eigenvalue problem with eigenvalues $\exp(-\lambda_\alpha^a\tau)$. We denote $\boldsymbol{f}'_p$ as the eigenvectors other than $\boldsymbol{d}_\alpha^a$. Because $\boldsymbol{d}_\alpha^{a\mathrm{T}}C'(t)\boldsymbol{f}'_p=\exp(-\lambda_\alpha^a(t-t_0))\boldsymbol{d}_\alpha^{a\mathrm{T}}\boldsymbol{f}'_p=0$, $C'(t)\boldsymbol{f}'_p=C(t)\boldsymbol{f}'_p$ holds. Therefore, $\boldsymbol{f}'_p$ are identical with the eigenvectors $\boldsymbol{f}_p=(f_{p,1}, f_{p,2}, ..., f_{p,3N})^{\mathrm{T}}$ of the generalized eigenvalue problem for $C(t_0+\tau)$ and $C(t_0)$ with the same eigenvalues $\exp(-\lambda_p\tau)$. Thus, $\boldsymbol{f}_p$ and $\exp(-\lambda_p\tau)$ can be obtained by solving the generalized eigenvalue problem for $C'(t_0+\tau)$ and $C'(t_0)$, which are real symmetric positive definite matrices.

After calculating the relaxation modes and rates, we confirm whether the slow relaxation modes and rates obtained using τ and t_0 are appropriate. For this purpose, the convergences of slow relaxation times as a function of τ are examined. The autocorrelation functions $C_{i,i}(t)$ are reconstructed from the calculated eigenvalues and eigenvectors and are compared with those directly calculated via simulation (especially the slow relaxation behavior). After examining the validity, we use the obtained relaxation modes and rates for analysis.

Differences between PCA and RMA

Here, we briefly describe the static analysis method, PCA, and the difference between PCA and RMA. PCA is a well-known method for analyzing the static properties of structural fluctuations obtained via a simulation [4,6,34–38]. In PCA, the eigenvalue problem is solved as

$$\sum_{j=1}^{3N} C_{i,j}(0)F_{n,j} = \Lambda_n F_{n,i} \quad \text{with} \quad \sum_{i=1}^{3N} F_{m,i}F_{n,i} = \delta_{m,n}, \tag{46}$$

where $C_{i,j}(0)$ is the component of the $3N\times 3N$ variance-covariance matrix. Here, we set the indices of the eigenvalues to ensure that the relationship $\Lambda_1 \geq \Lambda_2 \geq \ldots \geq \Lambda_{3N}$ holds. The eigenvector $\boldsymbol{F}_n$ with the eigenvalue Λ_n is referred to as the nth principal component axis. Note that $\Lambda_{3N-5} = \Lambda_{3N-4} = \ldots = \Lambda_{3N} = 0$ because the translational and rotational degrees of freedom are removed. The coordinate $\boldsymbol{R}$ can be expanded in terms of the PCA eigenvectors:

$$R_i = \sum_{n=1}^{3N-6} \Phi_n F_{n,i} \quad \text{with} \quad \Phi_n = \sum_{n=1}^{3N} F_{n,i}R_i. \tag{47}$$

Here, Φ_n is referred to as the nth principal component. The variance of Φ_n is given by Λ_n. In PCA, the dimensionless free energy surface as a function of Φ_p and Φ_q of Eq. (47) is calculated as

$$F(\Phi_p, \Phi_q) = -\ln P(\Phi_p, \Phi_q), \tag{48}$$

where $P(\Phi_p, \Phi_q)$ denotes the probability density of Φ_p and Φ_q. In RMA, the quantity Y_p playing the same role as Φ_p in PCA is defined by

$$Y_p = X_p|\tilde{\boldsymbol{g}}_p|\,. \tag{49}$$

Then, the dimensionless free energy surface as a function of Y_p and Y_q is calculated as

$$F(Y_p, Y_q) = -\ln P(Y_p, Y_q), \tag{50}$$

where $P(Y_p, Y_q)$ denotes the probability density of Y_p and Y_q. Here, X_p is calculated from $\boldsymbol{R}$ as follows. Because of Eqs. (33) and (37), $\sum_{i=1}^{3N} f_{p,i}\, g_{i,q} = \delta_{p,q}$ holds, which leads to $\sum_{i=1}^{3N} f_{p,i}\tilde{g}_{i,q} = \mathrm{e}^{\lambda_p t_0/2}\delta_{p,q}$. Therefore, by multiplying $\boldsymbol{f}_p^{\mathrm{T}}$ on both sides of Eq. (40), X_p is given as a function of $\boldsymbol{R}$ as

$$X_p \simeq \sum_{i=1}^{3N} \tilde{f}_{p,i}R_i \tag{51}$$

with

$$\tilde{f}_{p,i} = \mathrm{e}^{-\lambda_p t_0/2} f_{p,i}\,. \tag{52}$$

From several results of PCA and RMA, we confirmed the difference between them. PCA studies the static properties of the fluctuations of structures and extracts modes with large variances. In contrast, RMA studies the dynamical properties of the fluctuations of structures and extracts modes with slow relaxation. The schematic of the difference between PCA and RMA is shown in Figure 1. Figures 1(a) and 1(b) show the distribution of conformations. PCA obtains the mode with large variance, as shown in Figure 1(a). RMA obtains the mode with slow relaxation, as shown in Figure 1(b). It is thought that the local minimum-energy states are stable, so that the system remains in this state for a long time. The order parameters with slow relaxation may correspond to the directions between local minimum-energy states. Thus, slow relaxation modes may be suitable order parameters to classify local minimum-energy states and describe transitions between them, especially in the case of simulations with large conformational changes. When the simulation involves large structural changes, the difference between local minimum-energy states is relatively small compared with that between the folded and unfolded states. In this case, extracting the effective modes or order parameters for accurately identifying the local minimum-energy states is difficult for PCA.

Reference Interaction Site Models

Analysis method such as PCA and RMA can classify stable or meta-stable structure from the MD trajectory. We would like to evaluate the stability of those structures, including the effects of the water solvent around proteins. This is because not only the conformational energy of protein but also the influence of hydration around the protein is important for its stability. The conformational energy and the solvation free energy (SFE) compete with each other because the gains in the conformational energy and the SFE are mainly owing to the intramolecular hydrogen bonds in the protein and intermolecular hydrogen bonds between protein and water, respectively. However, incorporating the solvent effect in practice is difficult. One of the methods to deal with the solvent effect is liquid theory based on statistical mechanics, and we apply it to biomolecules such as proteins.

RISM theory is one of the most successful statistical mechanical theories for molecular liquids. Chandler, D., *et al.* developed the theory mainly in the early days [88–92]. Hirata, F., *et al.* developed a generalization of the RISM theory, which is called the XRISM theory, to polar and quadrupolar liquids [93] and to ions in a molecular polar solvent [94]. Pettitt and Rossky applied the theory to calculate the liquid state structure of water in several three-site models [95]. Being able to treat the water molecule paved the way for application of the RISM theory to biomolecules in water. Pettitt and Karplus calculated the Ramachandran plot of alanine dipeptide in aqueous solution to compare with the vacuum surface calculations [96]. Kitao, Hirata, and Go calculated the free energy profiles along normal modes of melittin by using the XRISM theory to investigate the solvent effects for protein stability [8]. Following these pioneering studies of proteins, Kinoshita, Okamoto, and Hirata applied the 1D-RISM theory to peptides to investigate stable structures of peptides immersed in the solvent at infinite dilution [97–99]. To perform simulations of peptides with the solvent effects, 1D-RISM were combined with the MC simulated annealing [100,101] and generalized-ensemble algorithms such as the multicanonical and the replica-exchange methods [84,102].

However, only small solute molecules can be handled correctly because the conventional RISM theory expresses solvent structures as radial distribution functions. Beglov and Roux extended the RISM theory to three dimensions [103], and by Kovalenko and Hirata around the same time [104]. The 3D-RISM theory treats solvation structures as three-dimensional distribution functions, not radial distribution functions. (see Ref. 16 for more details.) This makes it possible to correctly handle large solute molecules such as proteins. This review focuses on SFE calculations of proteins in the 3D-RISM theory. In particular, we describe the 3D-RISM theory with atomic decomposition (AD) method to calculate contributions for each amino acid.

Three-dimensional reference interaction site model theory

The structural stability of proteins is of interest for investigating protein folding and protein-protein interaction mechanisms. SFE, in particular, is one of the most important properties to investigate to understand the thermodynamic stability of biomolecules, including protein folding. To investigate the stability of proteins, we introduce the idea of total energy G, which is given by the sum of the conformational energy E and the SFE $\Delta\mu$:

$$G = E + \Delta\mu. \tag{53}$$

We can easily calculate the conformational energy using the MD software. To calculate SFE we employ the 3D-RISM, which is the statistical mechanical theory for molecular liquids [103–105]. For a solute-solvent system at infinite dilution, the 3D-RISM equation is written as follows:

$$h_\gamma(\mathbf{r}) = \sum_{\gamma'} c_{\gamma'}(\mathbf{r}) * [w^{vv}_{\gamma'\gamma}(r) + \rho_{\gamma'} H_{\gamma'\gamma}(r)], \tag{54}$$

where $h_\gamma(\mathbf{r})$ and $c_\gamma(\mathbf{r})$ are the total 3D and direct correlation functions of the solvent site γ around the solute, the asterisk denotes a convolution integral in the real space, $w^{vv}_{\gamma'\gamma}(r)$ is the site-site intramolecular correlation function of the solvent, and $H_{\gamma'\gamma}(r)$ represents the site-site total correlation functions of pure solvent. The site-site correlation functions of the solvent are obtained in advance from the 1D-RISM theory for pure solvent.

The 3D-RISM equation contains two unknown functions; then, it is complemented with a closure equation. The most basic closure relation is expressed as [16,106],

$$h_\gamma(\mathbf{r}) = \exp(-\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}) + b_\gamma(\mathbf{r})) - 1, \tag{55}$$

where $u_\gamma(\mathbf{r})$ is the interaction potential acting on the solvent site, γ, of position $\mathbf{r}$, and $b_\gamma(\mathbf{r})$ is a bridge function that is usually unknown. The case $b_\gamma(\mathbf{r})=0$ corresponds to the hypernetted-chain (HNC) closure equation [104,107],

$$h_\gamma(\mathbf{r}) = \exp(-\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})) - 1. \tag{56}$$

In addition, the repulsive bridge correction (RBC) [108] and its modification (chemical bond-RBC) [109] have been proposed as methods for evaluating $b_\gamma(\mathbf{r})$. However, the HNC closure has poor convergence with the 3D-RISM equation. To avoid this difficulty, Kovalenko and Hirata proposed the partial-linearized HNC (PLHNC) or Kovalenko-Hirata (KH) closure equation [105,110],

$$h_\gamma(\mathbf{r}) = \begin{cases} \exp(\chi_\gamma) - 1 & (\chi_\gamma < 0) \\ \chi_\gamma & (\chi_\gamma \geq 0). \end{cases}$$

$$\chi_\gamma = -\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r}) \tag{57}$$

The combination of the KH closure and the 3D-RISM equations shows stable and rapid convergence. Furthermore, Kast and Kloss proposed a partial series expansion of order n (PSE-n) of the HNC closure [111],

$$h_\gamma(\mathbf{r}) = \begin{cases} \exp(\chi_\gamma) - 1 & (\chi_\gamma < 0) \\ \sum_{i=1}^{n} \frac{(\chi_\gamma)^i}{i!} & (\chi_\gamma \geq 0). \end{cases} \tag{58}$$

The closures interpolate between the KH and HNC closures. If $n=1$, it becomes the KH closure, and $n \to \infty$ indicates the HNC closure and gives its convergence problem. The efficiency of the 3D-RISM theory with the PSE-3 closure was verified by calculating the SFE of neutral amino acid side chain analogue molecules [112]. In addition, Kobryn, Gusarov, and Kovalenko recently proposed a new closure (the KGK closure) suitable for polar and charged macromolecules in an electrolyte solution [113]. The KGK closure equation is expressed as follows:

$$h_\gamma(\mathbf{r}) = \max\{-1; -\beta u_\gamma(\mathbf{r}) + h_\gamma(\mathbf{r}) - c_\gamma(\mathbf{r})\}. \tag{59}$$

This closure levels out the distribution function inside the repulsive core, particularly in regions where there are strong depletions. The RISM theory, when combined with the KGK closure, can produce the solvation structure and thermodynamics of oligomeric polyelectrolytes and drug-like compounds in electrolyte solution. These are molecules for which no convergence can be obtained with other closures.

The correlation functions are converged by calculating the 3D-RISM equation and the closure equation alternately. In the past decade, various methods have been proposed to accelerate convergence. These include the dynamic relaxation technique [114], the modified direct inversion in iterative subspace (MDIIS) method [114], multigrid techniques [115,116], and the modified Anderson method [117]. In addition to these, it became possible to apply 3D-RISM theory to biomolecules such as proteins using a graphics processing unit [117] and using parallel computers [118].

After convergence, the 3D distribution function $g_\gamma(\mathbf{r})$ is defined from $h_\gamma(\mathbf{r})$ using

$$g_\gamma(\mathbf{r}) = h_\gamma(\mathbf{r}) + 1. \tag{60}$$

Calculation of SFE using 3D-RISM theory

One way of obtaining the SFE is to calculate the following Kirkwood charging formula [119],

$$\Delta\mu = \frac{1}{\beta}\sum_\gamma \rho_\gamma \int_0^1 d\lambda \int d\mathbf{r} \frac{\partial u_\gamma(\mathbf{r};\lambda)}{\partial\lambda} g_\gamma(\mathbf{r};\lambda). \tag{61}$$

The coupling parameter, λ, changes the interaction potential from no interaction (λ=0) to full interaction (λ=1). $u_\gamma(\mathbf{r};\lambda)$ varies according to λ, and $g_\gamma(\mathbf{r};\lambda)$ denotes the distribution function under $u_\gamma(\mathbf{r};\lambda)$. To evaluate this formula with a low calculation error, we need to perform 3D-RISM calculations at least 40 times. To avoid the necessity of numerically coupling parameter integrations, Singer and Chandler derived the closed form using RISM and HNC closure equations [120]. The Singer-Chandler formula can easily be extended to three dimensions. Its HNC functional is expressed as follows:

$$\Delta\mu_{\mathrm{HNC}} = \frac{1}{\beta}\sum_\gamma \rho_\gamma \int d\mathbf{r}\left[\frac{1}{2}h_\gamma^2(\mathbf{r}) - c_\gamma(\mathbf{r}) - \frac{1}{2}c_\gamma(\mathbf{r})h_\gamma(\mathbf{r})\right]. \tag{62}$$

Similarly, the Singer-Chandler KH functional is expressed as [121]

$$\Delta\mu_{\mathrm{KH}} = \frac{1}{\beta}\sum_\gamma \rho_\gamma \int d\mathbf{r}\left[\frac{1}{2}h_\gamma^2(\mathbf{r})\Theta(-h_\gamma(\mathbf{r})) - c_\gamma(\mathbf{r}) - \frac{1}{2}c_\gamma(\mathbf{r})h_\gamma(\mathbf{r})\right], \tag{63}$$

where Θ denotes Heaviside step function. The Singer-Chandler formula is also equivalent to the Kirkwood charging formula for each closure equation. The closed SFE form for the PSE-n closures is expressed as [111,112]

$$\Delta\mu_{\mathrm{PSE}-n} = \Delta\mu_{\mathrm{HNC}} - \frac{1}{\beta}\sum_\gamma \rho_\gamma \int d\mathbf{r}\left[\frac{(h_\gamma(\mathbf{r}))^{n+1}}{(n+1)!}\Theta(h_\gamma(\mathbf{r}))\right]. \tag{64}$$

For n=1, this gives the KH functional, while $n\to\infty$, the second term vanishes.

One modification to the Singer-Chandler formula is the so-called Gaussian fluctuations (GF) approximation [122],

$$\Delta\mu_{\mathrm{GF}} = \frac{1}{\beta}\sum_\gamma \rho_\gamma \int d\mathbf{r}\left[-c_\gamma(\mathbf{r}) - \frac{1}{2}c_\gamma(\mathbf{r})h_\gamma(\mathbf{r})\right]. \tag{65}$$

Despite dropping the first term of the Singer-Chandler formula, the GF approximation improved the solvation thermodynamics in aqueous solution [123,124]. This formula is also used in combination with the KGK closure equation.

Furthermore, because it is known that the SFE value obtained by the Singer-Chandler formula is an overestimation [125,126] various correction methods have been proposed to improve its accuracy. Kovalenko and Hirata proposed the repulsive bridge extension of the HNC functional [108]. They also applied the method to investigate stability of Met-enkephalin [127]. Kido, K., *et al.* evaluated the SFE of various solute molecules in chloroform and benzene solvents using the chemical bond-RBC [109]. Tanimoto, S., *et al.* showed that the SFE expressions, based on RBC and partial wave (PW) extensions, provide more accurate results than those of the HNC or KH functionals [128]. These results indicate that the inclusion for molecular orientation dependencies contributes to the improvement of the SFE. In addition, the following corrections based on the phenomenological partial molar volume (PMV) corrections have been proposed: the universal corrections [129–133], the structural descriptor correction [134], the bridge function correction [135], and the pressure correction [136–140]. These corrections include the PMV expressed as follows: [141,142]

$$V = \frac{\kappa_{\mathrm{T}}}{\beta}\left(1 - \sum_\gamma \rho_\gamma \int d\mathbf{r} c_\gamma(\mathbf{r})\right), \tag{66}$$

where κ_{T} is the pure solvent isothermal compressibility. These corrections take the form $\Delta\mu + aV + b$, where a and b are the fitting parameters. The method by which each correction determines a and b differs, and b=0 in some of them. However, the question of which of these correction methods is the best remains controversial.

On the other hand, we derived a new SFE functional based on density functional theory (DFT). We introduced a hard sphere (HS) reference system to the DFT for polyatomic molecular liquids from which to derive the SFE functional of a solute molecule in water. We denoted it the reference-modified density functional theory (RMDFT) [143,144]. The RMDFT functional is expressed as

$$
\begin{aligned}
\Delta\mu_{\mathrm{RMDFT}} = &-\frac{1}{\beta}\sum_{\gamma}\rho_{\gamma}\int d\mathbf{r}h_{\gamma}(\mathbf{r})\\
&+\frac{\rho}{\beta}\sum_{\gamma}\sum_{\gamma'}\rho_{\gamma}\int d\mathbf{r}\int d\mathbf{r}'C^{ex}_{\gamma\gamma'}(|\mathbf{r}-\mathbf{r}'|)h_{\gamma'}(\mathbf{r})\\
&+\frac{1}{2\beta}\sum_{\gamma}\sum_{\gamma'}\rho_{\gamma}\rho'_{\gamma}\int d\mathbf{r}\int d\mathbf{r}'C^{ex}_{\gamma\gamma'}(|\mathbf{r}-\mathbf{r}'|)h_{\gamma}(\mathbf{r})h_{\gamma'}(\mathbf{r}')\\
&+\Delta F^{\mathrm{HS}}[\rho_{\mathrm{O}}]-\rho_{\mathrm{O}}\int d\mathbf{r}\left[\frac{\delta F^{\mathrm{HS}}[\rho_{\mathrm{O}}]}{\delta(\rho_{\mathrm{O}}h_{\mathrm{O}}(\mathbf{r}))}(h_{\mathrm{O}}(\mathbf{r})+1)-\mu^{\mathrm{HS}}_{\mathrm{O}}\right],
\end{aligned}
\tag{67}
$$

where

$$
C^{ex}_{\alpha\beta}(|\mathbf{r}-\mathbf{r}'|) = \begin{cases} \bar{C}_{\mathrm{OO}}(|\mathbf{r}-\mathbf{r}'|)-C^{\mathrm{HS}}_{\mathrm{OO}}(|\mathbf{r}-\mathbf{r}'|) & (\alpha=\beta=\mathrm{O})\\ \bar{C}_{\alpha\beta}(|\mathbf{r}-\mathbf{r}'|) & (\mathit{otherwise})\end{cases}
$$

$$
\bar{C}_{\alpha\beta}(|\mathbf{r}-\mathbf{r}'|) = C_{\alpha\beta}(|\mathbf{r}-\mathbf{r}'|)+C^{\mathrm{IM}}_{\alpha\beta}(|\mathbf{r}-\mathbf{r}'|)
$$

$$
\tilde{C}_{\alpha\beta}(k) = \frac{\delta_{\alpha\beta}}{\rho}-\left[\delta_{\alpha\beta}\rho+(1-\delta_{\alpha\beta})\rho\frac{\sin(kL_{\alpha\beta})}{kL_{\alpha\beta}}\right]^{-1}. \tag{68}
$$

Here, ρ is the average number density of the water solvent, $C_{\alpha\beta}(r)$ is the site-site direct correlation function of pure solvent, $C^{\mathrm{IM}}_{\alpha\beta}(r)$ is the intramolecular direct correlation function [145], $L_{\alpha\beta}$ is the length of the bond between α and β sites, O denotes the oxygen site of water solvent, $C^{\mathrm{HS}}_{\mathrm{OO}}(r)$ corresponds to the site-site direct correlation function of the reference HS fluid, $F^{\mathrm{HS}}[\rho_{\mathrm{O}}]$ is the excess intrinsic free energy functional for HS fluid, and $\mu^{\mathrm{HS}}_{\mathrm{O}}$ is the excess chemical potential of the reference HS fluid. We demonstrated that using the RMDFT functional can improve the absolute values of the SFE for a set of neutral amino acid side-chain analogues and for 504 small organic molecules. We have also shown that the RMDFT functional has the same effect as the PMV correction [146]. Furthermore, we used the 3D-RISM theory with the RMDFT functional to investigate the structural stability of proteins during the folding process based on Anton's long simulation [22].

Differences in the SFE values due to the structures of the proteins, as obtained by the Singer-Chandler formula, agree well with those of the RMDFT functional. We calculated the difference in the SFE between the native structure of chignolin and its other structures using the RMDFT and the Singer-Chandler KH functionals and showed that they were in good agreement [143,144]. This means that the Singer-Chandler formula yields differences in SFE between different states are not significantly different from experimental or correctly computed values.

Atomic decomposition method

We can calculate the SFE of the whole protein using the Singer-Chandler formula with or without the correction or the RMDFT functional. In this section, we consider how to treat the SFE of individual atoms in the protein. We introduce the atomic decomposition (AD) method proposed by Chong and Ham to calculate such a property [147,148], once again using the Kirkwood charging formula (Eq. (61)) as the starting point. If the SFE given by $\Delta\mu$ of Eq. (61) is the SFE of the whole solute, then we consider the decomposition of SFE into contributions from the individual atoms. $u_{\gamma}(\mathbf{r})$ is represented by the sum of the potentials between the solute atomic site, α, and the solvent site, γ,

$$
u_{\gamma}(\mathbf{r}) = \sum_{\alpha=1}^{N}u_{\alpha\gamma}(|\mathbf{r}-\mathbf{r}_{\alpha}|), \tag{69}
$$

where $\mathbf{r}_{\alpha}$ is the position of the atomic site, α, and N is the number of atomic sites in the solute. Then, we can obtain the following basic expressions from Eqs. (61) and (69),

$$
\Delta\mu = \sum_{\alpha=1}^{N}\Delta\mu_{\alpha}, \tag{70}
$$

and

$$
\Delta\mu_{\alpha} = \sum_{\gamma}\rho_{\gamma}\int_{0}^{1}d\lambda\int d\mathbf{r}\frac{\partial u_{\alpha\gamma}(|\mathbf{r}-\mathbf{r}_{\alpha}|;\,\lambda)}{\partial\lambda}g_{\gamma}(\mathbf{r};\,\lambda). \tag{71}
$$

The most commonly used form of the solute-solvent interaction potential, $u_{\alpha\gamma}(r)$, is given by a sum of the Lennard-Jones (LJ) and Coulomb electrostatic terms, as follows:

$$
u_{\alpha\gamma}(r) = u^{\mathrm{LJ}}_{\alpha\gamma}(r)+u^{\mathrm{elec}}_{\alpha\gamma}(r). \tag{72}
$$

Here, $u^{\mathrm{LJ}}_{\alpha\gamma}(r)=4\epsilon_{\alpha\gamma}[(\sigma_{\alpha\gamma}/r)^{12}-(\sigma_{\alpha\gamma}/r)^{6})]$ and $u^{\mathrm{elec}}_{\alpha\gamma}(r)=q_{\alpha}q_{\gamma}/r$, where $\epsilon_{\alpha\gamma}$, $\sigma_{\alpha\gamma}$, q_{α}, and q_{γ} are the LJ parameters and the atomic charges of the solute site α and solvent site γ, respectively. In the charging formula, it is necessary to treat LJ parameter and electrostatic potentials separately by introducing two coupling parameters, λ_1 and λ_2. The parameters are selected to scale the LJ parameter, $\sigma_{\alpha\gamma}$, and the atomic charge in the solute, q_{α}, respectively. Then the solute-solvent interaction potential is as follows:

$$
u_{\alpha\gamma}(r;\,\lambda_1,\lambda_2) = u^{\mathrm{LJ}}_{\alpha\gamma}(r;\,\lambda_1)+u^{\mathrm{elec}}_{\alpha\gamma}(r;\,\lambda_2), \tag{73}
$$

where

$$
u^{\mathrm{LJ}}_{\alpha\gamma}(r;\,\lambda_1) = 4\epsilon_{\alpha\gamma}\left[\left(\frac{\sigma_{\alpha\gamma}\lambda_1}{r}\right)^{12}-\left(\frac{\sigma_{\alpha\gamma}\lambda_1}{r}\right)^{6}\right], \tag{74}
$$

and

$$
u^{\mathrm{elec}}_{\alpha\gamma}(r;\,\lambda_2) = \frac{q_{\alpha}q_{\gamma}\lambda_2}{r}. \tag{75}
$$

Thus, the integration path is as follows: first, LJ interaction is performed from 0 to 1 with λ_2=0, then the electrostatic interaction begins when the integral of λ_2 is changed from 0 to 1 with λ_1=1. $\Delta\mu_{\alpha}$ is finally expressed as below:

$$
\begin{aligned}
\Delta\mu_{\alpha} = \sum_{\gamma}\rho_{\gamma}\Bigg[&\int_{0}^{1}d\lambda_1\int d\mathbf{r}\frac{\partial u_{\alpha\gamma}(|\mathbf{r}-\mathbf{r}_{\alpha}|;\,\lambda_1,\lambda_2=0)}{\partial\lambda_1}g_{\gamma}(\mathbf{r};\,\lambda_1,\lambda_2=0)\\
&+\int_{0}^{1}d\lambda_2\int d\mathbf{r}\frac{\partial u_{\alpha\gamma}(|\mathbf{r}-\mathbf{r}_{\alpha}|;\,\lambda_1=1,\lambda_2)}{\partial\lambda_2}g_{\gamma}(\mathbf{r};\,\lambda_1=1,\lambda_2)\Bigg].
\end{aligned}
\tag{76}
$$

After calculating $\Delta\mu_\alpha$ once, we can reproduce the contributions of the main- and side-chains of amino acid residues.

To calculate $\Delta\mu_\alpha$, we require $g_\gamma(\mathbf{r};\lambda_1,\lambda_2)$ or $h_\gamma(\mathbf{r};\lambda_1,\lambda_2)$. We used the 3D-RISM theory to obtain $h_\gamma(\mathbf{r};\lambda_1,\lambda_2)$ under $u_\gamma(\mathbf{r};\lambda_1,\lambda_2)$. We calculated $g_\gamma(\mathbf{r};\lambda_1,\lambda_2)$ at every integration step in Eq. (76).

The Ham group applied the AD method to their studies about the amyloid-beta protein: hydrohobicity [149], dimerization [150], aggregation [151–153], and self-assembly [154]. They also studied about the amyloidogenic potential of β-2-Microglobulin mutant [155], and protein-ligand binding thermodynamics [156].

Another method uses the spatial decomposition analysis (SDA) method, proposed by Yamazaki and Kovalenko, to decompose the solvation thermodynamics quantities [157]. This estimates the contribution of individual groups on the solute using Voronoi tessellation. They applied the SDA method to the stability analysis of four small proteins (chignolin, CLN025, Trp-cage, and FSD-1) [158]. The SDA method was also applied to the analysis of ion-protein binding [159].

Results

Simulation at a transition temperature

Chignolin, an artificial mini-protein designed by Honda, S., *et al.*, is made up of the 10 amino acids GYDPETGTWG [20]. It has been widely used to test new simulation algorithms and analysis methods [21,158,160–170]. It is characterized by two stable states, a native state and a misfolded state, near room temperature at 1 atm in MD simulations [166,168,171–175]. Both states have a common β-turn structure from Asp3 to Thr6 but have slightly different hydrogen-bond patterns for the backbone atoms (see Figs. 7(a) and 7(b)). The denaturation temperature of chignolin is low (315 K); therefore, its atomic coordinates were determined only using NMR. To improve its stability, Honda, S., *et al.* mutated the N- and C- termini. The chignolin mutant CLN025 contains the mutations of two amino acids at both terminals (G1Y and G10Y). The crystal structure of CLN025 indicated that it had the same topology in aqueous solution [176]. The misfolded structure was not seen even in MD simulations of CLN025 [177–179].

We performed a 750 ns MD simulation of chignolin in aqueous solution at 450 K. Previously, we had performed two several-μs simulations of chignolin at 300 K. In one simulation, chignolin folded to the native structure while in the other simulation it folded to the misfolded structure. After folding to either structure once, the latter were maintained. To generate various structural changes during a several-hundred-ns-long simulation, we performed the simulation close to the transition temperature. The time series of RMSD of C_α atoms from the native structure is shown in Figure 2. The native structure is the first coordinate of 1UAO.pdb. Many transitions occurred among the native, misfolded, and unfolded structures during the simulation.

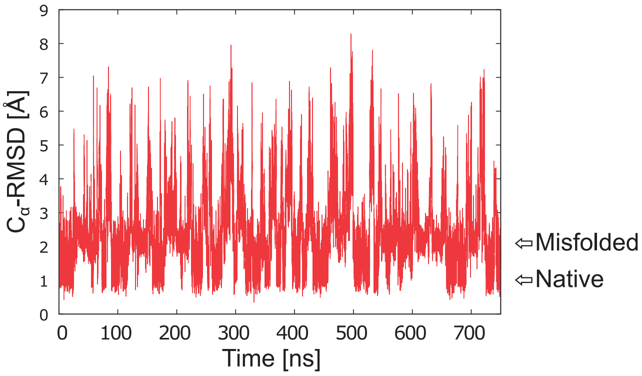

Figure 2 The time series of C_α-RMSD (Å) from the native structure of chignolin (PDB 1UAO model 1) near a transition temperature. There are several characteristic structures such as native structures (C_α-RMSD≈1 Å), misfolded structures (C_α-RMSD≈2 Å), and unfolded structures (C_α-RMSD≈5 Å). The figure was reproduced from Ref. 21.

Hereafter, we refer to the amide nitrogen atom and carbonyl oxygen atom on the main-chain of the *i*th amino acid as Xxx*i*N and Xxx*i*O, respectively. Here, Xxx is the three-letter code of the *i*th amino acid. Moreover, we refer to the nitrogen atom and oxygen atom on the side-chain of the *i*th amino acid as Xxx*i*N_s and Xxx*i*O_s, respectively. In addition, we refer to the hydrogen bond between atom A and atom B as H(A-B). We also refer to the distance between atom A and atom B as D(A-B).

Results of RMA

For analysis, we used the coordinates of C_α atoms on the backbone such that the number of the degrees of freedom was 30. PCA and RMA were carried out after the translational and rotational motions were removed from the coordinates of C_α atoms. For RMA, we set t_0 and τ to 10 ps and 20 ps, respectively. The suitable order parameters for identifying the native and misfolded structures in the chignolin system have been identified in previous studies as D(Asp3N-Gly7O) and D(Asp3N-Thr8O) [171]. As the native state includes H(Asp3N-Thr8O) and H(Gly1O-Gly10N) and the misfolded state includes H(Asp3N-Gly7O) and H(Glu1O-Thr9N), the order parameters were chosen based on these. Figure 3 shows the free-energy surfaces for the hydrogen bond distances (a), PCA (b), and RMA (c). The shape of the free-energy surface along D(Asp3N-Gly7O) and D(Asp3N-Thr8O) obtained by the present simulation is similar to that obtained by Refs. 166 and 171. These distances allow the native and misfolded states to be clearly distinguished. Although these distances are good order parameters, effective distances must be chosen and these depend on simulation systems. Figure 3(b) shows the free-energy surface obtained from PCA, where the PC modes were calculated automatically. The 1st and 2nd PC modes correspond to the directions with large variations in conformational fluctuation around an average structure. The native and mis-

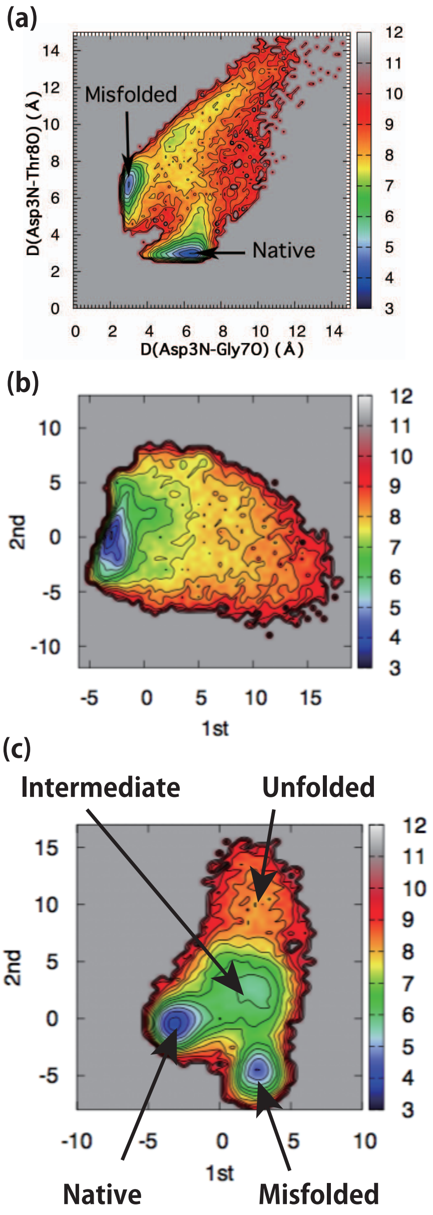

Figure 3 The free-energy surfaces along D(Asp3N-Gly7O) and D(Asp3N-Thr8O) (a), and along the first PC mode axis and the second PC mode axis (b), and along the slowest relaxation mode axis and the second slowest relaxation mode axis (c). The figure was reproduced from Ref. 21.

folded structures were not classified from the free-energy surface for the 1st and 2nd PC modes, because the conformational difference between them is low compared to the conformational fluctuations of the system. From RMA, we automatically obtained good order parameters to identify native and misfolded structures (Fig. 3(c)). The slow relaxation modes can be used to distinguish between the native and misfolded states. As the transition between the native and misfolded structures is slow, the slowest relaxation mode was found to be the axis distinguishing them. Interestingly, we could also identify the intermediate structure; by extracting the structures in the center part of the free-energy surface shown in Figure 3(c), a cluster was formed with a turn structure common to the native and misfolded structures (see Fig. 7 for structures). Thus, because the structures at both terminals fluctuate, a cluster of intermediate structures forming a turn is also obtained, while the fast relaxing movement of both terminals is ignored. The upper part of the free-energy surface shown in Figure 3(c) corresponded to the unfolded structure. The free-energy surface obtained from RMA shows the characteristic structures for the four states. RMA can identify the characteristic structure, even when it is only partially formed. In addition, it is evident that chignolin folds to the native or misfolded structures through the intermediate (turn) structure from the unfolded structures.

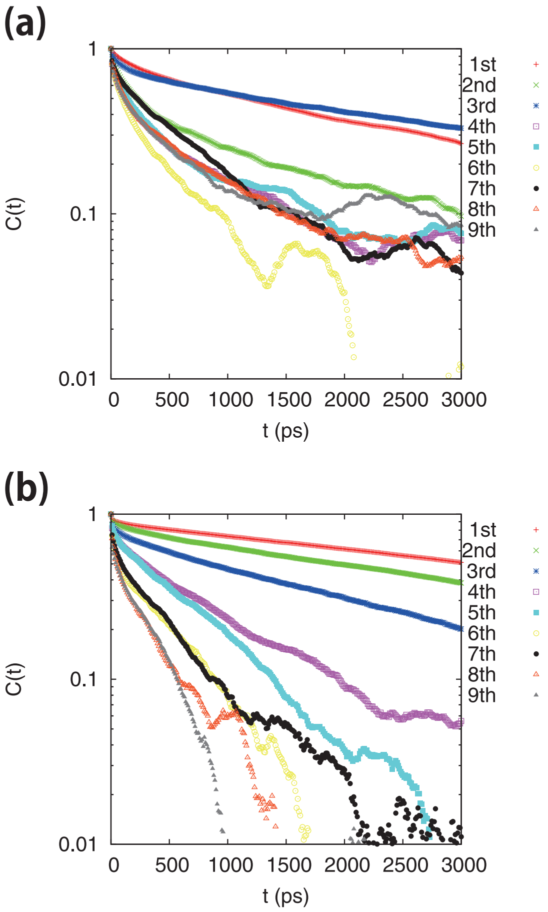

Figure 4 The time-displaced autocorrelation function for PCA (a) and RMA (b). The figure was reproduced from Ref. 21.

We calculated the time-displaced autocorrelation functions for PC modes and relaxation modes as shown in Figure 4. The 1st and 3rd PC modes (in red and blue, respectively) show slow relaxation, while the 2nd PC mode (in green) shows a relatively faster relaxation. The free energy surface for the 1st and 3rd PC modes distinguished the native and misfolded structures (see Fig. 2 of Ref. 21 for more detail),

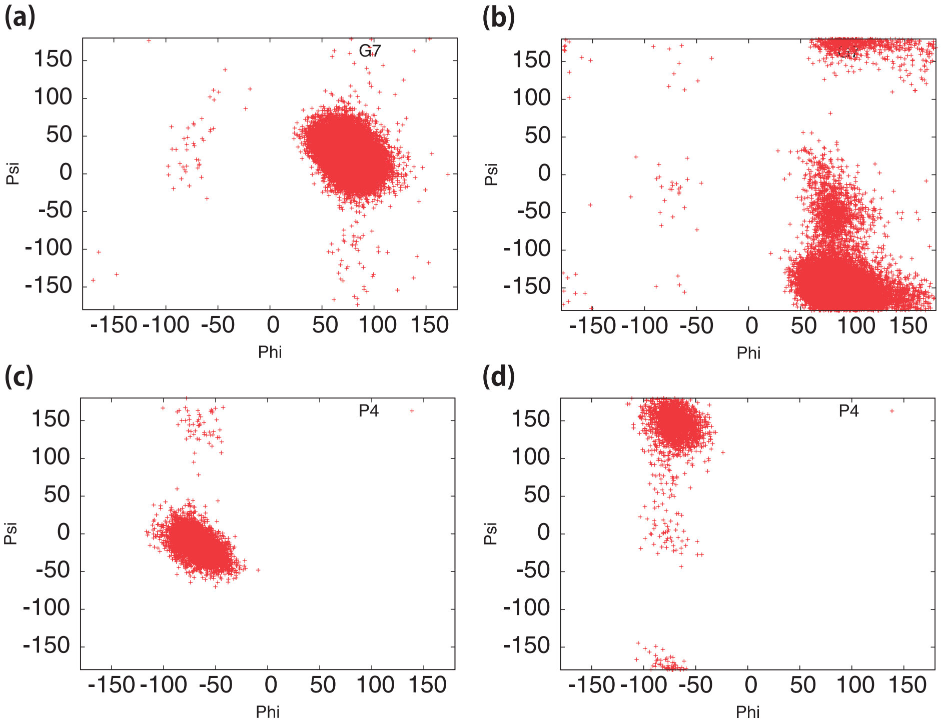

Figure 5 Ramachandran plots of Gly7 for the native (a) and misfolded (b) states and of Pro4 for the intermediate (c) and unfolded (d) structures. The figure was reproduced from Ref. 21.

indicating that it is more effective to use PC modes with slower relaxation rather than those with larger conformational fluctuations as the axes of the free-energy surface to classify the energy minimum states. The relaxation of the time-displaced autocorrelation function for the *p*th relaxation mode becomes gradually faster as *p* becomes larger. Thus, we succeeded in obtaining the order parameters with slow relaxation.

We examined the characteristic dihedral angles of the native, misfolded, intermediate, and unfolded states. The plots of each residue from Tyr2 to Thr6 for the native and misfolded states are similar to each other. The difference in the backbone dihedral angles of Gly7 causes the different hydrogen bond patterns observed between the native and misfolded states as shown in Figures 5(a) and 5(b). The plots of each residue from Asp3 to Glu5 for the intermediate state are also similar to those for the native and misfolded states, indicating the formation of a turn structure. These results demonstrate that the native, misfolded, and intermediate structures have the same turn structure. The main difference between the unfolded state and the other states is in the distribution of the dihedral angles of Pro4 as shown in Figures 5(c) and 5(d). This difference is responsible for the large RMSD value of residues from Asp3 to Glu5 of the unfolded state. Based on the structures of the four states obtained by RMA, we suggest that the dihedral angles are also good order parameters to classify the states in this system. In this study, we identified the characteristics of the native, misfolded, intermediate, and unfolded structures. Thus, we can calculate the stabilities of these states individually; we calculated their stabilities via solvent effect on these states.

Stability analysis using 3D-RISM theory

From the detailed analysis of the simulation near a transition temperature, we identify the native, misfolded, intermediate, and unfolded structures of chignolin. Thereafter, the relative stabilities between these structures using 3D-RISM theory at the amino acid level were examined (see Ref. 22 for more details). To examine detailed stabilities with solvent effect on these states, ensembles of these states at room temperature are needed. Thus, we performed a 5600-ns MD simulation at 1 atm and 298.15 K. The time series of RMSD of C_α from a native structure (1UAO.pdb model 1) is shown in Figure 6. The first-time transition from the unfolded state to the native state occurred around 100 ns in this trajectory. Next, the second-time transition from the native state to the misfolded state occurred around 1500 ns. After reaching to approximately RMSD≈5 Å, it settled in the misfolded state.

From the characteristics of these states, we extracted the obtained structures to the native, misfolded, intermediate, and unfolded states. The structures are shown in Figure 7. The red region indicates a side-chain of Tyr2, the blue region indicates that of Trp9, and the green region indicates side-chains of Thr6 and Thr8. The orange lines indicate intramolecular hydrogen bonds. The purple portion indicates a β-turn in the intermediate state. The native and misfolded

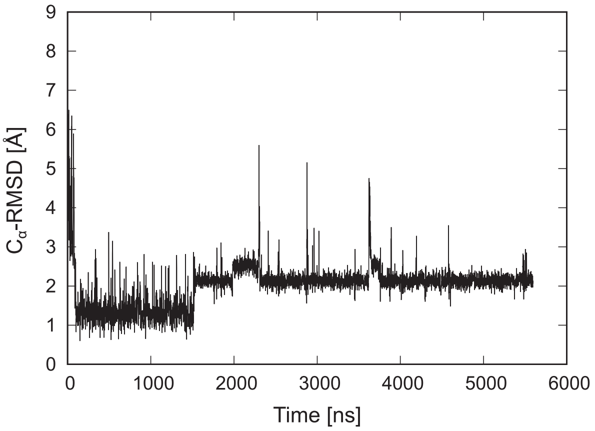

Figure 6 The time series of C_α-RMSD value from a native structure (PDB 1UAO model 1).

Table 1 Average values of total energy, G, conformational energy, E, and solvation free energy, $\Delta\mu$, of each state

Type	$G=E+\Delta\mu$	E	$\Delta\mu$
native	−171.1±9.4	−29.1±17.3	−142.0±13.4
misfolded	−171.2±9.0	−12.8±16.6	−158.4±13.7
intermediate	−158.5±8.5	47.4±29.9	−205.9±27.6
unfolded	−148.1±8.7	103.2±15.4	−251.3±13.4

Energy unit is kcal/mol.

states have a common β-turn structure from Asp3 to Thr6 but slightly different hydrogen bond patterns of the backbone. The differences between the native and misfolded states except for the different hydrogen bonds were observed owing to the arrangements of the side-chains of Thr8 and Trp9 because the value of ψ of the dihedral angle of Gly7 between the two states was different [21]. In addition, the side-chains of Thr6 and Thr8 in the native structure are located at the same side, while those of the misfolded structure are located at opposite sides. It forms the characteristic β-turn in the part from Asp3 to Thr6 for the intermediate state, even with an expanded shape as shown in Figure 7(c). This turn is a common structure of the native and misfolded states. This fact indicates that through the intermediated state, chignolin acquires a compact native or misfolded state (see Ref. 21). However, the unfolded state (Fig. 7(d)), is fully extended, and thus has no β-turn structure.

Next, we determine the stability by investigating the average values of the total energy given by the sum of the conformational energy and SFE. Table 1 shows the average values of the total energy, G, conformational energy, E, and SFE, $\Delta\mu$, of each state. The native (−171.1 kcal/mol) and misfolded (−171.2 kcal/mol) states have lower the G than the intermediate (−158.5 kcal/mol) and unfolded (−148.1 kcal/mol) states. This means that the native and misfolded states have similar stabilities and are more stable than the intermediate and unfolded states. The G value of the intermediate state is between that of the compact states

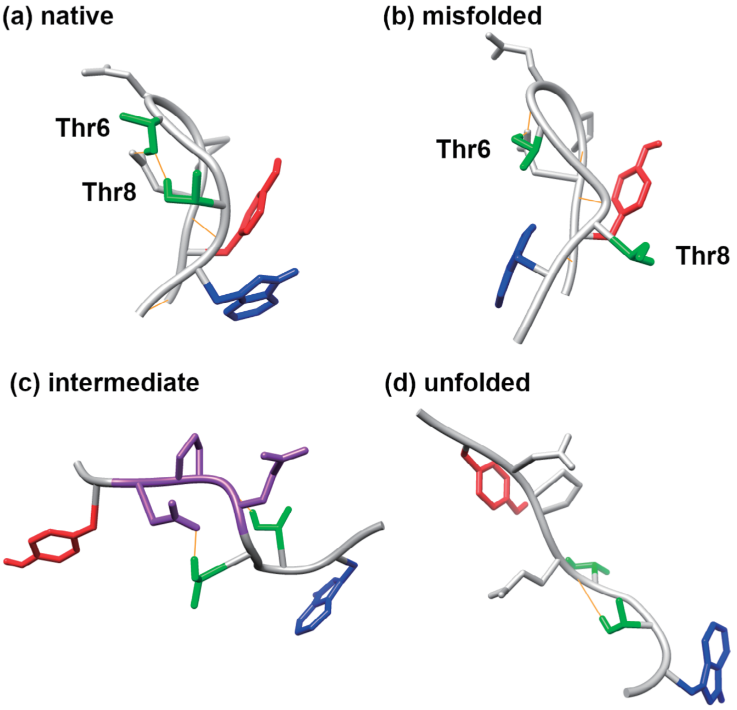

Figure 7 The characteristic structures of the native (a), misfolded (b), intermediate (c), and unfolded (d) structures. The red region indicates a side-chain of Tyr2, the blue region indicates that of Trp9, and the green region indicates side-chains of Thr6 and Thr8. In the intermediate state, the purple portion indicates a β-turn. The orange lines are the intramolecular hydrogen bonds.

(native and misfolded states) and the unfolded state. A clear inverse correlation between conformational energy, E, and SFE, $\Delta\mu$, is observed in Table 1. The compact states (native and misfolded states) have lower conformational energy than unfolded states (intermediate and unfolded states). The unfolded states have lower the SFE than the compact states. This is because the intermediate and unfolded states are extended and form intermolecular hydrogen bonds between the protein and water around the protein. On the other hand, the native and misfolded states are compact and form intramolecular hydrogen bonds in the protein. The results are similar to those of Refs. 84,179–181. There is a balance between the intramolecular hydrogen bonds in protein and the intermolecular hydrogen bonds between protein and water. The difference in total energy between different states is reduced by the competition between conformational energy and SFE. Therefore, the competition renders the free energy surface smooth. The solvation effect almost cancels the conformational stability, but contributes slightly to the stability of the native and misfolded states. Although the total energy differences between the native and misfolded states are similar to each other, the mechanisms of structural stability of the native and misfolded states differ. The average values of conformational energy of the native states are lower than those of the misfolded states. However, the SFE enhances the stability of misfolded states.

To investigate the stabilization mechanism of each amino acid during the process of folding, the differences between the average total energy (ΔG), average conformational energy (ΔE), and average SFE ($\Delta\Delta\mu$) of the main- and side-chains of each residue between the unfolded and intermediate states, the intermediate and misfolded states, and the intermediate and native states were calculated. The difference in average conformational energies may be affected by the difference in the structures of the two states. While the intermolecular hydrogen bond between protein and water contributes to the SFE, the intramolecular hydrogen bond in protein contributes to the conformational energy.

Figure 8(a) shows the energy differences between the unfolded and intermediate states of main- and side-chains for each amino acid. Black bars indicate total energy difference of each amino acid, ΔG, light gray ones represent conformational energy differences of main-chain for each amino acid, ΔE^{M}, dark gray ones represent the SFE differences of main-chain, $\Delta\Delta\mu^{\mathrm{M}}$, white ones denote conformational energy differences of side-chain for each amino acid, ΔE^{S}, and gray ones indicate the SFE differences of side-chain, $\Delta\Delta\mu^{\mathrm{S}}$. Here, the positive value indicates that the intermediate state is stable. For each residue, the competition between the conformational energy and the SFE is also observed. For example, the conformational energies for terminus have positive values while the SFE for terminus have negative values. As a result, the differences in total stability between the states at the both termini are small. In other words, the contribution to structural stability is small at both termini. The other residues also have similar competition between conformational energy and SFE; after considering both these parameters, we discuss the contribution to total stability for each residue. Figure 8(b) shows the energy differences between the unfolded and intermediate states for main-chain (black bar) and side-chain (white bar). The allows also indicates the intramolecular hydrogen bonds. The total energies of the main-chains of Thr6 and Gly7 are stabilized due to the hydrogen bonds, H(Asp3O-Thr6N) and H(Asp3O-Thr7N), and the total energy of the side-chain of Thr6 is stabilized due to the hydrogen bond between Asp3Os and Thr6Os. Pro4 in the intermediate state is more stable than in the unfolded state because of different dihedral angle of Pro4. The turn from Asp3 to Thr6 and the side-chain of Trp9 stabilizes the intermediate state. Intramolecular hydrogen bonds make structure stable even after considering both the energy terms.

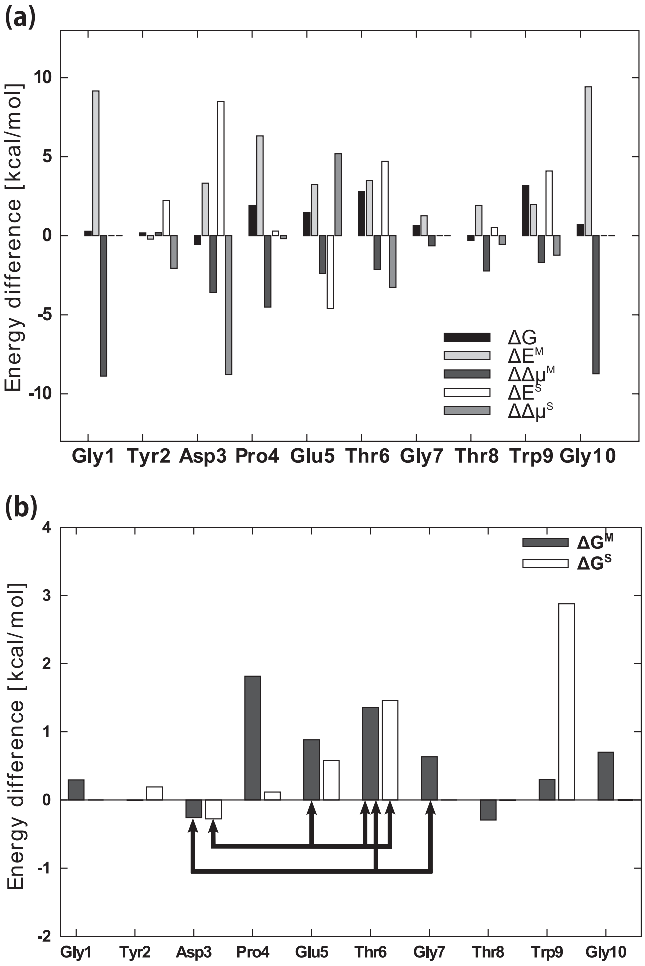

Figure 8 Differences in average total energy, ΔG, average conformational energy, ΔE, and average solvation free energy, $\Delta\mu$, of main chains (superscript M) and side chains (superscript S) of each residue of unfolded state from intermediate state (a). Main- and side-chain components of total energy difference of unfolded state from intermediate state (b). The arrows indicate hydrogen bonds.

Figure 9(a) shows the total energy differences between the intermediate and misfolded states, where the positive value

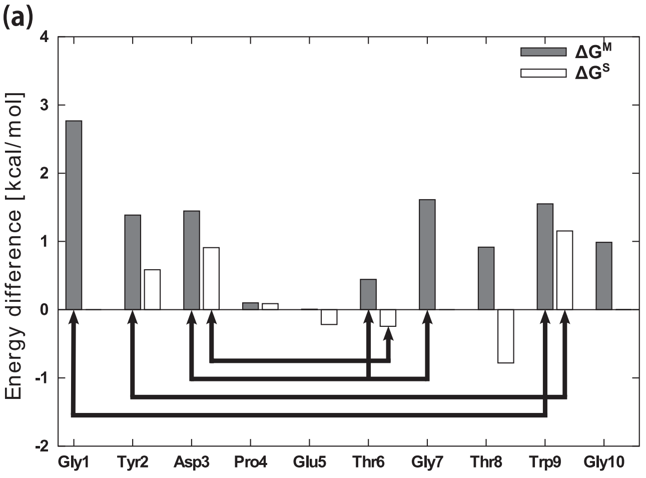

Figure 9 Main- and side-chain components of total energy difference, ΔG^{M} and ΔG^{S} of intermediate state from misfolded state (a), and intermediate state from native state (b). The arrows indicate hydrogen bonds.

indicates that the misfolded state is stable. The differences between the total energies of the main-chains from Pro4 to Thr6 are small because of turn formation from Asp3 to Thr6. Asp3 in the turn is more stable in the misfolded state than in the intermediate state. In addition, Gly1, Tyr2, Asp3, Gly7, and Trp9 are stable in the misfolded state rather than in the intermediate state.

The total energy differences between the intermediate and native states are shown in Figure 9(b). Here, the positive value indicates that the native state is stable. The differences between the total energies of the main-chains from Asp3 to Thr6 are also small because both the structures form the turn from Asp3 to Thr6. Thr6 is more stable in the native state, while Asp3 is more stable in the misfolded state. The total energy gain for the main-chains of Gly7 and Thr8 are due to the hydrogen bonds between the atoms of the main-chains, H(Asp3O-Gly7N) or H(Asp3O-Thr8N). Similary, that for the main-chains of Gly1 and Gly10 are due to the hydrogen bond, H(Gly1O-Gly10N), and that for the side-chains of Thr6 and Thr8 are due to the hydrogen bonds between the atoms of the side-chains, H(Asp3Os-Thr6Os) and H(Thr6Os-Thr8Os). Beside of no characteristic hydrogen bond in the side-chain of Tyr2, the conformational energy is more stabilized. This can be attributed to be the packing effect of Tyr2.

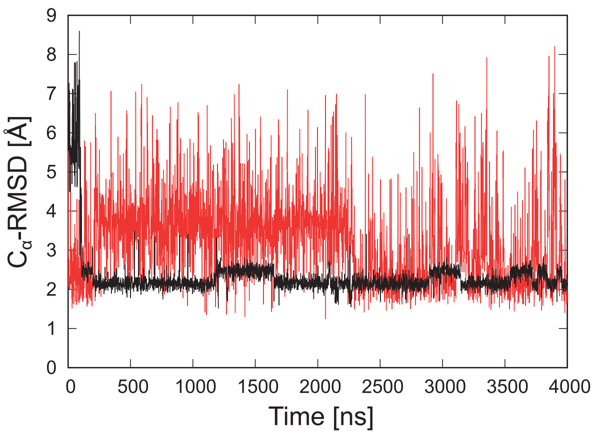

Figure 10 The time series of C_α-RMSD value of the T8P mutant from a native structure (PDB 1UAO model 1) at 298.15 K (black) and 420 K (red).

We propose the mutation of chignolin for the stabilization of the misfolded structure; we consider a residue, which stabilizes the native state but does not affect the stability of the misfolded state. The native structure is stabilized by Thr6 and Thr8 due to hydrogen bonding between the side chains of Thr6 and Thr8. In contrast, Thr8 was not involved in stabilization of the misfolded state (see Figs. 7(a), 7(b), and 9). We concluded that mutation of Thr8 to a neutral amino acid may improve the stability of the misfolded state. In addition, the mutation of Thr8 may be effective in changing the relative stabilization of the native and misfolded states. Thus, the mutation of Thr8 to a neutral amino acid may affect the relative stability between the misfolded state and the native state; the misfolded state becomes more stable than the native state. Then, we investigated the effects of Thr8 mutations on the structural stability of chignolin. We generated mutants in which Thr8 was mutated to 19 other amino acids and performed 4-μs MD simulations at room temperature. From these simulations, it was noted that five mutants (T8I, T8F, T8P, T8N, and T8Y) did not form the native structure; instead these favored the misfolded structure. In addition, MD simulations at 420 K were performed to increase sampling of the structures for these mutants. Among them, T8P formed the misfolded structure even at high temperatures. The time series of the C_α-RMSD values for T8P at 298.15 K and at 420 K are shown in Figure 10. The C_α-RMSD values of the mutant were stable with the lower limit being around 2.0 Å, which corresponds with the misfolded state. The lowest total energy structure of T8P mutant at 298.15 K is shown in Figure 11(a). It can be seen that the structure of

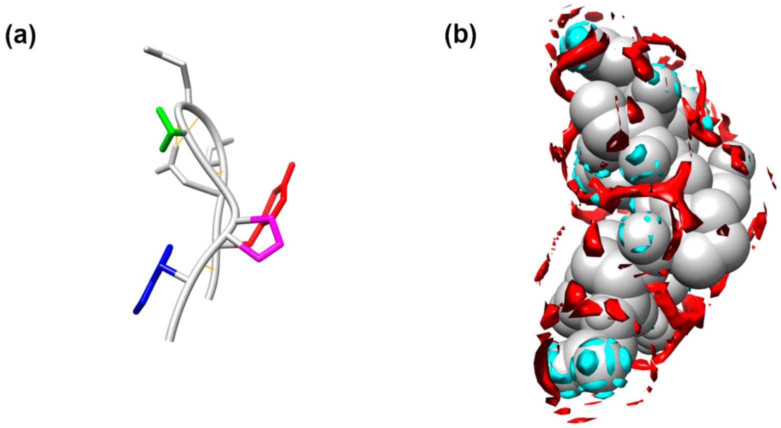

Figure 11 The lowest total energy structure of the T8P mutant at 298.15 K. In (a), the red region indicates a side-chain of Tyr2, the blue region indicates that of Trp9, the green region indicates that of Th6, and the purple region indicates that of Pro8. The orange lines are the intramolecular hydrogen bonds. In (b), the distribution of water oxygen (red) and hydrogen (cyan) of the structure is shown.

T8P mutant is similar to the misfolded structure in Figure 7(b). The side-chains of Thr6 and Pro8 are located at opposite sides, and β-turn structure from Asp3 to Thr6 is formed. Then, the side chain of Thr6 did not make a hydrogen bond with that of Pro8. The side-chain of proline has a cyclic structure and is not flexible. Therefore, owing to steric hindrance, the T8P mutant cannot take the dihedral angle to form the native structure. Based on these results, the misfolded structure is the most stable state for the T8P mutant. Finally, the distributions of water oxygen (red) and hydrogen (cyan) of the structure are shown in Figure 11(b). The threshold values are set to 3.0, which are three times the probability of bulk water. It can be seen that even small protein, chignolin, has complex hydration structures. Therefore, 3D-RISM theory can help study the effect of hydrogen bond between protein and water with higher specificity and accuracy unlike the continuum approximation.

Conclusions

Professor Nobuhiro Go is one of the pioneers who introduced the concepts of statistical mechanics into protein research using computers. His studies vastly contributed to the fields of computational chemistry and physics in protein simulations. More specifically, he introduced the idea of mode decomposition for the analysis protein motions. His works involving PCA proved significant in introducing RMA to protein systems in our works.

In this review, we introduce the analysis methods for molecular simulations of proteins based on statistical mechanics, RMA method and 3D-RISM theory. RMA investigates dynamics and kinetics of simulations with large conformational changes and extract characteristic states of proteins. Then we calculate the SFE of these states using 3D-RISM theory and investigate their stabilities with solvent effects. We review the results for designing amino-acid substitution of 10-residue peptide, chignolin, to stabilize the misfolded structure using these analytical methods. Furthermore, studies on the effects of mutations on structural stability of protein are critical for understanding changes in protein function. Structural information of meta-stable states is useful in altering the relative stability between the native and meta-stable structures and for designing new structures. Using the powerful analysis methods, we suggested possible mutations in chignolin that could stabilize the misfolded structure (i.e., a meta-stable state) [22]. In future works, we aim to apply the computational approaches to larger proteins and design new structures based on information of meta-stable structures.

Acknowledgments

This work was supported by the Japan Science and Technology Agency, Precursory Research for Embryonic Science and Technology (JST PRESTO) (JPMJPR13LB). Numerical calculations were performed in part using the Highly Accelerated- Parallel Advanced System for Computer Sciences (HA-PACS) at the Center for Computational Sciences (CCS), University of Tsukuba. Molecular graphics for Figures 7 and 11 were designed using UCSF Chimera, which was developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIGMS P41-GM103311) [182].

The authors would like to thank Dr. Toshiki Nagai, Mr. Taku Yamamoto, Mr. Yuta Koizumi, Dr. Satoshi Natori, and Dr. Naoyuki Karasawa of Keio University for fruitful discussions regarding RMA.

Conflicts of interest

The authors declare no competing financial interests.

Author contributions

Y. M., H. T., A. M. wrote the manuscript.

References

[1] Go, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210 (1983).

[2] Braun, W. & Go, N. Calculation of protein conformations by proton-proton distance constraints: A new efficient algorithm. *J. Mol. Biol.* **186**, 611–626 (1985).

[3] Go, N., Noguti, T. & Nishikawa, T. Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl. Acad. Sci. USA* **80**, 3696–3700 (1983).

[4] Kitao, A., Hirata, F. & Go, N. The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chem. Phys.* **158**, 447–472 (1991).

[5] Kitao, A., Hirata, F. & Go, N. Effects of solvent on the conformation and the collective motions of a protein. 2. structure of hydration in melittin. *J. Phys. Chem.* **97**, 10223–10230 (1993).

[6] Hayward, S., Kitao, A., Hirata, F. & Go, N. Effect of solvent on collective motions in globular protein. *J. Mol. Biol.* **234**, 1207–1217 (1993).

[7] Kitao, A. & Go, N. Investigating protein dynamics in collective coordinate space. *Curr. Opin. Struct. Biol.* **9**, 164–169 (1999).

[8] Kitao, A., Hirata, F. & Go, N. Effects of solvent on the conformation and the collective motions of a protein. 3. free energy analysis by the extended RISM theory. *J. Phys. Chem.* **97**, 10231–10235 (1993).

[9] Kitao, A., Hayward, S. & Go, N. Energy landscape of a native protein: jumping-among-minima model. *Proteins* **33**, 496–517 (1998).

[10] Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., *et al.* Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **51**, 91–97 (2008).

[11] Shaw, D. E., Grossman, J. P., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., *et al.* Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. in *Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis.* pp. 41–53 (Piscataway, NJ, USA, IEEE Press, 2014).

[12] Ohmura, I., Morimoto, G., Ohno, Y., Hasegawa, A. & Taiji, M. MDGRAPE-4: a special-purpose computer system for molecular dynamics simulations. *Philos. Trans. A Math. Phys. Eng. Sci.* **372**, 20130387 (2014).

[13] Takano, H. & Miyashita, S. Relaxation modes in random spin systems. *J. Physical Soc. Japan* **64**, 3688–3698 (1995).

[14] Koseki, S., Hirao, H. & Takano, H. Monte Carlo study of relaxation modes of a single polymer chain. *J. Physical Soc. Japan* **66**, 1631–1637 (1997).

[15] Hirao, H., Koseki, S. & Takano, H. Molecular dynamics study of relaxation modes of a single polymer chain. *J. Physical Soc. Japan* **66**, 3399–3405 (1997).

[16] Hirata, F. ed. *Molecular Theory of Solvation* (Kluwer Academic Publishers, Dordrecht, 2003).

[17] Mitsutake, A., Iijima, H. & Takano, H. Relaxation mode analysis of homopolymer systems. *Bussei Kenkyu* **85**, 376–381 (2005). (in Japanese)

[18] Mitsutake, A., Iijima, H. & Takano, H. Principal component analysis and relaxation mode analysis of a peptide. *Seibutsu Butsuri* **45 Supplement**, S214 (2005). (Abstract for the 43th Annual meeting, The biophysical society of Japan) (in Japanese)

[19] Mitsutake, A., Iijima, H. & Takano, H. Relaxation mode analysis of a peptide system: comparison with principal component analysis. *J. Chem. Phys.* **135**, 164102 (2011).

[20] Honda, S., Yamasaki, K., Sawada, Y. & Morii, H. 10 residue folded peptide designed by segment statistics. *Structure* **12**, 1507–1518 (2004).

[21] Mitsutake, A. & Takano, H. Relaxation mode analysis and Markov state relaxation mode analysis for chignolin in aqueous solution near a transition temperature. *J. Chem. Phys.* **143**, 124111 (2015).

[22] Maruyama, Y. & Mitsutake, A. Analysis of structural stability of chignolin. *J. Chem. Phys. B* **122**, 3801–3814 (2018).

[23] Mitsutake, A. & Takano, H. Relaxation mode analysis for molecular dynamics simulations of proteins. *Biophys. Rev.* **10**, 375–389 (2018).

[24] Brooks, B. & Karplus, M. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* **80**, 6571–6575 (1983).

[25] Levitt, M., Sander, C. & Stern, P. S. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.* **181**, 423–447 (1985).

[26] Lamm, G. & Szabo, A. Langevin modes of macromolecules. *J. Chem. Phys.* **85**, 7334–7348 (1986).

[27] Kottalam, J. & Case, D. A. Langevin modes of macromolecules: applications to crambin and DNA hexamers. *Biopolymers* **29**, 1409–1421 (1990).

[28] Tirion, M. M. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908 (1996).

[29] Baher, I., Atilgan, A. R. & Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **2**, 173–181 (1997).

[30] Tama, F. & Sanejouand, Y.-H. Conformational change of proteins arising from normal mode calculations. *Protein Eng.* **14**, 1–6 (2001).

[31] Cui, Q. & Bahar, I. eds. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (Chapman and Hall/CRC, Boca Raton, 2005).

[32] Miyashita, O. & Tama, F. Coarse-Grained Normal Mode Analysis to Explore Large-Scale Dynamics of Biological Molecules. in *Coarse-graining of condensed phase and biomolecular systems.* (Voth, G. A. ed.) (CRC Press, New York, 2008).

[33] Tama, F. & Brooks 3rd, C. L. The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.* **318**, 733–747 (2002).

[34] Levy, R. M., Srinivasan, A. R., Olson, W. K. & McCammon, J. A. Quasi- harmonic method for studying very low frequency modes in proteins. *Biopolymers* **23**, 1099–1112 (1984).

[35] Ichiye, T. & Karplus, M. Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Protein* **11**, 205–217 (1991).

[36] Abagyan, R. & Argos, P. Optimal protocol and trajectory visualization for conformational searches of peptides and proteins. *J. Mol. Biol.* **225**, 519–532 (1992).

[37] García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **68**, 2696–2699 (1992).

[38] Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **17**, 412–425 (1993).

[39] Lange, O. F. & Grubmüller, H. Full correlation analysis of conformational protein dynamics. *Proteins* **70**, 1294–1312 (2007).

[40] Sakuraba, S., Joti, Y. & Kitao, A. Detecting coupled collective motions in protein by independent subspace analysis. *J. Chem. Phys.* **133**, 185102 (2010).

[41] Kamada, M., Toda, M., Sekijima, M., Takata, M. & Joe, K. Analysis of motion features for molecular dynamics simula-

tion of proteins. *Chem. Phys. Lett.* **502**, 241–247 (2011).

[42] Moritsugu, K., Koike, R., Yamada, K., Kato, H. & Kidera, A. Motion tree delineates hierarchical structure of protein dynamics observed in molecular dynamics simulation. *PLoS One* **10**, e0131583 (2015).

[43] Matsunaga, Y., Kidera, A. & Sugita, Y. Sequential data assimilation for single-molecule FRET photon-counting data. *J. Chem. Phys.* **142**, 214115 (2015).

[44] Iba, Y., Fujisaki, H. & Matsunaga, Y. Special topic: conformational fluctuations and dynamics of biomolecules—statistical analusis of computer simulation and experimental data. *Proceedings of the Institute of Statistical Mathematics* **62**, 163–170 (2014). (in Japanese)

[45] Ito, R. & Yoshidome, T. An accurate computational method for an order parameter with a Markov state model constructed using a manifold-learning technique. *Chem. Phys. Lett.* **691**, 22–27 (2018).

[46] Fujisaki, H., Moritsugu, K., Mitsutake, A. & Suetani, H. Conformational change of a biomolecule studied by the weighted ensemble method: Use of the diffusion map method to extract reaction coordinates. *J. Chem. Phys.* **149**, 134112 (2018).

[47] Zuckerman, D. M. *Statistical Physics of Biomolecules: An Introduction* (CRC Press, New York, 2010).

[48] Komatsuzaki, T., Berry, R. S. & Leitner, D. M. *Advancing theory for kinetics and dynamics of complex, many-dimensional systems* (Wiley, Canada, 2011).

[49] Bowman, G. R., Pande, V. S. & Noé, F. ed. *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* (Springer, Dordrecht, 2014).

[50] Naritomi, Y. & Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *J. Chem. Phys.* **134**, 065101 (2011).

[51] Naritomi, Y. & Fuchigami, S. Slow dynamics of a protein backbone in molecular dynamics simulation revealed by time-structure based independent component analysis. *J. Chem. Phys.* **139**, 215102 (2013).

[52] Pérez-Hernández, G., Paul, F., Giorgino, T., Fabritiis, D. G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).

[53] Schwantes, C. R. & Pande, V. S. Improvements in Markov state model construction reveal many non-native interactions in the folding of NTL9. *J. Chem. Theory Comput.* **9**, 2000–2009 (2013).

[54] Mori, T. & Saito, S. Dynamic heterogeneity in the folding/unfolding transitions of FiP35. *J. Chem. Phys.* **142**, 135101 (2015).

[55] Mori, T. & Saito, S. Molecular mechanism behind the fast folding/unfolding transitions of villin headpiece subdomain: Hierarchy and heterogeneity. *J. Phys. Chem. B* **120**, 11683–11691 (2016).

[56] Schütte, Ch., Fischer, A., Huisinga, W. & Deuflhard, P. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.* **151**, 146–168 (1999).

[57] Swope, W. C., Pitera, J. W. & Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. theory. *J. Phys. Chem. B* **108**, 6571–6581 (2004).

[58] Singhal, N., Snow, C. D. & Pande, V. S. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* **121**, 415–425 (2004).

[59] Chodera, J. D., Swope, W. C., Pitera, J. W. & Dill, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.* **5**, 1214–1226 (2006).

[60] Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A. & Swope, W. C. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101 (2007).

[61] Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).

[62] Noé, F., Horenko, I., Schütte, C. & Smith, J. C. Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* **126**, 155102 (2007).

[63] Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromloecules. *Curr. Opin. Struct. Biol.* **18**, 154–162 (2008).

[64] Noé, F. & Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **43**, 141–147 (2017).

[65] Buchete, N. & Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **112**, 6057–6069 (2008).

[66] Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., *et al.* Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).

[67] Schwantes, C. R., McGibbon, R. T. & Pande, V. S. Perspective: Markov models for long-timescale biomolecular dynamics. *J. Chem. Phys.* **141**, 090901 (2014).

[68] Wu, H., Nüske, F., Paul, F., Klus, S., Koltai, P. & Noé, F. Variational Koopman models: slow collective variables and molecular kinetics from short off-equilibrium simulations. *J. Chem. Phys.* **146**, 154104 (2017).

[69] Schütte, C., Noé, F., Lu, J., Sarich, M. & Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **134**, 204105 (2011).

[70] Hagita, K. & Takano, H. Relaxation mode analysis of a single polymer chain in a melt. *J. Physical Soc. Japan* **71**, 673–676 (2002).

[71] Saka, S. & Takano, H. Relaxation of a single knotted ring polymer. *J. Physical Soc. Japan* **77**, 034001 (2008).

[72] Iwaoka, N., Hagita, K. & Takano, H. Estimation of relaxation modulus of polymer melts by molecular dynamics simulations: application of relaxation mode analysis. *J. Physical Soc. Japan* **84**, 044801 and references therein (2015).

[73] Natori, S. & Takano, H. Two-step relaxation mode analysis with multiple evolution times: application to a single [n]polycatenane. *J. Physical Soc. Japan* **86**, 043003 (2017).

[74] de Gennes, P. G. *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1984).

[75] Doi, M. & Edwards, S. F. *The Theory of Polymer Dynamics* (Oxford University Press, Oxford, 1986).

[76] Nagai, T., Mitsutake, A. & Takano, H. Relaxation mode analysis of a biopolymer system by molecular dynamics. in *Biophysics* **vol. 49 Supplement S75** (Abstract for the 47th Annual meeting, The biophysical society of Japan, 2009).

[77] Nagai, T., Mitsutake, A. & Takano, H. Principal component relaxation mode analysis of an all-atom molecular dynamics simulation of human lysozyme. *J. Physical Soc. Japan* **82**, 023803 (2013).

[78] Mitsutake, A. & Takano, H. Folding pathways of NuG2—a designed mutant of protein G—using relaxation mode analysis. *J. Chem. Phys.* **151**, 044117 (2019).

[79] Karasawa, N., Mitsutake, A. & Takano, H. Two-step relaxation mode analysis with multiple evolution times applied to all-atom molecular dynamics protein simulation. *Phys. Rev. E* **96**, 062408 (2017).

[80] Karasawa, N., Mitsutake, A. & Takano, H. Identification of slow relaxation modes in a protein trimer via positive definite

relaxation mode analysis. *J. Chem. Phys.* **150**, 084113 (2019).

[81] Natori, S. & Takano, H. Dynamic properties of densely grafted polymer brushes investigated by multistep relaxation mode analysis. *J. Physical Soc. Japan* **87**, 104003 (2018).

[82] Risken, H. *The Fokker-Planck equation: Methods of Solution and Applications.* 2nd Ed. (Springer-Verlag, Berlin, Heidelberg, 1989).

[83] Zwanzig, R. *Nonequilibrium Statistical Mechanics* (Oxford university press, New York, 2001).

[84] Mitsutake, A., Kinoshita, M., Okamoto, Y. & Hirata, F. Combination of the replica-exchange monte carlo method and the reference interaction site model theory for simulating a peptide molecule in aqueous solution. *J. Phys. Chem. B* **108**, 19002–19012 (2004).

[85] Mitsutake, A., Hansmann, U. H. E. & Okamoto, Y. Temperature dependence of distributions of conformations of a small peptide. *J. Mol. Graph. Model.* **16**, 226–238 (1998).

[86] Eckart, C. Some studies concerning rotating axes and polyatomic molecules. *Phys. Rev.* **47**, 552–558 (1935).

[87] McLachlan, A. D. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **128**, 49–79 (1979).

[88] Chandler, D. & Andersen, H. C. Optimized cluster expansions for classical fluids. II. theory of molecular liquids. *J. Chem. Phys.* **57**, 1930–1937 (1972).

[89] Ladanyi, B. M. & Chandler, D. New type of cluster theory for molecular fluids: Interaction site cluster expansion. *J. Chem. Phys.* **62**, 4308–4324 (1975).

[90] Chandler, D. Cluster diagrammatic analysis of the RISM equation. *Mol. Phys.* **31**, 1213–1223 (1976).

[91] Pratt, L. R. & Chandler, D. Interaction site cluster series for the Helmholtz free energy and variational principle for chemical equilibria and intramolecular structures. *J. Chem. Phys.* **66**, 147–151 (1977).

[92] Chandler, D. The dielectric constant and related equilibrium properties of molecular fluids: Interaction site cluster theory analysis. *J. Chem. Phys.* **67**, 1113–1124 (1977).

[93] Hirata, F. & Rossky, P. J. An extended rism equation for molecular polar fluids. *Chem. Phys. Lett.* **83**, 329–334 (1981).

[94] Hirata, F., Pettitt, B. M. & Rossky, P. J. Application of an extended RISM equation to dipolar and quadrupolar fluids. *J. Chem. Phys.* **77**, 509–520 (1982).

[95] Pettitt, B. M. & Rossky, P. J. Integral-equation predictions of liquid state structure for waterlike intermolecular potentials. *J. Chem. Phys.* **77**, 1451–1457 (1982).

[96] Pettitt, B. M. & Karplus, M. The potential of mean force surface for the alanine dipeptide in aqueous solution: a theoretical approach. *Chem. Phys. Lett.* **121**, 194–201 (1985).

[97] Kinoshita, M., Okamoto, Y. & Hirata, F. Calculation of hydration free energy for a solute with many atomic sites using the RISM theory: A robust and efficient algorithm. *J. Comput. Chem.* **18**, 1320–1326 (1997).

[98] Kinoshita, M., Okamoto, Y. & Hirata, F. Solvation structure and stability of peptides in aqueous solutions analyzed by the reference interaction site model theory. *J. Chem. Phys.* **107**, 1586–1599 (1997).

[99] Kinoshita, M., Okamoto, Y. & Hirata, F. Calculation of solvation free energy using RISM theory for peptide in salt solution. *J. Comput. Chem.* **19**, 1724–1735 (1998).

[100] Kinoshita, M., Okamoto, Y. & Hirata, F. First-principle determination of peptide conformations in solvents: Combination of Monte Carlo simulated annealing and RISM Theory. *J. Am. Chem. Soc.* **120**, 1855–1863 (1998).

[101] Kinoshita, M., Okamoto, Y. & Hirata, F. Analysis on conformational stability of C-peptide of ribonuclease A in water using the reference interaction site model theory and Monte Carlo simulated annealing. *J. Chem. Phys.* **110**, 4090–4100 (1999).

[102] Mitsutake, A., Kinoshita, M., Okamoto, Y. & Hirata, F. Multicanonical algorithm combined with the RISM theory for simulating peptides in aqueous solution. *Chem. Phys. Lett.* **329**, 295–303 (2000).

[103] Beglov, D. & Roux, B. An integral equation to describe the solvation of polar molecules in liquid water. *J. Phys. Chem. B* **101**, 7821–7826 (1997).

[104] Kovalenko, A. & Hirata, F. Three-dimensional density profiles of water in contact with a solute of arbitrary shape: A RISM approach. *Chem. Phys. Lett.* **290**, 237–244 (1998).

[105] Kovalenko, A. & Hirata, F. Potential of mean force between two molecular ions in a polar molecular solvent: A study by the three-dimensional reference interaction site model. *J. Phys. Chem. B* **103**, 7942–7957 (1999).

[106] Hansen, J. P. & McDonald, I. R. *Theory of Simple Liquids*, third ed. (Elsevier/Academic Press, London, 2006).

[107] Cortis, C. M., Rossky, P. J. & Friesner, R. A. A three-dimensional reduction of the Ornstein-Zernicke equation for molecular liquids. *J. Chem. Phys.* **107**, 6400–6414 (1997).

[108] Kovalenko, A. & Hirata, F. Hydration free energy of hydrophobic solutes studied by a reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method. *J. Chem. Phys.* **113**, 2793–2805 (2000).

[109] Kido, K., Yokogawa, D. & Sato, H. A modified repulsive bridge correction to accurate evaluation of solvation free energy in integral equation theory for molecular liquids. *J. Chem. Phys.* **137**, 024106 (2012).

[110] Kovalenko, A. & Hirata, F. Potentials of mean force of simple ions in ambient aqueous solution. I. Three-dimensional reference interaction site model approach. *J. Chem. Phys.* **112**, 10391–10402 (2000).

[111] Kast, S. M. & Kloss, T. Closed-form expressions of the chemical potential for integral equation closures with certain bridge functions. *J. Chem. Phys.* **129**, 236101 (2008).

[112] Luchko, T., Blinov, N., Limon, G. C., Joyce, K. P. & Kovalenko, A. SAMPL5: 3D-RISM partition coefficient calculations with partial molar volume corrections and solute conformational sampling. *J. Comput. Aided Mol. Des.* **30**, 1115–1127 (2016).

[113] Kobryn, A. E., Gusarov, S. & Kovalenko, A. A closure relation to molecular theory of solvation for macromolecules. *J. Phys. Condens. Matter* **28**, 404003 (2016).

[114] Kovalenko, A., Ten-no, S. & Hirata, F. Solution of three-dimensional reference interaction site model and hypernetted chain equations for simple point charge water by modified method of direct inversion in iterative subspace. *J. Comput. Chem.* **20**, 928–936 (1999).

[115] Sergiievskyi, V. P. & Fedorov, M. V. 3DRISM multigrid algorithm for fast solvation free energy calculations. *J. Chem. Theory Comput.* **8**, 2062–2070 (2012).

[116] Gusarov, S., Pujari, B. S. & Kovalenko, A. Efficient treatment of solvation shells in 3D molecular theory of solvation. *J. Comput. Chem.* **33**, 1478–1494 (2012).

[117] Maruyama, Y. & Hirata, F. Modified Anderson method for accelerating 3D-RISM calculations using graphics processing unit. *J. Chem. Theory Comput.* **8**, 3015–3021 (2012).

[118] Maruyama, Y., Yoshida, N., Tadano, H., Takahashi, D., Sato, M. & Hirata, F. Massively parallel implementation of 3D-RISM calculation with volumetric 3D-FFT. *J. Comput. Chem.* **35**, 1347–1355 (2014).

[119] Ben-Naim, A. *Molecular Theory of Solutions* (Oxford University Press, New York, 2006).

[120] Singer, S. J. & Chandler, D. Free energy functions in the extended RISM approximation. *Mol. Phys.* **55**, 621–625 (1985).

[121] Kovalenko, A. & Hirata, F. Self-consistent description of a metal-water interface by the Kohn-Sham density functional theory and the three-dimensional reference interaction site model. *J. Chem. Phys.* **110**, 10095–10112 (1999).

[122] Chandler, D., Singh, Y. & Richardson, D. M. Excess electrons in simple fluids. I. General equilibrium theory for classical hard sphere solvents. *J. Chem. Phys.* **81**, 1975–1982 (1984).

[123] Ichiye, T. & Chandler, D. Hypernetted chain closure reference interaction site method theory of structure and thermodynamics for alkanes in water. *J. Phys. Chem.* **92**, 5257–5261 (1988).

[124] Lee, P. H. & Maggiora, G. M. Solvation thermodynamics of polar molecules in aqueous solution by the XRISM method. *J. Phys. Chem.* **97**, 10175–10185 (1993).

[125] Ten-no, S. Free energy of solvation for the reference interaction site model: Critical comparison of expressions. *J. Chem. Phys.* **115**, 3724–3731 (2001).

[126] Sato, K., Chuman, H. & Ten-no, S. Comparative study on solvation free energy expressions in reference interaction site model integral equation theory. *J. Phys. Chem. B* **109**, 17290–17295 (2005).

[127] Kovalenko, A., Hirata, F. & Kinoshita, M. Hydration structure and stability of Met-enkephalin studied by a three-dimensional reference interaction site model with a repulsive bridge correction and a thermodynamic perturbation method. *J. Chem. Phys.* **113**, 9830–9836 (2000).

[128] Tanimoto, S., Yoshida, N., Yamaguchi, T., Ten-no, S. L. & Nakano, H. Effect of molecular orientational correlations on solvation free energy computed by reference interaction site model theory. *J. Chem. Info. Model.* **59**, 3770–3781 (2019).

[129] Chuev, G. N. & Fedorov, M. V. Reference interaction site model study of self-aggregating cyanine dyes. *J. Chem. Phys.* **131**, 074503 (2009).

[130] Palmer, D. S., Sergiievskyi, V. P., Jensen, F. & Fedorov, M. V. Accurate calculations of the hydration free energies of druglike molecules using the reference interaction site model. *J. Chem. Phys.* **133**, 044104 (2010).

[131] Palmer, D. S., Frolov, A. I., Ratkova, E. L. & Fedorov, M. V. Towards a universal method for calculating hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J. Phys. Condens. Matter* **22**, 492101 (2010).

[132] Palmer, D. S., Frolov, A. I., Ratkova, E. L. & Fedorov, M. V. Toward a universal model to calculate the solvation thermodynamics of druglike molecules: the importance of new experimental databases. *Mol. Pharm.* **8**, 1423–1429 (2011).

[133] Ratkova, E. L. & Fedorov, M. V. On a relationship between molecular polarizability and partial molar volume in water. *J. Chem. Phys.* **135**, 244109 (2011).

[134] Frolov, A. I., Ratkova, E. L., Palmer, D. S. & Fedorov, M. V. Hydration thermodynamics using the reference interaction site model: speed or accuracy? *J. Phys. Chem. B* **115**, 6011–6022 (2011).

[135] Truchon, J.-F., Pettitt, B. M. & Labute, P. A cavity corrected 3D-RISM functional for accurate solvation free energies. *J. Chem. Theory Comput.* **10**, 934–941 (2014).

[136] Sergiievskyi, V. P., Jeanmairet, G., Levesque, M. & Borgis, D. Fast computation of solvation free energies with molecular density functional theory: Thermodynamic-ensemble partial molar volume corrections. *J. Phys. Chem. Lett.* **5**, 1935–1942 (2014).

[137] Sergiievskyi, V. P., Jeanmairet, G., Levesque, M. & Borgis, D. Solvation free-energy pressure corrections in the three dimensional reference interaction site model. *J. Chem. Phys.* **143**, 184116 (2015).

[138] Misin, M., Fedorov, M. V. & Palmer, D. S. Communication: Accurate hydration free energies at a wide range of temperatures from 3D-RISM. *J. Chem. Phys.* **142**, 091105 (2015).

[139] Misin, M., Fedorov, M. V. & Palmer, D. S. Hydration free energies of molecular ions from theory and simulation. *J. Phys. Chem. B* **120**, 975–983 (2016).

[140] Misin, M., Palmer, D. S. & Fedorov, M. V. Predicting solvation free energies using parameter-free solvent models. *J. Phys. Chem. B* **120**, 5724–5731 (2016).

[141] Imai, T., Kinoshita, M. & Hirata, F. Theoretical study for partial molar volume of amino acids in aqueous solution: Implication of ideal fluctuation volume. *J. Chem. Phys.* **112**, 9469–9478 (2000).

[142] Imai, T., Harano, Y., Kovalenko, A. & Hirata, F. Theoretical study for volume changes associated with the helix-coil transition of peptides. *Biopolymers* **59**, 512–519 (2001).

[143] Sumi, T., Mitsutake, A. & Maruyama, Y. A solvation-free-energy functional: A reference-modified density functional formulation. *J. Comput. Chem.* **36**, 1359–1369 (2015).

[144] Sumi, T., Mitsutake, A. & Maruyama, Y. Erratum: "A solvation-free-energy functional: A reference-modified density functional formulation" [*J. Comput. Chem.* **36**, 1359–1369 (2015)]. *J. Comput. Chem.* **36**, 2009–2011 (2015).

[145] Chandler, D., Mccoy, J. D. & Singer, S. J. Density functional theory of nonuniform polyatomic systems. I. General formulation. *J. Chem. Phys.* **85**, 5971–5976 (1986).

[146] Maruyama, Y. Correction terms for the solvation free energy functional of three-dimensional reference interaction site model based on the reference-modified density functional theory. *J. Mol. Liq.* **291**, 111160 (2019).

[147] Chong, S.-H. & Ham, S. Atomic decomposition of the protein solvation free energy and its application to amyloid-beta protein in water. *J. Chem. Phys.* **135**, 034506 (2011).

[148] Chong, S.-H. & Ham, S. Component analysis of the protein hydration entropy. *Chem. Phys. Lett.* **535**, 152–156 (2012).

[149] Chong, S.-H. & Ham, S. Site-directed analysis on protein hydrophobicity. *J. Comput. Chem.* **35**, 1364–1370 (2014).

[150] Chong, S.-H. & Ham, S. Impact of chemical heterogeneity on protein self-assembly in water. *Proc. Natl. Acad. Sci. USA* **109**, 7636–7641 (2012).

[151] Chong, S.-H., Park, M. & Ham, S. Structural and thermodynamic characteristics that seed aggregation of amyloid-β protein in water. *J. Chem. Theory Comput.* **8**, 724–734 (2012).

[152] Chong, S.-H. & Ham, S. Interaction with the surrounding water plays a key role in determining the aggregation propensity of proteins. *Angew. Chem. Int. Ed. Engl.* **53**, 3961–3964 (2014).

[153] Chong, S.-H. & Ham, S. Distinct role of hydration water in protein misfolding and aggregation revealed by fluctuating thermodynamics analysis. *Acc. Chem. Res.* **48**, 956–965 (2015).

[154] Lin, Y., Im, H., Diem, L. T. & Ham, S. Characterizing the structural and thermodynamic properties of Aβ42 and Aβ40. *Biochem. Biophys. Res. Commun.* **510**, 442–448 (2019).

[155] Chong, S.-H., Hong, J., Lim, S., Cho, S., Lee, J. & Ham, S. Structural and thermodynamic characteristics of amyloidogenic intermediates of β-2-microglobulin. *Sci. Rep.* **5**, 13631 (2015).

[156] Chong, S.-H. & Ham, S. A new computational method for protein–ligand binding thermodynamics. *Bull. Korean Chem. Soc.* **40**, 180–185 (2019).

[157] Yamazaki, T. & Kovalenko, A. Spatial decomposition analysis of the thermodynamics of cyclodextrin complexation. *J. Chem. Theory Comput.* **5**, 1723–1730 (2009).

[158] Yamazaki, T. & Kovalenko, A. Spatial decomposition of solvation free energy based on the 3D integral equation theory of molecular liquid: application to miniproteins. *J. Phys. Chem. B* **115**, 310–318 (2011).

[159] Kiyota, Y. & Takeda-Shitaka, M. Molecular recognition study on the binding of calcium to calbindin D9k based on 3D reference interaction site model theory. *J. Phys. Chem. B* **118**, 11496–11503 (2014).

[160] van der Spoel, D. & Seibert, M. M. Protein folding kinetics and thermodynamics from atomistic simulations. *Phys. Rev. Lett.* **96**, 238102 (2006).

[161] Xu, W., Lai, T., Yang, Y. & Mu, Y. Reversible folding simulation by hybrid Hamiltonian replica exchange. *J. Chem. Phys.* **128**, 175105 (2008).

[162] Zacharias, M. Combining elastic network analysis and molecular dynamics simulations by Hamiltonian replica exchange. *J. Chem. Theory Comput.* **4**, 477–487 (2008).

[163] Kier, B. L. & Andersen, N. H. Probing the lower size limit for protein-like fold stability: ten-residue microproteins with specific, rigid structures in water. *J. Am. Chem. Soc.* **130**, 14675–14683 (2008).

[164] Roy, S., Goedecker, S., Field, M. J. & Penev, E. A minima hopping study of all-atom protein folding and structure prediction. *J. Phys. Chem. B* **113**, 7315–7321 (2009).

[165] Moritsugu, K., Terada, T. & Kidera, A. Scalable free energy calculation of proteins via multiscale essential sampling. *J. Chem. Phys.* **133**, 224105 (2010).

[166] Harada, R. & Kitao, A. Exploring the folding free energy landscape of a β-hairpin miniprotein, chignolin, using multiscale free energy landscape calculation method. *J. Phys. Chem. B* **115**, 8806–8812 (2011).

[167] Okumura, H. Temperature and pressure denaturation of chignolin: Folding and unfolding simulation by multibaric-multithermal molecular dynamics method. *Proteins* **80**, 2397–2416 (2012).

[168] Harada, R., Nakamura, T., Takano, Y. & Shigeta, Y. Protein folding pathways extracted by OFLOOD: Outlier FLOODing method. *J. Comput. Chem.* **36**, 97–102 (2015).

[169] Harada, R., Takano, Y. & Shigeta, Y. Enhanced conformational sampling method for proteins based on the TaBoo SeArch algorithm: application to the folding of a mini-protein, chignolin. *J. Comput. Chem.* **36**, 763–772 (2015).

[170] Nishimoto, Y. & Fedorov, D. G. The fragment molecular orbital method combined with density-functional tight-binding and the polarizable continuum model. *Phys. Chem. Chem. Phys.* **18**, 22047–22061 (2016).

[171] Satoh, D., Shimizu, K., Nakamura, S. & Terada, T. Folding free-energy landscape of a 10-residue mini-protein, chignolin. *FEBS Lett.* **580**, 3422–3426 (2006).

[172] Suenaga, A., Narumi, T., Futatsugi, N., Yanai, R., Ohno, Y., Okimoto, N., *et al.* Folding dynamics of 10-residue β-hairpin peptide chignolin. *Chem. Asian J.* **2**, 591–598 (2007).

[173] Kitao, A. Transform and relax sampling for highly anisotropic systems: application to protein domain motion and folding. *J. Chem. Phys.* **135**, 045101 (2011).

[174] Kührová, P., De Simone, A., Otyepka, M. & Best, R. B. Force-field dependence of chignolin folding and misfolding: comparison with experiment and redesign. *Biophys. J.* **102**, 1897–1906 (2012).

[175] Shao, Q. Folding or misfolding: the choice of β-hairpin. *J. Phys. Chem. B* **119**, 3893–3900 (2015).

[176] Honda, S., Akiba, T., Kato, Y. S., Sawada, Y., Sekijima, M., Ishimura, M., *et al.* Crystal structure of a ten-amino acid protein. *J. Am. Chem. Soc.* **130**, 15327–15331 (2008).

[177] Hatfield, M. P. D., Murphy, R. F. & Lovas, S. Molecular dynamics analysis of the conformations of a beta-hairpin miniprotein. *J. Phys. Chem. B* **114**, 3028–3037 (2010).

[178] Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).

[179] Maruyama, Y. & Mitsutake, A. Stability of unfolded and folded protein structures using a 3D-RISM with the RMDFT. *J. Phys. Chem. B* **121**, 9881–9885 (2017).

[180] Imai, T., Harano, Y., Kinoshita, M., Kovalenko, A. & Hirata, F. A theoretical analysis on hydration thermodynamics of proteins. *J. Chem. Phys.* **125**, 024911 (2006).

[181] Maruyama, Y. & Harano, Y. Does water drive protein folding? *Chem. Phys. Lett.* **581**, 85–90 (2013).

[182] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).