Statistical Mechanics of Protein Allostery: Roles of Backbone and Side-Chain Structural Fluctuations

Kazuhito Itoh*
Department of Applied Physics, Nagoya University, Nagoya, 464-8603, Japan,

Masaki Sasai
Department of Applied Physics, Nagoya University, Nagoya, 464-8603, Japan,
Korea Institute for Advanced Study, Seoul 130-722, Korea, and
Okazaki Institute for Integrative Bioscience, Okazaki 444-8787, Japan

A statistical mechanical model of allosteric transition of proteins is developed by extending the structure-based model of protein folding to cases that a protein has two different native conformations. Partition function is calculated exactly within the model and free-energy surfaces associated with allostery are derived. In this paper, the model of allosteric transition proposed in a previous paper (*Proc. Natl. Acad. Sci. USA*, **107**, 7775-7780 (2010)) is reformulated to describe both fluctuation in side-chain configurations and that in backbone structures in a balanced way. The model is applied to example proteins, Ras, calmodulin, and CheY: Ras undergoes the allosteric transition between GDP-bound and GTP-bound forms, and the model results show that the GDP-bound form is stabilized enough to prevent unnecessary signal transmission, but the conformation in the GTP-bound state bears large fluctuation in side-chain configurations, which may help to bind multiple target proteins for multiple pathways of signalling. The calculated results of calmodulin show the scenario of sequential ordering in Ca^{2+} binding and the associated allosteric conformational change, which are realized through the sequential appearing of pre-existing structural fluctuations, i.e., fluctuations to show structures suitable to bind Ca^{2+} before its binding. Here, the pre-existing fluctuations to accept the second and third Ca^{2+} ions are dominated by the side-chain fluctuation. In CheY, the calculated side-chain fluctuation of Tyr106 is coordinated with the backbone structural change in the $\beta 4$–$\alpha 4$ loop, which explains the pre-existing Y-T coupling process in this protein. Ability of the model to explain allosteric transitions of example proteins supports the view that the large entropic effects lower the free energy barrier of allosteric transition.

I. INTRODUCTION

Allosteric transition of protein conformation plays essential roles in many biological regulatory processes, and its physical mechanism has been intensively studied[1–6]. There remain, however, largely unresolved problems on the mechanism of allosteric transitions: Consider the case, for example, that the protein undergoes allosteric transition between two different conformations. Here, we should note that such allosteric proteins may intrinsically fluctuate among different conformations as shown in NMR measurment[2, 3, 5], so that in a more precise way of saying, we should consider the transition between one ensemble of conformations fluctuating around a certain typical structure and the other ensemble of conformations fluctuating around different typical structure. It has been proposed that in spite of such fluctuations, the mechanism of transition might be explained by the structural change along one or a few number of definite routes which connect those two typical structures [7–10]. These selected routes, if they are main routes of the transiton, have to be sufficiently stabilized in energy to have a statistical significance. In molecular systems under the influence of intense thermal noise, however, a long time duration is required to search for the energetically favorable states [11], or sometimes they might be unaccessible when large fluctuations predominate, and hence kinetic accessibility to such a few number of routes has to be carefully examined. To search accessible pathways connecting two structures, dynamical simulations of conformational fluctuation have been performed, and it has been shown that the low energy vibration modes or the principal components of fluctuation around each of two structures largely determine the transition pathway [12–15]. Even in the case that such a few number of modes are important around two structures, the nonlinear features should become more explicit at around the boundary of two basins of free energy, i.e., at the transition state of the allosteric conformational change.

Miyashita et al. [16] examined energetics of the transition state and showed that the nonlinear effects should be indeed important. They suggested that the simple extension of normal modes from each of two structures leads to the prohibitively high energy state at the transition state region, so that some entropic effects are necessary to compensate such energy rise. In this view, the large entropy at the transition state region can lower the free energy barrier, or in other words, not a few number of routes but the exponentially large number of possible routes increase the probability of the transition, which compensates the energy rise along each route. They attributed this large entropy to the partial unfolding or "cracking" of protein

* kazuhito@tbp.cse.nagoya-u.ac.jp

at around the transition state [16, 17].

Importance of entropic effects at the transition state region has been under the debate: In one view, the non-native interactions, i.e., interactions between atoms which do not reside in proximity in either of two structures become important and lower the energy of the transition state without much increase in entropy [8]. In the other view, native interactions, i.e., interactions between atoms which are close in space in either of two structures dominate the structural change in the time scale of milliseconds or longer as in folding processes, so that the loss of native interactions at the transition state should lead to the large energy rise in this time scale, which has to be compensated by increase in entropy [16]. In order to resolve this problem and to get a unified picture, one has to develop theoretical models based on these ideas to critically examine whether they are consistent with the experimental data.

In a previous paper [18], the present authors developed a statistical mechanical model of allosteric transition by extending the model of protein folding, which has been proposed by Wako and Saito [19, 20] and actively studied more recently [21–24]. This folding model neglects the non-native interactions but quantitatively explains the thermodynamic and kinetic features of folding processes of many proteins [22, 25–29]. A merit to use this model is that the partition function can be derived exactly within the model, so that free energy surfaces and other important quantities are readily calculated in a transparent way. The extension of this model for application to the problems of allosteric transitions inherits these distinctive features allowing the detailed quantitative calculation of allosteric transitions.

In Ref.18, this statistical mechanical model was applied to the allosteric transition of an example protein, the receiver domain of nitrogen regulatory protein C (NtrC). The dominant conformation of NtrC is its inactive form before phosphorylation, which turns into the active form upon phosphorylation. The NMR data showed that NtrC exhibits the large amplitude pre-existing fluctuation [8, 30], i.e., fluctuation between active and inactive conformations before phosphorylation, which implies that the allosteric transition of NtrC is dominated by the "population-shift" or "conformational selection" mechanism rather than the "induced-fit" mechanism: conformations near the active form are the pre-existing metastable ones, which are visited before phosphorylation, but are more emphasized upon phosphorylation and become to have a large statistical weight, inducing the shift in population of conformations from inactive to active forms [5, 8, 30]. The results calculated with the statistical mechanical model quantitatively explained these NMR data [18].

With the model of Ref.18, the pre-existing fluctuation of NtrC is explained by the lowering of the free energy barrier through the entropic effects; At the transition state region, the local structures of some parts of the protein resemble to active structures, while those of other parts resemble to inactive structures, and the protein can take the combinatorially large number of mosaic patterns of those coexisting local structures. This large number of structural patterns contribute to the large entropy at the transition state, which lowers the free energy barrier and brings about the large amplitude pre-existing structural fluctuation. With this appearance of mosaic patterns, some native interactions specific to active or inactive structure are lost, which induces fluctuating formation and loss of native interactions during allosteric transition. This disordering of interactions is different from "cracking" introduced in Ref.16 in its literal sense, but shares the features of fluctuating interactions [31] to induce entropic effects, whose importance was also confirmed in the atomistic molecular dynamics simulation of a short protein chain [32]. This statistical mechanical model further predicts how the free energy is modified by the local perturbation at individual residues and how the structural change spreads from one structure to the other within a protein with the atomistic resolution [18]. These predictions should provide opportunities to check the validity of the model, and hence further exploration with this model should be meaningful to resolve the debate on the mechanism of allosteric transition.

In application of this statistical mechanical method to many proteins, we should be aware of the facts that in many proteins difference of two structures before and after the allosteric transition is characterized not only by difference in the backbone structure but also by difference in side-chain configuration and packing [33]. We should, therefore, describe allosteric transitions in terms of both backbone structure and side-chain configuration. In this paper, we reformulate the statistical mechanical model of allosteric transition from that proposed in Ref.18 to treat the two aspects, backbone structure and side-chain configuration, in a more balanced and systematic way and apply this model to example proteins, Ras, calmodulin, and CheY, to compare the calculated results with experimental data. We focus on how fluctuations in the side-chain degrees of freedom contribute to the entropic effects to explain allosteric transitions of these example proteins.

In Sec.II we introduce the statistical mechanical model of the allosteric transition. The calculation methods of the model are explained in Sec.III, and reaction coordinates of free energy surfaces and order parameters of the conformational transition are explained in Sec.IV. In Sec.V we explain the entropic effects of allosteric transition. We apply the model to three proteins, Ras, calmodulin, and CheY as examples in Sec.VI, and discuss the allosteric mechanism in Sec.VII

II. MODEL

A. Definition of conformational states

We consider a protein which undergoes allosteric transition between two free-energy basins, one around 'active' (A) conformation and the other around 'inactive' (I) conformation, and these two native conformations, A and I, are used as references to describe other conformations. For A and I of example proteins which we discuss later on in this paper, X-ray crystal structures are used in the model, but we should note that protein structure largely fluctuates in solution as evidenced by NMR or shown with the shallow valleys in the calculated free energy surfaces, so that allosteric transitions we consider are not the transitions between single frozen conformations but transitions between ensembles of conformations fluctuating around those reference conformations: A and I are used as input data to Hamiltonian to describe transitions in thermally activated fluctuating systems.

When both the backbone and side-chain configurations at the kth residue in the conformation A and those in the conformation I are similar to each other, we classify the kth site into type-'common' (C) site. We consider that a residue of the type C site takes either of two kinds of configurations: when both the backbone and side-chain structures of the kth residue are close to the common structures of the A and I conformations, we write $m_k = \mathrm{C}$ and otherwise, $m_k = \mathrm{D}$. The configuration of a residue which is not type C (a type $\bar{\mathrm{C}}$ site), on the other hand, is distinguished by three letters: $m_k = \mathrm{A(I)}$, when the backbone and the side-chain structures of the kth residue in the conformation examined are close to those in the A(I) conformation, and $m_k = \mathrm{D}$ otherwise. Then, the conformation of the protein is represented by

$$\mathbf{m} = (m_1, m_2, \cdots, m_N), \quad m_k = \begin{cases} \mathrm{C, D} & k \in \mathcal{C} \\ \mathrm{A, I, D} & k \in \bar{\mathcal{C}}, \end{cases}$$

where N is the total number of residues and $\mathcal{C}$ ($\bar{\mathcal{C}}$) is a set of type C ($\bar{\mathrm{C}}$) sites.

The precise definition of the type C residue is given by dihedral angles, ϕ_k, ψ_k, and $\chi(a_k)$: The kth site with the amino-acid type a_k is a type C site, when $\Theta_k = \Theta_{\mathrm{b}k}\Theta_{\mathrm{s}k} = 1$ with

$$\Theta_{\mathrm{b}k}(\eta_{\mathrm{b}}) = \theta(\eta_{\mathrm{b}} - |\Delta\phi_k| - |\Delta\psi_k|) \tag{1}$$

$$\Theta_{\mathrm{s}k}(\eta_{\mathrm{s}}) = \theta(\eta_{\mathrm{s}} - \Delta\chi(a_k)), \tag{2}$$

where $\theta(x)$ is a step function of $\theta(x) = 1$ for $x \geq 0$ and $\theta(x) = 0$ for $x < 0$, and η_{b} and η_{s} are cut-off angles. $\Delta\phi_k$ and $\Delta\psi_{k-1}$ are differences between A and I structures in dihedral angles between the $k-1$ and kth residues of the main chain. We define $\Delta\phi_1 = \Delta\psi_N = 0$ for convenience. $\Delta\chi(a_k) = 0$ for $a_k =$ Gly, Ala, and Pro, and otherwise $\Delta\chi(a_k) = \sum_{j\in r(a_k)} |\Delta\chi_{jk}|$, where $\Delta\chi_{jk}$ is difference between A and I structures in the jth χ angle of the side chain of the kth residue, and $r(a_k)$ is a set of χ angles of the amino-acid type a_k.

In order to write down Hamiltonian in an explicit way, we introduce projection functions of configuration at the kth residue, $p_k^{\mathrm{A}}(m_k)$, $p_k^{\mathrm{I}}(m_k)$, $p_k^{\mathrm{N}}(m_k)$, and $p_k^{\mathrm{D}}(m_k)$. $p_k^{\mathrm{A}}(m_k)$ is the projection onto the A structure:

$$\begin{aligned} p_k^{\mathrm{A}}(\mathrm{C}) = 1, \quad p_k^{\mathrm{A}}(\mathrm{D}) = 0 \quad \text{for } k \in \mathcal{C}, \\ p_k^{\mathrm{A}}(\mathrm{A}) = 1, \quad p_k^{\mathrm{A}}(\mathrm{I}) = p_k^{\mathrm{A}}(\mathrm{D}) = 0 \quad \text{for } k \in \bar{\mathcal{C}}, \end{aligned}$$

and $p_k^{\mathrm{I}}(m_k)$ is the projection onto the I structure:

$$\begin{aligned} p_k^{\mathrm{I}}(\mathrm{C}) = 1, \quad p_k^{\mathrm{I}}(\mathrm{D}) = 0 \quad \text{for } k \in \mathcal{C}, \\ p_k^{\mathrm{I}}(\mathrm{I}) = 1, \quad p_k^{\mathrm{I}}(\mathrm{A}) = p_k^{\mathrm{I}}(\mathrm{D}) = 0 \quad \text{for } k \in \bar{\mathcal{C}}. \end{aligned}$$

We define $p_k^{\mathrm{N}}(m_k)$ as

$$\begin{aligned} p_k^{\mathrm{N}}(\mathrm{C}) = 1, \quad p_k^{\mathrm{N}}(\mathrm{D}) = 0 \quad \text{for } k \in \mathcal{C}, \\ p_k^{\mathrm{N}}(m_k) = p_k^{\mathrm{A}}(m_k) + p_k^{\mathrm{I}}(m_k) \quad \text{for } k \in \bar{\mathcal{C}}, \end{aligned}$$

and we write $p_k^{\mathrm{D}}(m_k) = 1 - p_k^{\mathrm{N}}(m_k)$.

With these definitions of projection operators, a segment from the ith to jth residues satisfying

$$p_{i-1}^{\mathrm{D}}(m_{i-1}) \prod_{k=i}^{j} p_k^{\mathrm{N}}(m_k) p_{j+1}^{\mathrm{D}}(m_{j+1}) = 1,$$

in which residues take A, I, or C structure, is called N-stretch (i,j) and a segment of

$$p_{i-1}^{\mathrm{N}}(m_{i-1}) \prod_{k=i}^{j} p_k^{\mathrm{D}}(m_k) p_{j+1}^{\mathrm{N}}(m_{j+1}) = 1,$$

in which all resides take structures which are different from those found in either of A or I conformation, is called D-stretch (i,j). A segment from the ith to jth residues which take the A(I) structure is referred to as A(I)-stretch (i,j) when

$$(1 - p_{i-1}^{\mathrm{A(I)}}(m_{i-1})) \prod_{k=i}^{j} p_k^{\mathrm{A(I)}}(m_k) (1 - p_{j+1}^{\mathrm{A(I)}}(m_{j+1})) = 1.$$

Here, we conveniently define $p_0^{\sigma}(m_0) = p_{N+1}^{\sigma}(m_{N+1}) = 1$ ($\sigma =$ N, D, A, and I). In this way, the protein conformation is hierarchically described by a mosaic of D- and N-stretches and a mosaic of A- and I-stretches in each N-stretch (Fig.1(a))

B. Hamiltonian

We assume that interactions in the A or I conformation are described by contacts between residues, which are referred to as native contacts. In the present structure-based model, interactions in the conformation examined are also described in terms of those native contacts and effects of non-native contacts are neglected.

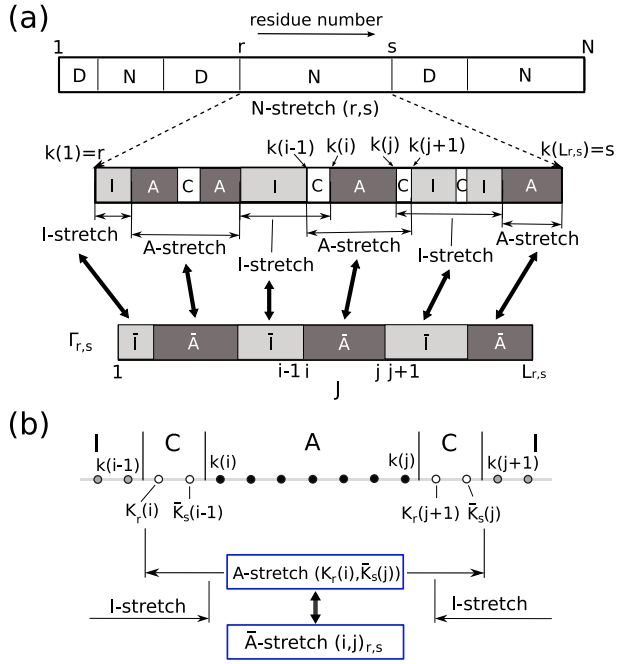

FIG. 1. The hierarchical representation of a protein conformation. (a) A conformation of a protein chain is described by a mosaic of D- and N-stretches and each N-stretch is described by a mosaic of A- and I-stretches and type C residues. By counting the contribution of type C residues at first, the reduced representation of the N-stretch (r,s) is possible in terms of $\bar{\rm A}$- and $\bar{\rm I}$-stretches on a one-dimensional lattice $\Gamma_{r,s}$, which has $L_{r,s}$ sites, where $L_{r,s}$ is the number of type C residues in the N-stretch (r,s). (b) $\bar{\rm A}(\bar{\rm I})$-stretch $(i,j)_{r,s}$ and A(I)-stretch $(K_r(i),\bar{K}_s(j))$ are in one-to-one correspondence. (See Sec.IIIA.)

Allosteric sites are specific sites of the protein, which directly interact with the effector. We use a suffix α to designate the state of allosteric sites; 'α = holo or apo', or we may consider cases such as 'α = GTP bound or GDP bound', or 'α = phosphorylated or dephosphorylated' depending on types of allosteric proteins under consideration. Hamiltonian of the state α is given by extending the statistical mechanical model of protein folding [19–24, 26–28] as

$$\mathcal{H}_\alpha(\mathbf{m}) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left[\mathcal{H}^{\rm N}_{i,j}+\mathcal{H}^{\rm A}_{i,j}+\mathcal{H}^{\rm I}_{i,j}\right]+\mathcal{V}_\alpha, \tag{3}$$

$$\mathcal{H}^{\rm N}_{i,j} = \varepsilon^{\rm N}_{i,j}\prod_{k=i}^{j} p^{\rm N}_k(m_k), \tag{4}$$

$$\mathcal{H}^{\rm A}_{i,j} = \varepsilon^{\rm A}_{i,j}\prod_{k=i}^{j} p^{\rm A}_k(m_k), \tag{5}$$

$$\mathcal{H}^{\rm I}_{i,j} = \varepsilon^{\rm I}_{i,j}\prod_{k=i}^{j} p^{\rm I}_k(m_k). \tag{6}$$

Here, $\mathcal{H}^{\rm N}_{i,j}$ is the energy gain due to contacts which are common to A and I structures with energy $\varepsilon^{\rm N}_{i,j}$ between the ith and jth residues. This energy is gained only when all the residues from i to j belong to the same N-stretch. $\mathcal{H}^{\rm A(I)}_{i,j}$ is the energy gain due to contacts which are not common to A and I but are specific to either of A or I structure with energy $\varepsilon^{\rm A(I)}_{i,j}$, which arises when all the residues from i to j belong to the same A(I)-stretch.

$\mathcal{V}_\alpha$ represents interaction energies around the allosteric sites in the state α. $\Delta\mathcal{V}_{\alpha\alpha'} = \mathcal{V}_{\alpha'} - \mathcal{V}_\alpha$ is the energy change associated with the state switching from α to α'. We may consider the case, for example, that $\alpha(\alpha')$ is the effector-free (binding) state. Then, upon the effector binding, $\Delta\mathcal{V}_{\alpha\alpha'} \neq 0$ modulates the statistical weight of stretches, which can lead to the global conformational transition of the protein from I to A. We here adopt the following form of $\mathcal{V}_\alpha$;

$$\mathcal{V}_\alpha = \sum_{i\in\mathcal{L}}\sum_{\sigma={\rm N,A,I}}\varepsilon^\sigma_{L\alpha i}p^\sigma_i(m_i) + \sum_{\substack{\langle i,j\rangle\in\mathcal{B}\\ i<j}}\sum_{\sigma={\rm N,A,I}}\varepsilon^\sigma_{\alpha i,j}\prod_{k=i}^{j}p^\sigma_k(m_k), \tag{7}$$

where $\mathcal{L}$ is a set of allosteric sites, i.e., residues which can directly interact with the effector. $\mathcal{B}$ is a set of residue pairs $\langle i,j\rangle$ whose interactions are directly modulated upon switching at the allosteric sites. We consider that members of $\mathcal{B}$ are limited to residue pairs which are spatially close to the allosteric sites. $\varepsilon^\sigma_{\alpha i,j}$ is the energy modulation of the contact energy $\varepsilon^\sigma_{i,j}$ at the α state, and the total contact energy of the pair $\langle i,j\rangle$ is $\varepsilon^\sigma_{\alpha i,j}+\varepsilon^\sigma_{i,j}$. $\varepsilon^\sigma_{L\alpha i}$ with σ = A(I) is the interaction energy between the effector and the ith residue when the ith residue is involved in the A(I)-stretch. $\varepsilon^{\rm N}_{L\alpha i}$ is energy of the interaction which is common to the A and I structures.

C. Contact energies

The contact energy between the ith and jth residues in the A(I) structure $\varepsilon^{\rm A(I)}_{{\rm tot}\,i,j} = \varepsilon^{\rm N}_{i,j} + \varepsilon^{\rm A(I)}_{i,j}$ is defined as

$$\varepsilon^{\rm A(I)}_{{\rm tot}\,i,j} = \varepsilon\sum_{\substack{k\in Res(i)\\ l\in Res(j)}}\theta(R_{\rm c}-R^{\rm A(I)}_{kl}), \tag{8}$$

with $\varepsilon\le 0$ for $j>i+2$ and $\varepsilon^{\rm A(I)}_{{\rm tot}\,i,j}=0$ for $j\le i+2$, where $Res(i)$ is a set of heavy atoms in the ith residues. $R^{\rm A(I)}_{kl}$ is a distance between the kth heavy atom in the ith residue and the lth heavy atom in the jth residue in the A(I) structure, and $R_{\rm c}$ is a threshold distance. $\varepsilon^{\rm N}_{i,j}$ and $\varepsilon^{\rm A(I)}_{i,j}$ for $j>i+2$ are given by $\varepsilon^{\rm N}_{i,j}=\varepsilon q^{\rm N}_{i,j}$ and $\varepsilon^{\rm A(I)}_{i,j}=\varepsilon q^{\rm A(I)}_{i,j}$

with

$$q_{i,j}^{\rm N} = \sum_{\substack{k\in Res(i)\\ l\in Res(j)}} \theta(R_{\rm c}-R_{kl}^{\rm A})\theta(R_{\rm c}-R_{kl}^{\rm I}), \quad (9)$$

$$q_{i,j}^{\rm A(I)} = \sum_{\substack{k\in Res(i)\\ l\in Res(j)}} \theta(R_{\rm c}-R_{kl}^{\rm A(I)})\theta(R_{kl}^{\rm I(A)}-R_{\rm c}). \quad (10)$$

The energy modulation $\varepsilon_{\alpha i,j}^{\sigma}$ of Eq.7 for $i>j$ (σ = N, A, or I) is written as

$$\varepsilon_{\alpha i,j}^{\sigma} = \gamma_{\alpha i,j}^{\sigma}\varepsilon q_{i,j}^{\sigma}, \quad \langle i,j\rangle \in \mathcal{B}, \quad (11)$$

where $\gamma_{\alpha i,j}^{\sigma}$ is a parameter of the interaction modulation. Thus, the contact energy of the modulated pair $\langle i,j\rangle \in \mathcal{B}$ is written by

$$\varepsilon_{i,j}^{\sigma} + \varepsilon_{\alpha i,j}^{\sigma} = \bar{\gamma}_{\alpha i,j}^{\sigma}\varepsilon_{i,j}^{\sigma}, \quad (12)$$

with $\bar{\gamma}_{\alpha i,j}^{\sigma} = 1+\gamma_{\alpha i,j}^{\sigma}$.

D. Partition function

The partition function of the state α is obtained by summing over all the configurations of $\mathbf{m}$ as

$$Z_{\alpha} = \sum_{\mathbf{m}} g(\mathbf{m})\exp[-\beta\mathcal{H}_{\alpha}(\mathbf{m})], \quad (13)$$

with $\beta = 1/k_{\rm B}T$, where

$$\begin{aligned} g(\mathbf{m}) &= \prod_{k=1}^{N}[\Omega_k^{\rm N}]^{p_k^{\rm N}(m_k)}[\Omega_k^{\rm D}]^{p_k^{\rm D}(m_k)} \\ &= \Omega^{\rm N}\prod_{k=1}^{N}[g_k]^{p_k^{\rm D}(m_k)} = \Omega^{\rm D}\prod_{k=1}^{N}[1/g_k]^{p_k^{\rm N}(m_k)}, \end{aligned} \quad (14)$$

with $\Omega^{\rm N(D)} = \prod_{k=1}^{N}\Omega_k^{\rm N(D)}$ and $g_k = \Omega_k^{\rm D}/\Omega_k^{\rm N}$. $\Omega_k^{\rm N}$ is the phase-space volume that the kth residue having a native configuration can take. Here, we assume that the A and I configurations have the equal phase-space volume. $\Omega_k^{\rm D}$ is the phase-space volume that the kth residue having non-native configurations can take. $k_{\rm B}\ln g_k$ is the entropic gain to have the D configurations compared to having the native configurations at the kth residue. We write g_k as

$$g_k = \begin{cases} \bar{g}_k & k\in\mathcal{C} \\ \bar{g}_k - 1 & k\in\bar{\mathcal{C}}. \end{cases} \quad (15)$$

In the present model, the conformational transition is characterized by the trade-off among energy of attractive interactions in N-stretches, entropy arising from the larger number of non-native configurations in D-stretches, and entropy arising from the combinatorial number of arrangements of A-, I-, and D-stretches.

III. CALCULATION OF PARTITION FUNCTION

We calculate the partition function hierarchically in two steps: At the first step, the statistical weight of each N-stretch is calculated by solving a two-valued problem of A and I in each N-stretch by borrowing the technique of a two-valued problem of folding [23], and at the second step, the partition function of the entire protein chain is calculated by solving the two-valued problem of N and D.

In this section, to avoid complicate expressions, we suppress the suffix α denoting the state around the allosteric sites: In the α state with σ =N, A, or I, $\varepsilon_{k,l}^{\sigma}$ expressed in this section denotes $\bar{\gamma}_{\alpha k,l}^{\sigma}\varepsilon_{k,l}^{\sigma}$ for the modulated pair $\langle k,l\rangle\in\mathcal{B}$, and $\varepsilon_{Lk}^{\sigma}$ denotes $\varepsilon_{L\alpha k}^{\sigma}$ for $k\in\mathcal{L}$. We define $\varepsilon_{l,l}^{\sigma}=0$ and $\varepsilon_{Lk}^{\sigma}=0$ for $k\notin\mathcal{L}$.

A. Hierarchical description of partition function

m_k in the N-stretch has a value either of m_k = A or I at the type $\bar{\rm C}$ site or m_k = C at the type C site. In the N-stretch (r,s), when the number of the type $\bar{\rm C}$ sites, $L_{r,s}=\sum_{k=r}^{s}(1-\Theta_k)$, is zero, $\mathcal{H}_{i,j}^{\rm A}$ and $\mathcal{H}_{i,j}^{\rm I}$ for $r\le i<j\le s$ should vanish, and thus the weight of the N-stretch (r,s), $w_{r,s}^{\rm N}$, is given by

$$w_{r,s}^{0} = \exp\left[-\beta\left(\sum_{k=r}^{s-1}\sum_{l=k+1}^{s}\varepsilon_{k,l}^{\rm N} + \sum_{k=r}^{s}\Theta_k\varepsilon_{Lk}^{\rm N}\right)\right]. \quad (16)$$

When $L_{r,s}\neq 0$, the weight of N-stretch (r,s) is given by

$$w_{r,s}^{\rm N} = w_{r,s}^{0}\sum_{\mu_{r,s}^{\rm AI}}\exp[-\beta\mathcal{H}_{r,s}^{\rm AI}]. \quad (17)$$

In Eq.17, because the contributions from the C-type sites are already included in the factor $w_{r,s}^{0}$, the residual part is calculated by introducing the effective Hamiltonian $\mathcal{H}_{r,s}^{\rm AI}$ defined on a one-dimensional lattice with $L_{r,s}$ sites (We refer to this lattice as $\Gamma_{r,s}$. See Fig.1.):

$$\mathcal{H}_{r,s}^{\rm AI} = \sum_{i=1}^{L_{r,s}}\sum_{j=i}^{L_{r,s}}\sum_{\sigma={\rm A,I}} e_{r,s}^{\sigma}(i,j)\prod_{J=i}^{j}p_J^{\sigma}(m_J^{r,s}). \quad (18)$$

Here, $m_J^{r,s} = m_{k(J)}$, $J=1,2,\cdots,L_{r,s}$ with $k(J)\in\bar{\mathcal{C}}$, where $k(J)$ is ordered as $r\le k(1)<k(2)<\cdots<k(L_{r,s})\le s$. Thus, the summation in Eq.17 is calculated over the set of $\mu_{r,s}^{\rm AI}=\{m_J^{r,s}={\rm A,I}|1\le J\le L_{r,s}\}$. We rewrite $p_{k(J)}^{\rm A(I)}$ as $p_J^{\rm A(I)}$. $e_{r,s}^{\sigma}(i,j)$ is an effective interaction energy given by

$$e_{r,s}^{\sigma}(i,j) = \sum_{l=K_r(i)}^{k(i)}\sum_{l'=k(j)}^{\bar{K}_s(j)}\varepsilon_{l,l'}^{\sigma} + \delta_{i,j}\varepsilon_{Lk(i)}^{\sigma}, \quad (19)$$

with

$$K_r(i) = \begin{cases} r & i = 1 \\ k(i-1)+1 & i \neq 1, \end{cases} \tag{20}$$

$$\bar{K}_s(j) = \begin{cases} s & j = L_{r,s} \\ k(j+1)-1 & j \neq L_{r,s}. \end{cases} \tag{21}$$

Thus, $w^{\rm N}_{r,s}/w^0_{r,s}$ can be regarded as a partition function of the effective Hamiltonian $\mathcal{H}^{\rm AI}_{r,s}$ on $\Gamma_{r,s}$. In the following, a segment from the ith to jth sites on $\Gamma_{r,s}$ is referred to as $\bar{\rm A}(\bar{\rm I})$-stretch $(i,j)_{r,s}$ when

$$p^{\rm I(A)}_{i-1}(m^{r,s}_{i-1}) \prod_{J=i}^{j} p^{\rm A(I)}_J(m^{r,s}_J) p^{\rm I(A)}_{j+1}(m^{r,s}_{j+1}) = 1.$$

$\bar{\rm A}(\bar{\rm I})$-stretch $(i,j)_{r,s}$ and A(I)-stretch $(K_r(i), \bar{K}_s(j))$ are in one-to-one correspondence as illustrated in Fig.1(b).

The weight of the N-stretch (r,s) with $L_{r,s} \neq 0$ can be written by summing up $2^{L_{r,s}}$ configurations of the $\bar{\rm A}$- and $\bar{\rm I}$-stretches $\mathcal{M}^{\rm AI}_{r,s}$ as,

$$w^{\rm N}_{r,s} = w^0_{r,s} \sum_{M \in \mathcal{M}^{\rm AI}_{r,s}} w^{\rm AI}_{r,s}(M), \tag{22}$$

where $w^{\rm AI}_{r,s}(M)$ is a product of $w^{(r.s){\rm A}}_{k,l}$ and $w^{(r.s){\rm I}}_{k,l}$ corresponding to the configuration $M \in \mathcal{M}^{\rm AI}_{r,s}$. Here, $w^{(r.s){\rm A(I)}}_{k,l}$ is the weight of the $\bar{\rm A}(\bar{\rm I})$-stretch $(i,j)_{r,s}$ given by

$$w^{(r.s)\sigma}_{k,l} = \exp\left(-\beta \sum_{i=k}^{l} \sum_{j=i}^{l} e^{\sigma}_{r,s}(i,j) \right), \tag{23}$$

where $\sigma =$ A or I, and $1 \leq k \leq l \leq L_{r,s}$. The weight of D-stretch (r,s) is written by

$$w^{\rm D}_{r,s} = \prod_{k=r}^{s} g_k. \tag{24}$$

Since expressions for $w^{\rm N}_{r,s}$ and $w^{\rm D}_{r,s}$ are obtained as in Eq.22 and Eq.24, we now can write the partition function as

$$Z = \Omega^{\rm N} \sum_{M \in \mathcal{M}_{\rm ND}} w^{\rm ND}(M), \tag{25}$$

where $\mathcal{M}_{\rm ND}$ is a set of 2^N configurations of N- and D-stretches and $w^{\rm ND}(M)$ is the statistical weight of the configuration $M \in \mathcal{M}_{\rm ND}$, which is a product of $w^{\rm N}_{r,s}$ and $w^{\rm D}_{r,s}$.

B. Generating function and constrained partition function

It is informative to calculate expectation values under the constraints of various physical conditions. For example, we consider d_1 kinds of physical quantities n_k with $k = 1, 2, ..., d_1$, which are defined in the A- or I-stretches of the protein chain (i.e., n_k is the quantity defined on the $\bar{\rm C}$-type sites or defined as a function of native interactions specific to A or I conformation) and $d_2 = d - d_1$ kinds of quantities n_k with $k = d_1+1, d_1+2, ..., d$, which are defined in the N-stretches (i.e., n_k is defined regardless of whether it is common or specific to A and I conformations). Fixing values of $\{n_k\}$ thus provides a constraint to calculate other expectation values, so that $\{n_k\}$ can be used as a d-dimensional reaction coordinate of the protein conformational change.

We consider $n_k(p,q)$ with $k = 1, 2, ..., d_1$, which are defined in the A- or I-stretch (p,q) and $n_k(r,s)$ with $k = d_1+1, d_1+2, ..., d$, defined in the N-stretch (r,s). n_k is a sum of $n_k(p,q)$ for $k \in \{1, 2, ..., d_1\}$ over A or I-stretches or a sum of $n_k(r,s)$ for $k \in \{d_1+1, d_1+2, ..., d\}$ over N-stretches along the protein chain. Then, the partition function $Z(\mathbf{n}) = Z(n_1, n_2, \cdots, n_d)$ under the constraint of the d-dimensional reaction coordinate can be derived from the generating function as explained below.

We introduce an auxiliary variable X_k for the coordinate n_k with $1 \leq k \leq d$. Defining

$$W^{(r.s)\sigma}_{i,j}(\bar{\mathbf{X}}) = w^{(r.s)\sigma}_{i,j} \prod_{k=1}^{d_1} X_k^{n_k(K_r(i), \bar{K}_s(j))}, \tag{26}$$

with $\sigma =$ A or I in the N-stretch (r,s), and replacing $w^{(r.s)\sigma}_{i,j}$ in Eq.22 with $W^{(r.s)\sigma}_{i,j}(\bar{\mathbf{X}})$, the function of $\bar{\mathbf{X}} = (X_1, X_2, \cdots, X_{d_1})$, $w^{\rm N}_{r,s}(\bar{\mathbf{X}})$, is obtained. We further define $W^{\rm N}_{r,s}(\mathbf{X})$ as

$$W^{\rm N}_{r,s}(\mathbf{X}) = w^{\rm N}_{r,s}(\bar{\mathbf{X}}) \prod_{k=d_1+1}^{d} X_k^{n_k(r,s)}. \tag{27}$$

Replacing $w^{\rm N}_{r,s}$ in Eq.25 with $W^{\rm N}_{r,s}(\mathbf{X})$, we obtain the generating function $\mathcal{G}(\mathbf{X})$ as

$$\mathcal{G}(\mathbf{X}) = \sum_{\mathbf{n}} Z(\mathbf{n}) \prod_{k=1}^{d} X_k^{n_k}, \tag{28}$$

Using the generating function, $Z(\mathbf{n})$ is obtained from

$$Z(\mathbf{n}) = \prod_{k=1}^{d} \frac{1}{n_k!} \frac{\partial^{n_k}}{\partial X_k^{n_k}} \mathcal{G}(\mathbf{X}) \Big|_{\mathbf{X}=\mathbf{0}}. \tag{29}$$

The free energy function in the d-dimensional reaction coordinate is written as

$$F(\mathbf{n}) = -k_{\rm B}T \ln Z(\mathbf{n}). \tag{30}$$

C. Transfer matrix method

As shown in previous subsections, the partition function or generating functions are hierarchically described

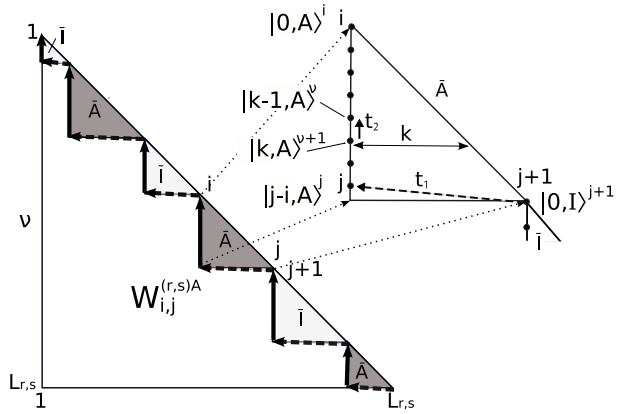

FIG. 2. A transfer pathway generated by transfer matrices on the lattice $\Gamma_{r,s}$. The arrow t_1 indicates a transfer from $|0,\mathrm{I}\rangle^{j+1}$ to $|j-i,\mathrm{A}\rangle^{j}$ by the transfer matrix $\mathcal{R}^{j}_{j+1}$, and the arrow t_2 indicates a transfer from $|k,\mathrm{A}\rangle^{\nu+1}$ to $|k-1,\mathrm{A}\rangle^{\nu}$ by $\mathcal{T}^{\nu}_{\nu+1}$ of $\mathcal{R}^{\nu}_{\nu+1}$. The transfer path from $|0,\mathrm{I}\rangle^{j+1}$ to $|0,\mathrm{A}\rangle^{i}$ creates the weight function of $\bar{\mathrm{A}}$-stretch $(i,j)_{r,s}$, $W^{(r.s)\mathrm{A}}_{i,j}(\bar{\mathbf{X}})$.

by using the weight of D-, N-, $\bar{\mathrm{A}}$-, and $\bar{\mathrm{I}}$-stretches. Utilizing this hierarchy, the exact generating function of the model can be obtained by extending Bruscolini-Pelizzola's transfer-matrix method [23] and by using the method twice, first to calculate Eq.22 and second to calculate Eq.25.

The first step is to calculate the function $w^{\mathrm{N}}_{r,s}(\bar{\mathbf{X}})$. For that purpose, we consider a triangle of size $L_{r,s}$ with respect to the lattice $\Gamma_{r,s}$. See Fig.2. A row in this triangle is designated by ν with $1 \le \nu \le L_{r,s}$. The transfer matrix $\mathcal{R}^{\nu-1}_{\nu}$ which transforms the row ν to the row $\nu-1$ is defined by its action on a vector $|k,\sigma\rangle^{\nu}$ ($k = 0,1,\cdots,\nu-1$, $\sigma = \mathrm{A},\mathrm{I}$). $|k,\mathrm{A}\rangle^{\nu}$ and $|k,\mathrm{I}\rangle^{\nu}$ represent $\bar{\mathrm{A}}$-stretch and $\bar{\mathrm{I}}$-stretch extending from the $(\nu-k)$th to νth sites on $\Gamma_{r,s}$. $|k,\sigma\rangle^{\nu}$ is regarded as a basis vector of the 2ν-dimensional space and ${}^{\nu}\langle k,\sigma| = (|k,\sigma\rangle^{\nu})^{t}$:

$$ {}^{\nu}\langle k,\sigma|k',\sigma'\rangle^{\nu} = \delta_{k,k'}\delta_{\sigma,\sigma'}. $$

The transfer matrix $\mathcal{R}^{\nu}_{\nu+1}$ is given by

$$ \begin{aligned} \mathcal{R}^{\nu}_{\nu+1}(\bar{\mathbf{X}}) = &\sum_{\sigma=\mathrm{A},\mathrm{I}}\sum_{k=1}^{\nu} |k-1,\sigma\rangle^{\nu}\,{}^{\nu+1}\langle k,\sigma| \\ &+\sum_{k=0}^{\nu-1} W^{(r.s)\mathrm{A}}_{\nu-k,\nu}(\bar{\mathbf{X}})|k,\mathrm{A}\rangle^{\nu}\,{}^{\nu+1}\langle 0,\mathrm{I}| \\ &+\sum_{k=0}^{\nu-1} W^{(r.s)\mathrm{I}}_{\nu-k,\nu}(\bar{\mathbf{X}})|k,\mathrm{I}\rangle^{\nu}\,{}^{\nu+1}\langle 0,\mathrm{A}|. \end{aligned} \tag{31} $$

$|k,\sigma\rangle^{\nu}\,{}^{\nu+1}\langle k',\sigma'| \equiv |k,\sigma\rangle^{\nu} \otimes {}^{\nu+1}\langle k',\sigma'|$ is a $2\nu \times 2(\nu+1)$ matrix and transfers the state $|k',\sigma'\rangle^{\nu+1}$ to the state $|k,\sigma\rangle^{\nu}$.

Using transfer matrices, the function $w^{\mathrm{N}}_{r,s}(\bar{\mathbf{X}})$ is given by

$$ w^{\mathrm{N}}_{r,s}(\bar{\mathbf{X}}) = w^{0}_{r,s}\,{}^{1}\langle 0|\mathcal{R}^{1}_{2}\mathcal{R}^{2}_{3}\cdots\mathcal{R}^{L_{r,s}}_{L_{r,s}+1}|0\rangle^{L_{r,s}+1}. \tag{32} $$

with $|0\rangle^{\nu} = \sum_{\sigma=\mathrm{A},\mathrm{I}}|0,\sigma\rangle^{\nu}$, where we introduced an auxiliary row of $\nu = L_{r,s}+1$. Note that

$$ \begin{aligned} {}^{i}\langle 0,\sigma|\mathcal{T}^{i}_{i+1}\mathcal{T}^{i+1}_{i+2}\cdots\mathcal{T}^{j-1}_{j}\mathcal{R}^{j}_{j+1}|0,\sigma'\rangle^{j+1} \\ = W^{(r.s)\sigma}_{i,j}(\bar{\mathbf{X}})(1-\delta_{\sigma,\sigma'}), \end{aligned} \tag{33} $$

where $\mathcal{T}^{\nu}_{\nu+1}$ is the first line of the right-hand side of Eq.31 and $\mathcal{T}^{\nu-1}_{\nu} = \mathcal{R}^{\nu-1}_{\nu}\mathcal{U}_{\nu}$ with

$$ \mathcal{U}_{\nu} = \sum_{\sigma=\mathrm{A},\mathrm{I}}\sum_{k=1}^{\nu-1} |k,\sigma\rangle^{\nu}\,{}^{\nu}\langle k,\sigma|. \tag{34} $$

By inserting the ν-dimensional identity matrix $\mathbf{1}_{\nu} = \mathcal{U}_{\nu} + |0,\mathrm{A}\rangle^{\nu\,\nu}\langle 0,\mathrm{A}| + |0,\mathrm{I}\rangle^{\nu\,\nu}\langle 0,\mathrm{I}|$ between $\mathcal{R}^{\nu-1}_{\nu}$ and $\mathcal{R}^{\nu}_{\nu+1}$ in Eq.32 for each row ν ($2 \le \nu \le L_{r,s}$), one can verify the right hand side of Eq.32 to be $w^{0}_{r,s}\sum_{M\in\mathcal{M}^{\mathrm{AI}}_{r,s}} W^{\mathrm{AI}}_{r,s}(M)$ with $W^{\mathrm{AI}}_{r,s}(M)$ being the product of $W^{(r.s)\mathrm{A}}_{i,j}(\bar{\mathbf{X}})$ and $W^{(r.s)\mathrm{I}}_{i,j}(\bar{\mathbf{X}})$ corresponding to the configuration M as in Eq.22.

In a similar way, the generating function is obtained by using Bruscolini-Pelizzola's transfer matrices [23] as

$$ \mathcal{G}(\mathbf{X}) = \Omega^{\mathrm{D}}\,{}^{0}\langle 0|\mathcal{Q}^{0}_{1}\mathcal{Q}^{1}_{2}\cdots\mathcal{Q}^{N}_{N+1}|0\rangle^{N+1}, \tag{35} $$

where $\Omega^{\mathrm{D}} = \Omega^{\mathrm{N}}w^{\mathrm{D}}_{1,N}$. Here, $|k\rangle^{\mu}$ is a $\mu+1$-dimensional basis vector, ${}^{\mu}\langle k|k'\rangle^{\mu} = \delta_{k,k'}$ ($0 \le k \le \mu$). $\mathcal{Q}^{\mu}_{\mu+1}$ is the matrix to transfer the row $\mu+1$ to the row μ ($0 \le \mu \le N+1$):

$$ \begin{aligned} \mathcal{Q}^{\mu}_{\mu+1}(\mathbf{X}) = &\sum_{k=0}^{\mu} |k\rangle^{\mu}\,{}^{\mu+1}\langle k+1| \\ &+\sum_{k=0}^{\mu} W_{\mu-k+1,\mu}(\mathbf{X})|k\rangle^{\mu}\,{}^{\mu+1}\langle 0|, \end{aligned} \tag{36} $$

with $W_{r,s}(\mathbf{X}) = W^{\mathrm{N}}_{r,s}(\mathbf{X})/w^{\mathrm{D}}_{r,s}$, where we defined $W_{r+1,r}(\mathbf{X}) = 1$ and introduced the two auxiliary rows of $\mu = 0$ and $N+1$.

By rewriting Eqs.32 and 35 in recurrent equations respectively, we calculate the generating function and obtain the constrained partition function. The detail of the recurrent equations is explained in Appendix A.

IV. FREE ENERGY SURFACE AND CONFORMATIONAL TRANSITION

In this article we illustrate free energy surfaces using four reaction coordinates n_f, n_a, n_b, and n_s. n_f is the number of residues that take the native (A, I, or C) configuration, $n_f = \sum_{k=1}^{N} p^{\mathrm{N}}_{k}(m_k)$, which is a reaction coordinate of the folding transition. At $n_f = N$ all residues take the A, I, or their common structures in an N-stretch, and at $n_f = 0$ all residues take non-native structures. n_a is the number of residues that take the structure specific to the A conformation, $n_a = \sum_{k=1}^{N}(1-\Theta_k)p^{\mathrm{A}}_{k}(m_k)$,

TABLE I. Definition and the maximum values of the reaction coordinates used in this article.

coordinate	definition	maximum value
n_f	$\sum_k p_k^{\rm N}(m_k)$	N
n_a	$\sum_k(1-\Theta_k)p_k^{\rm A}(m_k)$	$N_a=\sum_k(1-\Theta_k)$
n_b	$\sum_k(1-\Theta_{{\rm b}k})p_k^{\rm A}(m_k)$	$N_b=\sum_k(1-\Theta_{{\rm b}k})$
n_s	$\sum_k(1-\Theta_{{\rm s}k})p_k^{\rm A}(m_k)$	$N_s=\sum_k(1-\Theta_{{\rm s}k})$

which is a reaction coordinate of the allosteric transition. $n_{b(s)}$ is the number of residues that take the backbone (side-chain) configuration specific to the A conformation, $n_{b(s)}=\sum_{k=1}^N(1-\Theta_{{\rm b(s)}k})p_k^{\rm A}(m_k)$, which is a reaction coordinate of the backbone(side-chain) changes in the allosteric transition. $n_a=n_b=n_s=0$ and $n_f=N$ for the I conformation and $n_a=N_a$, $n_b=N_b$, $n_s=N_s$, and $n_f=N$ for the A conformation, where $N_a=\sum_{k=1}^N(1-\Theta_k)=L_{1,N}$ is the number of type $\bar{\rm C}$ residues and $N_{b(s)}=\sum_{k=1}^N(1-\Theta_{{\rm b(s)}k})$ is the number of residues whose backbone (side-chain) configuration of the native structure is not common to A and I. The reaction coordinates are summarized in Table I.

Free energy change due to switching from the state α to the state α' at $\mathbf{n}$ is given by

$$\begin{aligned}\Delta F_{\alpha\alpha'}(\mathbf{n}) &= F_{\alpha'}(\mathbf{n})-F_\alpha(\mathbf{n})\\ &= -k_{\rm B}T\ln\langle\exp[-\beta\Delta\mathcal{V}_{\alpha\alpha'}]\rangle_{\mathbf{n}}^{\alpha},\end{aligned}\tag{37}$$

where $\langle\cdots\rangle_{\mathbf{n}}^{\alpha}$ is the average taken with $\mathcal{H}_\alpha$ under the constraint of $\mathbf{n}$ and $\Delta\mathcal{V}_{\alpha\alpha'}=\mathcal{V}_{\alpha'}-\mathcal{V}_\alpha$.

We define two order parameters, $\rho_{\alpha i}^\sigma(\mathbf{n})$ and $\xi_{\alpha i}^\sigma(\mathbf{n})$, of how well the $\sigma=$ N, A, or I structure is developed at the ith site with a given $\mathbf{n}$ as

$$\rho_{\alpha i}^\sigma(\mathbf{n})=\langle p_i^\sigma(m_i)\rangle_{\mathbf{n}}^\alpha,\quad i\in\bar{\mathcal{C}},\tag{38}$$

and

$$\xi_{\alpha i}^\sigma(\mathbf{n})=\sum_j q_{i,j}^\sigma\langle p_{ij}^\sigma\rangle_{\mathbf{n}}^\alpha/\sum_j q_{i,j}^\sigma,\tag{39}$$

where $p_{ij}^\sigma=\prod_{k=i}^j p_k^\sigma(m_k)$. $\langle p_{ij}^\sigma\rangle_{\mathbf{n}}^\alpha$ is the probability that the fragment extending from the ith to jth sites are in the σ stretch under the constraint of $\mathbf{n}$ in the state α. Note that $\xi_{\alpha i}^\sigma(\mathbf{n})$ is defined only at the ith residue which satisfies $\sum_j q_{i,j}^\sigma\neq 0$. $\rho_{\alpha i}^\sigma$ represents the statistical average of the backbone and side-chain structure formation of the ith residue. $\xi_{\alpha i}^\sigma(\mathbf{n})$ represents the statistical average of the contact formation of the ith residue, and satisfies $0\le\xi_{\alpha i}^\sigma(\mathbf{n})\le 1$. When $\xi_{\alpha i}^{\rm A(I)}(\mathbf{n})=1$, all neighbors of the ith residue are in the same A(I) stretch as the ith residue.

Change of $\langle p_{ij}^\sigma\rangle_{\mathbf{n}}^\alpha$ due to switching from the state α to the state α' at $\mathbf{n}$ is given by a correlation between p_{ij}^σ and $\exp[-\beta(\mathcal{V}_{\alpha'}-\mathcal{V}_\alpha)]$:

$$\begin{aligned}\Delta p_{ij}^\sigma(\alpha,\alpha';\mathbf{n}) &= \langle p_{ij}^\sigma\rangle_{\mathbf{n}}^{\alpha'}-\langle p_{ij}^\sigma\rangle_{\mathbf{n}}^\alpha\\ &= \frac{\langle p_{ij}^\sigma e^{-\beta\Delta\mathcal{V}}\rangle_{\mathbf{n}}^\alpha-\langle p_{ij}^\sigma\rangle_{\mathbf{n}}^\alpha\langle e^{-\beta\Delta\mathcal{V}}\rangle_{\mathbf{n}}^\alpha}{\langle e^{-\beta\Delta\mathcal{V}}\rangle_{\mathbf{n}}^\alpha},\end{aligned}\tag{40}$$

where $\Delta\mathcal{V}=\mathcal{V}_{\alpha'}-\mathcal{V}_\alpha$.

V. ENTROPIC MECHANISM OF ALLOSTERIC TRANSITION

A. Free energy decomposition

The free energy is decomposed into energy and entropy. The summation over $\mathbf{m}$ can be described by

$$\sum_{\mathbf{m}}=\sum_{\mathbf{n}}\sum_{\mathbf{m}\in\Lambda(\mathbf{n})},$$

where $\Lambda(\mathbf{n})$ is a set of $\mathbf{m}$ constrained at $\mathbf{n}$. The partition function fixed at $\mathbf{n}$, $Z_\alpha(\mathbf{n})$, is written as

$$Z_\alpha(\mathbf{n})=\sum_{\mathbf{m}\in\Lambda(\mathbf{n})}g(\mathbf{m})\exp[-\beta\mathcal{H}_\alpha(\mathbf{m})].\tag{41}$$

$Z_\alpha(\mathbf{n})$ can be rewritten as

$$Z_\alpha(\mathbf{n})=\Omega(\mathbf{n})\left[\langle\exp(\beta\mathcal{H}_\alpha)\rangle_{\mathbf{n}}^\alpha\right]^{-1},\tag{42}$$

where $\Omega(\mathbf{n})=\sum_{\mathbf{m}\in\Lambda(\mathbf{n})}g(\mathbf{m})$ is the number of conformations constrained at $\mathbf{n}$, and

$$\langle\cdots\rangle_{\mathbf{n}}^\alpha=\frac{1}{Z_\alpha(\mathbf{n})}\sum_{\mathbf{m}\in\Lambda(\mathbf{n})}(\cdots)g(\mathbf{m})\exp[-\beta\mathcal{H}_\alpha].\tag{43}$$

Thus, the free energy at $\mathbf{n}$ can be expressed by $F_\alpha(\mathbf{n})=E_\alpha(\mathbf{n})-TS_\alpha(\mathbf{n})$ with the energy $E_\alpha(\mathbf{n})=\langle\mathcal{H}_\alpha\rangle_{\mathbf{n}}^\alpha$ and the entropy

$$S_\alpha(\mathbf{n})=S_{\rm c}(\mathbf{n})+S_{{\rm e}\alpha}(\mathbf{n}),\tag{44}$$

where $S_{\rm c}(\mathbf{n})=k_{\rm B}\ln\Omega(\mathbf{n})$ and

$$\begin{aligned}S_{{\rm e}\alpha}(\mathbf{n}) &= -k_{\rm B}\ln\langle\exp[\beta(\mathcal{H}_\alpha-E_\alpha(\mathbf{n}))]\rangle_{\mathbf{n}}^\alpha\\ &= -k_{\rm B}\sum_{k=2}^\infty\frac{\beta^k}{k!}c_k(\mathcal{H}_\alpha;\mathbf{n})\le 0.\end{aligned}\tag{45}$$

Here, $c_k(\mathcal{H}_\alpha;\mathbf{n})$ is the kth order cumulant of $\mathcal{H}_\alpha$ under the constraint of $\mathbf{n}$, e.g., $c_2(\mathcal{H}_\alpha;\mathbf{n})=\langle\mathcal{H}_\alpha^2\rangle_{\mathbf{n}}^\alpha-(\langle\mathcal{H}_\alpha\rangle_{\mathbf{n}}^\alpha)^2$ and $c_3(\mathcal{H}_\alpha;\mathbf{n})=\langle\mathcal{H}_\alpha^3\rangle_{\mathbf{n}}^\alpha-3\langle\mathcal{H}_\alpha^2\rangle_{\mathbf{n}}^\alpha\langle\mathcal{H}_\alpha\rangle_{\mathbf{n}}^\alpha+2(\langle\mathcal{H}_\alpha\rangle_{\mathbf{n}}^\alpha)^3$. $S_{{\rm e}\alpha}(\mathbf{n})$ is negative and represents the entropic cost due to the effective reduction of the number of available contact patterns [18]. Energy dependence of $S_{{\rm e}\alpha}(\mathbf{n})$ reflects the cooperativity of forming nonlocal correlation of interactions in $\mathcal{H}_\alpha$. Reduction of the number of possible backbone and side-chain configurations due to the constraint of fixed $\mathbf{n}$ is represented by this term.

B. Entropic mechanism

The conformational entropy $S_{\rm c}(\mathbf{n})$ is further decomposed as

$$S_{\rm c}(\mathbf{n})=S_{\rm cf}(\mathbf{n})+S_{\rm d}(\mathbf{n})+k_{\rm B}\ln\Omega^{\rm N}.\tag{46}$$

$S_{\rm cf}(\mathbf{n}) = k_{\rm B}\ln[\sum_{\mathbf{m}\in\Lambda(\mathbf{n})} 1]$ is entropy of a combinatorial number of configurations of N and D stretches and A and I stretches in the each N stretch constrained at $\mathbf{n}$, and $S_{\rm d}(\mathbf{n})$ is the entropic gain of non-native conformations in the D-stretches.

In the cracking mechanism proposed by Miyashita *et al.* [16], a transient local unfolding of non-native backbone configurations which are different from either A or I structure lowers the free energy barrier of the allosteric transition. With this mechanism, the free energy lowering is due to the entropic gain $S_{\rm d}(\mathbf{n})$. In the reaction coordinate $\mathbf{n} = (n_a, n_f)$, $(N-n_f)k_{\rm B}\ln g_{\rm min} \le S_{\rm d}(\mathbf{n}) \le (N-n_f)k_{\rm B}\ln g_{\rm max}$ with $g_{\rm min(max)}$ being the minimum(maximum) value of g_k, and $N-n_f$ is a number of the unfolding residues (sites with $m_k = {\rm D}$). If the local unfolding occurs at some specific sites, however, the cracking mechanism might be accompanied with the large entropic cost of $S_{{\rm e}\alpha}(\mathbf{n})$ due to the bias for selecting those specific sites.

Other possible entropic mechanism of the allosteric transition is the free energy lowering due to the entropic gain of $S_{\rm cf}(\mathbf{n})$. This gain is obtained even in the case of $n_f \approx N$: At $(n_a, n_f) = (n_a, N)$, for example, $S_{\rm d}(n_a, N) = 0$ and $S_{\rm cf}(n_a, N)$ is the entropy arising from a large number of possible configurations of A and I stretches given by

$$S_{\rm cf}(n_a, N) = k_{\rm B}\ln\binom{N_a}{n_a}. \tag{47}$$

This large entropy comes from a combinatorially large number of mosaic configurations of A and I stretches along the protein chain. In other words, there are combinatorially many possible routes from I to A conformations without invoking D-stretches through the local unfolding. Though $S_{\rm cf}$ is compensated by the negative entropy of $S_{{\rm e}\alpha}$, i.e., the conformations of the large energy with many disrupted native contacts are excluded from the count, the remaining entropy can be still large in Eq.44 to lower the free energy barrier and induces the large pre-existing conformational fluctuation. This mechanism of entropic lowering of free energy is intrinsically nonlinear associated with the backbone and side-chain structural fluctuations ranging over the entire protein and with the fluctuating formation and disruption of native interactions as have been observed in molecular dynamics simulations [31, 32].

VI. EXAMPLE PROTEINS

A. Model parameters

In order to apply the model to example proteins, we first explain the parameters used in this application. The number of residues taking the backbone configuration which is common to A and I structures, $\sum_k \Theta_{{\rm b}k}(\eta_{\rm b})$, depends on the cut-off angle, $\eta_{\rm b}$ in Eq.1, and the number of residues taking the side-chain configuration common to A and I structures, $\sum_k \Theta_{{\rm s}k}(\eta_{\rm s})$, depends on the cut-off angle, $\eta_{\rm s}$ in Eq.2. Dependences of $\sum_k \Theta_{{\rm b}k}(\eta_{\rm b})/N$ and $\sum_k \Theta_{{\rm s}k}(\eta_{\rm s})/N$ on cut-off angles are shown in Fig.3. We can see that $\sum_k \Theta_{{\rm b}k}(\eta_{\rm b})/N$ and $\sum_k \Theta_{{\rm s}k}(\eta_{\rm s})/N$ depend only mildly on $\eta_{\rm b}$ and $\eta_{\rm s}$ when $\eta_{\rm b}$ or $\eta_{\rm s}$ is larger than about 45°. The backbone configuration is typically classified into three rotamer states, so that it is reasonable to regard a residue as type-C when its difference in dihedral angles is less than 60° between A and I configurations. Thus, we set the cut-off angles to be $\eta_{\rm b} = \eta_{\rm s} = 60°$. We here note that this fairly large value for $\eta_{\rm b}$ and $\eta_{\rm s}$ implies that the vibrational backbone and side-chain fluctuations of nanoseconds or faster are coarse-grained in this representation, so that a variety of structures are collectively treated here as a "single configuration". Through this coarse-graining, slow fluctuations of micro to milliseconds are highlighted with the present model.

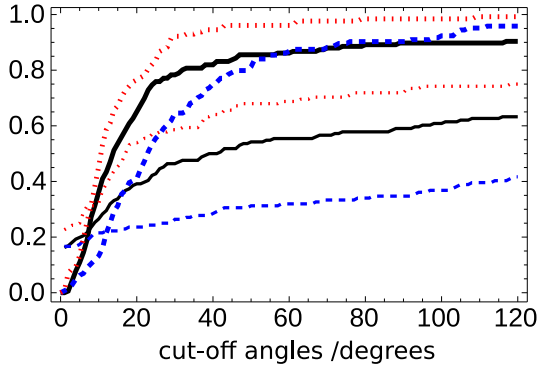

FIG. 3. The ratio of the number of residues with $\Theta_{{\rm b(s)}k} = 1$ to the total number of residues, $\sum_k \Theta_{{\rm b}k}(\eta_{\rm b})/N$ (thick lines) and $\sum_k \Theta_{{\rm s}k}(\eta_{\rm s})/N$ (thin lines), are shown for the example proteins; Ras (solid black lines), calmodulin (dashed blue lines), and CheY (dotted red lines).

To define the native contacts, we assume the threshold distance to be $R_{\rm c} = 5$Å. For simplicity, the parameter of the contact energy ε is set to be common to all contact pairs. Then, the strength of contact between a pair of residues is proportional to the number of pairs of heavy atoms which are in contact between two residues. The partition function is a function of a variable $\varepsilon/k_{\rm B}T$, which is a dimension-less parameter to control the stability of native states.

Residues with the bulky side-chain should bear greater steric hindrance, and residues occupying the larger phase-space volume $\Omega_k^{\rm D}$ in the non-native state should tend to have the larger phase-space volume $\Omega_k^{\rm N}$ in the native state. We therefore anticipate that the residue dependence of $g_k = \Omega_k^{\rm D}/\Omega_k^{\rm N}$ is weak. In this paper, by reference to values of the entropic cost to form the native structure in the single-native-structure model of the protein folding [22, 23, 29], we choose the values of $\bar{g}_k$ to be $\bar{g}_k = 2.0$ for glycine, alanine, and proline, and $\bar{g}_k = 4.5$ otherwise. Using these values of $\bar{g}_k$ and the contact energies defined in Sec.II, we can confirm that the folding/unfolding free en-

ergy surfaces of example proteins are properly described with the present model.

B. Allosteric transitions of example proteins

We apply the proposed statistical mechanical model to allosteric transitions of three example proteins, Ras, calmoduline, and CheY. The data of the example proteins are summarized in Table II. Definitions of allosteric sites $\mathcal{L}$ and the modulated pairs $\mathcal{B}$, and the parameter values of interactions in $\mathcal{L}$ and $\mathcal{B}$ such as $\bar{\gamma}^{\sigma}_{\alpha i,j}$ and $\varepsilon^{\sigma}_{L\alpha i}$ depend on specific features of each protein, which are briefly explained below together with the simulated results, but their details are given in Appendix B.

Reference conformations A and I of example proteins are depicted in Fig.4, most of which are X-ray crystal structures. Here, to avoid possible misinterpretation, we should note again that the allosteric transition which we focus on is not the transitions between individual conformations A and I but is the transition between the free-energy basin around A and that around I. Conformations A and I are used as references to describe other conformations and as the input data to Hamiltonian to calculate fluctuation around them. When the fluctuation is anisotropic in conformational space, the averaged structure over those fluctuating structures in equilibrium could differ from A or I, which may explain differences between X-ray structures and NMR or single-molecular data in solution as will be discussed in the following examples.

1. *Ras*

Ras GTPase is a conformational switch controlling cell proliferation, differentiation, and development [34]. The inactive conformation (I) is a GDP-bound form and the active conformation (A) is a GTP-bound form [35, 36] (Fig.4(a)). In the GTP-bound form, Ras can interact with various target proteins, triggering respective downstream signaling processes. Two regions, switch I (residues 25-40) and Switch II (residues 57-75), at which oncogenic mutations frequently occur, undergo major conformational changes upon GDP/GTP exchange. Observations of binding of Ras to multiple target proteins raised intriguing questions on whether Ras in the GTP-bound state can take multiple structures to fit different proteins, and such questions promoted the experimental examinations on the structural fluctuation in the GTP-bound state [37–39]. In this subsection, we show that the present statistical mechanical model explains the allosteric transition of Ras in a way consistent with the experimental data, and discuss the problems of structural fluctuation in the GTP-bound state with emphasis on the roles of fluctuations in side-chain configuration.

The allosteric sites $\mathcal{L}$ of Ras are residues which directly interact with GDP or GTP. Because definition of $\mathcal{L}$ depends on whether residues in $\mathcal{L}$ bind GTP (in the following, $\alpha = t$ designates the GTP-bound state) or GDP ($\alpha = d$ designates the GDP-bound state), we use the notation of $\mathcal{L}_\alpha$. Interactions around $\mathcal{L}_\alpha$ should also depend on α, so that we write modulated pairs around $\mathcal{L}_\alpha$ as $\mathcal{B}_\alpha$. $\mathcal{B}_{d(t)} = \{\langle i,j\rangle\}$ is a set of contact pairs in which at least either of i or j belongs to $\mathcal{L}_{d(t)}$. See Appendix B for more details.

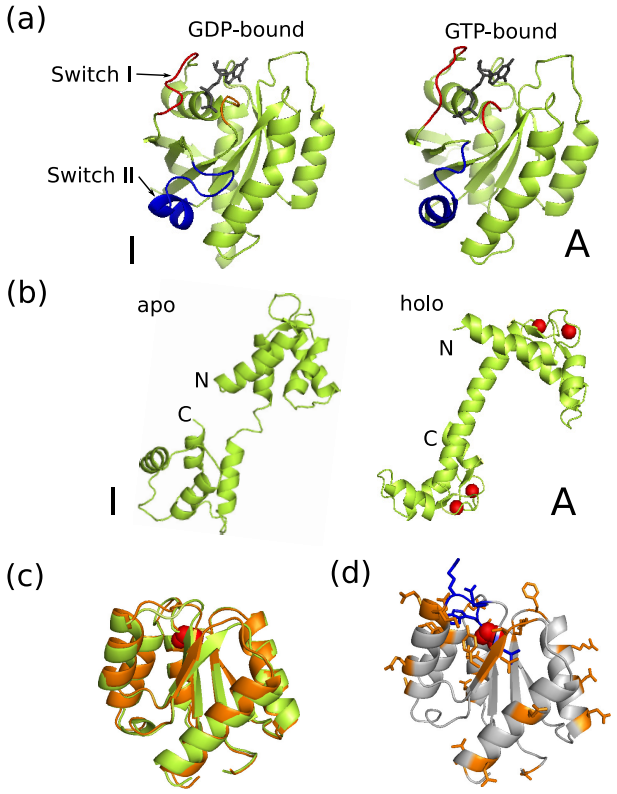

FIG. 4. The inactive (I) and active (A) conformations of example proteins. (a) Ras in the GDP- and GTP-bound forms, (b) calmodulin apo and holo (Ca^{2+}-binding) forms, and (c) a superposition of dephosphorylated (I:green) and phosphorylated (A:orange) CheY conformations are illustrated. Calmodulin apo structure is a solution structure obtained by NMR measurment, and all other structures in this figure are crystal structures obtained by X-ray diffraction analyses. See Table II. In (b) red spheres represent Ca^{2+} ions. (d) The residues taking the conformations which are not common to the dephosphorylated and phosphorylated CheY are illustrated; blue colored represent the residues which change their backbone configurations ($\Theta_{bk} = 0$) and orange colored represent the residues which largely change their side-chain configurations with only minor changes in backbone configurations ($\Theta_{bk} = 1$ and $\Theta_{sk} = 0$). In (c) and (d) a red colored residue is the phosphorylation site.

In Fig.5 the calculated free energy surfaces $F_\alpha(n_a, n_f)$ in GDP- and GTP-bound states are shown in the two-dimensional space of n_f and n_a. The transition involving I and A conformations occurs in the native basin of $n_f > 160$, and the native state is separated from the unfolded state at around $n_f \approx 60$ by a large free energy

TABLE II. Example proteins. $N_a = \sum_{k=1}^{N}(1-\Theta_k)$, $N_b = \sum_{k=1}^{N}(1-\Theta_{bk})$, and $N_s = \sum_{k=1}^{N}(1-\Theta_{sk})$ with $\eta_b = \eta_s = 60°$

protein	inactive (PDB ID)	active (PDB ID)	N	N_a	N_b	N_s	activation
Ras	2q21[a]	5p21	166	82	23	74	GDP/GTP exchange
Calmodulin	1cfd[b]	1cll	144	104	19	98	Ca^{2+} binding
CheY	3chy	1fqw	128	31	5	28	phosphorylation

[a] To fit to the size of the shorter-length structure, C-terminal 5 residues are truncated.
[b] To fit to the size of the shorter-length structure, N-terminal 3 residues and the C-terminal 1 residue are truncated.

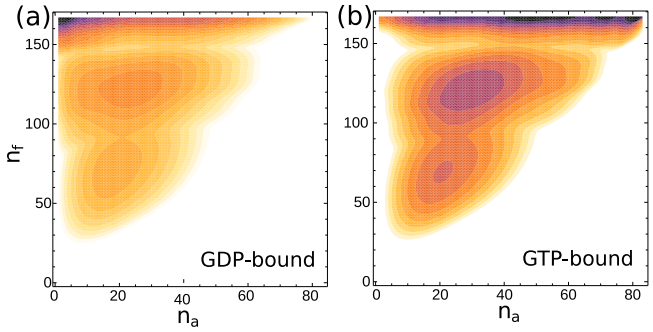

FIG. 5. Free energy surfaces in the space of the reaction coordinate (n_a, n_f) of Ras. Here, n_f is the number of residues that take the native (A, I, or C) configuration, i.e., the coordinate of folding process, and n_a is the number of residues that take the structure specific to the A conformation, i.e., the coordinate of allosteric transformation. $F_\alpha(n_a, n_f)$ in the GDP-bound (a) and GTP-bound (b) states are presented. $\varepsilon/k_BT = -0.052$. Contour is drawn in every $2k_BT$.

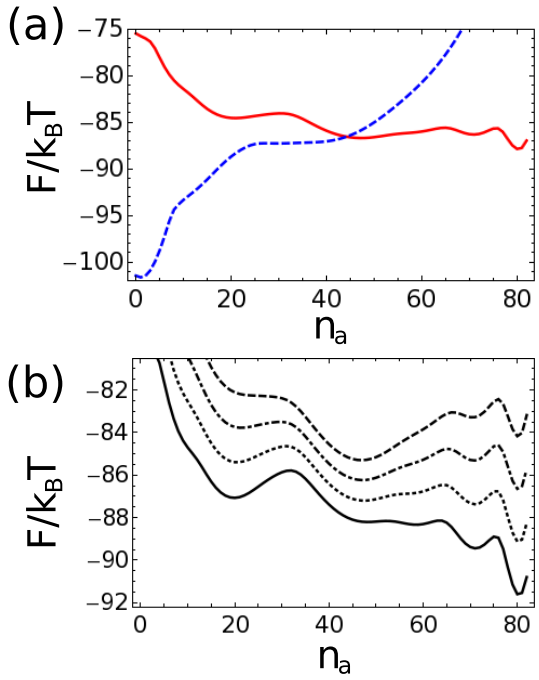

FIG. 6. Free energy surfaces of Ras in the fully folded state (the coordinate of folding is fixed at $n_f = N$) are shown as functions of the coordinate of allosteric transformation, n_a. (a) The free energy surfaces $F_\alpha(n_a, n_f = N)$ in the GTP-bound ($\alpha = t$: solid red line) and GDP-bound ($\alpha = d$: dashed blue line) states at $\varepsilon/k_BT = -0.052$ are shown. (b) By increasing temperature (decreasing $|\varepsilon/k_BT|$) the active state of the GTP-bound form at $n_a \approx 80$ becomes less stable and the transition to an intermediate state at $n_a \approx 45$ takes place. $F_\alpha(n_a, n_f = N)$ at $\varepsilon/k_BT = -0.055$ (solid line), -0.053 (dotted line), -0.051 (dot-dashed line), and -0.049 (dashed line) are shown.

barrier, so that the free energy surfaces of the allosteric conformational change can be approximately described along the coordinate fixed at $n_f = N = 166$. At $n_f = N$, as shown in Fig.6(a) for the GDP-bound case ($\alpha = d$), there is a distinct free energy minimum at $n_a = 0$ showing the sufficient stability of the I conformation in the GDP-bound state. With activation of about $10k_BT$ within this free energy valley, however, n_a changes to 15, leading to considerable structural fluctuations around I conformation as has been observed with NMR in the GDP-bound state [40]. In contrast, in the GTP-bound state ($\alpha = t$), the free energy minimum at $n_a = 80$ corresponding to the A conformation is shallow and the free energy barrier between the A state at $n_a = 80$ and an intermediate state at $n_a = 70$ is less than a few k_BT. As shown in Fig.6(b), by decreasing ε or increasing temperature, the A state at around $n_a = 80$ becomes less stable and the transition to an intermediate state at around $n_a = 45$ takes place. The calculated shallow free energy basin in Fig.6 suggests the large structural fluctuation with varying n_a: Conformations with different n_a would appear in the GTP-bound state by overpassing free energy barriers of a few k_BT, which should correspond to the motion in timescale of microseconds to milliseconds assuming the Arrhenius law with the pre-factor $10^5 s^{-1}$ to $10^6 s^{-1}$ [22, 25]. The suggested large conformational fluctuation is consistent with the observed large fluctuation in a single-molecule FRET measurement [38] and the NMR data [39] of the GTP-bound Ras .

Relative importance of the side-chain-configuration change can be illustrated by plotting a free energy surface in the space of the two-dimensional coordinate (n_b, n_s). Shown in Fig.7 is the free energy surface in the space of (n_b, n_s) in the GTP-bound state at $n_f = N$. On this surface, $(n_b, n_s) = (0, 0)$ corresponds to the I conformation and $(n_b, n_s) = (N_b, N_s) = (23, 74)$ corresponds to the A conformation. From the A state at $(n_b, n_s) = (N_b, N_s)$ to the intermediate at around $(n_b, n_s) = (18, 42)$, variation of n_b is much smaller than that of n_s, showing that

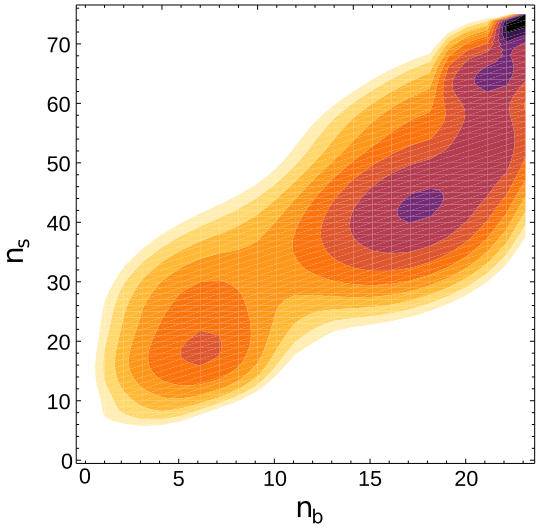

FIG. 7. A free energy surface in the space of the reaction coordinate (n_b, n_s) of Ras. Here, $n_{b(s)}$ is the number of residues that take the backbone (side-chain) configuration specific to the A conformation. $F_\alpha(n_b, n_s)$ fixed at $n_f = N$ in the GTP-bound ($\alpha = t$) state is presented. $\varepsilon/k_{\rm B}T = -0.052$. Contour is drawn in every $k_{\rm B}T$.

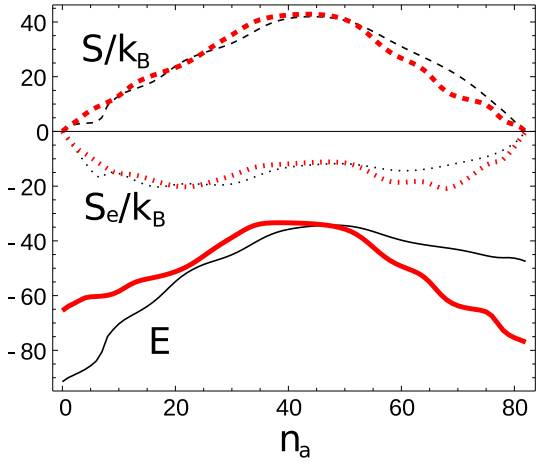

FIG. 8. Energy and entropy of Ras in the GTP-bound (thick red lines) and GDP-bound (thin black lines) states. Energy E_α (solid lines), entropy S_α (dashed lines), and the Hamiltonian dependent part of entropy $S_{e\alpha}$ (dotted lines), fixed in the fully folded state at $n_f = N$, are represented along the reaction coordinate of allosteric transformation, n_a. $\varepsilon/k_{\rm B}T = -0.052$. Values of energy are scaled by $k_{\rm B}T$.

this structural change is almost dominated by the side-chain rearrangement. On the other hand, in fluctuation between the intermediates at around $(n_b, n_s) = (18, 42)$ and the state at $(n_b, n_s) = (6, 18)$, the backbone and side-chain structural changes take place in a concomitant way.

Free energy surfaces can be further analyzed by decomposing them into energetic and entropic parts. Shown in Fig.8 are the one-dimensional surfaces of energy E_α, entropy S_α, and the Hamiltonian-dependent part of entropy $S_{e\alpha}$ in $\alpha = d$ and t, plotted in the coordinate n_a with the fixed condition of $n_f = N$. In the GTP-bound state, change in TS_α in the region for $30 < n_a < 82$ is comparable with the change in energy in that region, showing that the cancellation between energy and entropy leads to the gentle free energy landscape for $n_a > 35$ as was shown in Fig.6. The entropic cost $S_{e\alpha}(n_a)$ is relatively small for $35 < n_a < 50$, suggesting the coexistence of a relatively large number of conformations in each energy range in this region. At $n_a = 70$ in the GTP-bound state, $|S_{e\alpha}(n_a)|$ is relatively large, suggesting that the smaller number of specific pathways are dominant along the direction of energy decrease toward the A state from the intermediate state at $n_a = 70$.

Precise structural information is obtained by calculating structural order parameters $\rho^\sigma_{\alpha i}$ and $\xi^\sigma_{\alpha i}$ with $\sigma =$ A or I. Figs.9(a)-9(d) show $\rho^\sigma_{ti}(n_a, n_f)$ and $\xi^\sigma_{ti}(n_a, n_f)$ at $n_a = 70$ and 45 with the fixed condition of $n_f = N$ in the GTP-bound state. At $n_a = 70$ (Figs.9(a) and 9(b)) $\rho^{\rm A}_{ti}$ and $\xi^{\rm A}_{ti}$ are small in the C-terminal region ($i > 112$), while they remain large for $i \leq 112$. The side-chain configuration of the helix $\alpha 4$ (residue 127-137) is similar to that in the I conformation with $\rho^{\rm I}_{ti} > 0.5$, $\xi^{\rm I}_{ti} > 0.5$ and $\rho^{\rm A}_{ti} < 0.2$, $\xi^{\rm A}_{ti} < 0.2$. Thus, by leaving from the A state to the intermediate state at $n_a = 70$, the A structure is replaced by the I structure for $i > 112$, while the A structure remains for $i \leq 112$.

By further shifting conformations to the other intermediate state at $n_a = 45$ (Figs.9(c) and 9(d)), ρ^σ_{ti} and ξ^σ_{ti} ($\sigma = {\rm A}, {\rm I}$) for $i > 112$ are not much changed from those at $n_a = 70$, but $\rho^{\rm A}_{ti}$ and $\xi^{\rm A}_{ti}$ decrease for $i \leq 112$ except at the center of Switch I region. In the regions $1 \leq i \leq 5$, $22 \leq i \leq 52$, $58 \leq i \leq 59$, $74 \leq i \leq 91$, and $105 \leq i \leq 109$ (see Fig.10), each type-$\bar{\rm C}$ residue takes the A and I structures with nearly equal probability as $\rho^{\rm A}_{ti} \approx \rho^{\rm I}_{ti} \approx 0.5$ ($i \in \bar{\mathcal{C}}$), and the native contacts which are specific to the A or I structure are almost broken as $\xi^{\rm A}_{ti}$, $\xi^{\rm I}_{ti} < 0.2$. In other words, a large number of conformations having patterns of mixed A and I stretches coexist in this state at $n_a = 45$, which leads to the large configuration entropy of $\approx 40 k_{\rm B}$ and stabilizes this state. Since $\Theta_{kb} = 1$ and $\Theta_{ks} = 0$ at many residues of $\rho^{\rm A}_{ti} \approx \rho^{\rm I}_{ti} \approx 0.5$ (Figs.9(c) and (d)), this large number of structural patterns and the associated large entropy mainly come from the side-chain rearrangements. It is interesting to see that these regions having the mixed A and I structures (green-colored regions in Fig.10) overlap with the regions where Ras binds the target protein [36]. In the intermediate state at around $n_a = 20$, the Switch II region tends to take I configurations. The side-chain rearrangements in the intermediate at around $n_a = 45$ and the backbone fluctuation at around $n_a = 45$ to 20 may correspond to the appearance of multiple conformational states of *regional polysterism* proposed in ref.[37].

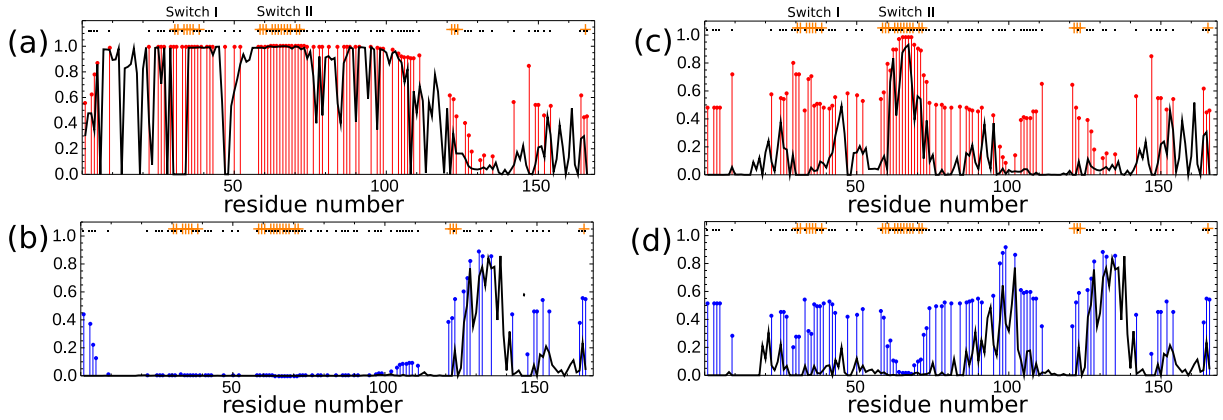

FIG. 9. Order parameter of formation of the σ (= A or I) structure at the ith residue, $\rho^{\sigma}_{\alpha i}(n_a, N)$, and order parameter of formation of contacts specific to the σ structure at around the ith residue, $\xi^{\sigma}_{\alpha i}(n_a, N)$, of the GTP-bound Ras ($\alpha = t$). The order parameters at $n_a = 70$ for $\sigma =$ A (a) and $\sigma =$ I (b), and at $n_a = 45$ for $\sigma =$ A (c) and $\sigma =$ I (d) are shown. Points with vertical lines represent ρ^{σ}_{ti} and connected lines represent ξ^{σ}_{ti}. "+" and dot "." presented at top of each panel indicate the sites with $\Theta_{kb} = 0$ and $\Theta_{ks} = 0$, respectively. $\varepsilon/k_{\rm B}T = -0.052$.

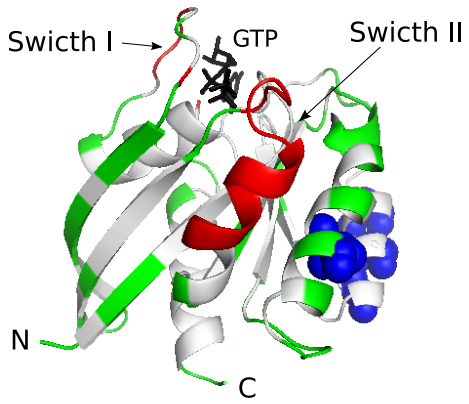

FIG. 10. Order parameter of formation of the σ (= A or I) structure, $\rho^{\sigma}_{\alpha i}(n_a, N)$, in the intermediate state ($n_a = 45$) of the GTP-bound ($\alpha = t$) Ras. Red colored indicate residues of $\rho^{\rm A}_{ti} \geq 0.7$, and green colored indicate residues of $0.4 \leq \rho^{\rm A}_{ti} \leq 0.6$ and $0.4 \leq \rho^{\rm I}_{ti} \leq 0.6$ with small $\xi^{\rm A}_{ti}$ and $\xi^{\rm I}_{ti}$. Blue spheres are residues of $\rho^{\rm I}_{ti} \geq 0.7$.

2. *Calmodulin*

Calmodulin (CaM) consists of two structurally similar domains (N- and C-domains) connected by a flexible linker. Each domain contains two EF-hand (helix-loop-helix) Ca^{2+}-binding motifs. The inactive closed conformation (I) is switched to the active open conformation (A) by binding Ca^{2+} ions (Fig.4(b)) [41, 42]. The linker has a loop structure in I conformation, whereas the linker forms a helix in A conformation.

A set of Ca^{2+}-binding residues in the kth EF-hand motif are denoted by EF_k ($k = 1, 2, 3, 4$) with $\mathcal{L} = \cup_{k=1}^{4}\mathrm{EF}_k$, where k is numbered from the N- to C-termini along the protein chain. We express the Ca^{2+}-binding states of $\mathcal{L}$ as $\alpha =$ "$\alpha_1\alpha_2\alpha_3\alpha_4$" with $\alpha_k = 0(1)$ representing the Ca^{2+}-free(binding) state of EF_k, so that the Ca^{2+}-free (apo) state is $\alpha = 0000$, the state in which two Ca^{2+} ions bind to the N(C)-domain is $\alpha = 1100(0011)$, and the state in which four Ca^{2+} ions bind to N- and C-domains (holo) is $\alpha = 1111$. $\mathcal{B}_k$ ($k = 1, 2, 3, 4$) is a set of pairs $\langle i, j\rangle$, where at least either i or j is the Ca^{2+}-binding site of EF_k. See appendix B for more details.

In Fig.11(a) the free energy surface $F_\alpha(n_a, n_f)$ in the Ca^{2+}-free state $\alpha = 0000$ is shown in the two-dimensional space of n_f and n_a. While the I conformation at $(n_a, n_f) = (8, 144)$ is most stable for $\varepsilon/k_{\rm B}T < -0.07$, there are two other free energy minima of metastable states at $(n_a, n_f) = (25, 136)$ and at $(n_a, n_f) = (55, 144)$.

Conformation is partially unfolded at the state around $(n_a, n_f) = (25, 136)$; About ten residues of the C-domain take nonnative configurations with $\sum_{i=84}^{147}\rho^{\rm N}_{\alpha i}/64 \approx 0.85$, while $\rho^{\rm N}_{\alpha i} \approx 1$ for residues of the N-domain and the linker region. Especially, the loop region including Ca^{2+}-binding site EF_4 is unfolded with the probability of 50% as $\sum_{i=129}^{138}\rho^{\rm N}_{\alpha i}/10 \approx 0.5$. In this way, the partial unfolding plays a role in the allosteric transition of CaM. This result for partial unfolding or "cracking" at the C-domain is consistent with the results calculated with a variational model [43, 44]. In other states than $\alpha = 0000$, however, there is no free energy minimum at which apparent unfolding is induced for $\varepsilon/k_{\rm B}T < -0.07$ as shown in Figs.11(b) and 11(c). Thus, in a unified analysis of all the Ca-binding states, it is convenient to plot free energy surfaces with the fixed n_f at $n_f = N$ as shown in Fig.12.

An interesting scenario of the Ca^{2+} binding process is suggested by analyzing structures appearing in Fig.12. The state of $\alpha = 0000$ has a shallow free energy minimum at $(n_a, n_f) = (35, 144)$. Both this state at $(n_a, n_f) = (35, 144)$ and the partial unfolded state at $(n_a, n_f) = (25, 136)$ are separated from the I conformation by free energy barriers of $\approx 7k_{\rm B}T$, whereas the free energy barrier between states at $(n_a, n_f) = (35, 144)$ and $(25, 136)$ is $\approx 3k_{\rm B}T$. At this minimum of $(n_a, n_f) = (35, 144)$, as

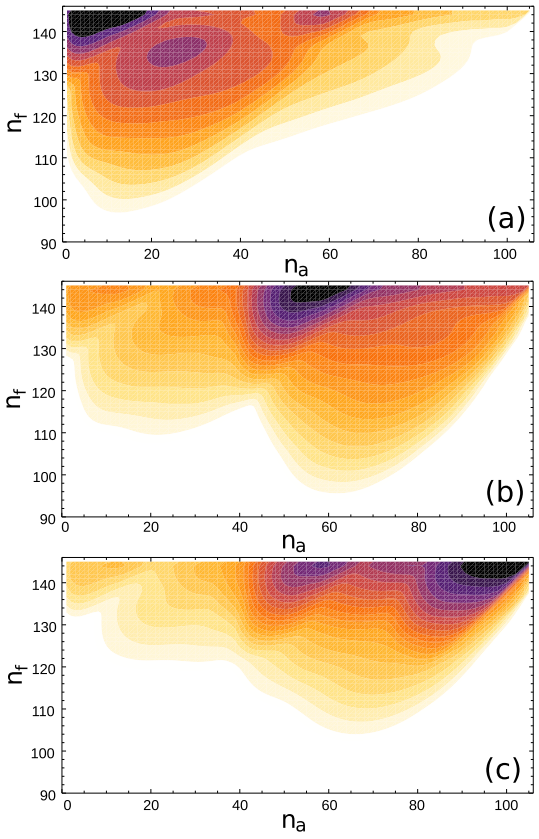

FIG. 11. Free energy surfaces in the space of the reaction coordinate (n_a, n_f) of calmodulin. Here, n_f is the number of residues that take the native (A, I, or C) configuration, i.e., the coordinate of folding process, and n_a is the number of residues that take the structure specific to the A conformation, i.e., the coordinate of allosteric transformation. $F_\alpha(n_a, n_f)$ in the states $\alpha = 0000$ (a), $\alpha = 0011$ (b), and $\alpha = 1111$ (c) are presented. $\varepsilon/k_BT = -0.08$. Contour is drawn in every $2k_BT$.

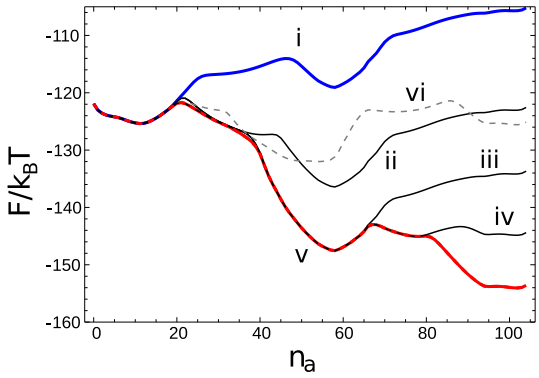

FIG. 12. Free energy surfaces of calmodulin in the fully folded state (the coordinate of folding is fixed at $n_f = N$) are shown as functions of the coordinate of allosteric transformation, n_a. The free energy surfaces $F_\alpha(n_a, n_f = N)$ in $\alpha = 0000$ (i), $\alpha = 0001$ (ii), $\alpha = 0011$ (iii), $\alpha = 0111$ (iv), $\alpha = 1111$ (v), and $\alpha = 1100$ (vi) are shown. $\varepsilon/k_BT = -0.08$

shown in Fig.13(a), the A structures are developed in the C-terminal half of the C-domain, $0.4 < \rho^{\rm A}_{0000i} < 0.6$ for $i > 110$, and $\xi^{\rm A}_{0000i}$ has relatively large values at around $n_a = 120$ and 140. These large values of $\rho^{\rm A}_{0000i}$ and $\xi^{\rm A}_{0000i}$ imply that the A structures emerge at around EF_4 before Ca^{2+} binding. This pre-existence of the A structures is consistent with the NMR data of the fragment of the C-terminal domain [45]. By the thermal excitation of several k_BT from the I conformation or from the intermediate state at $(n_a, n_f) = (25, 136)$ to the intermediate at $(n_a, n_f) = (35, 144)$, the pre-existing A structures appear at around EF_4, which promotes Ca^{2+} binding to the site EF_4, changing the state of CaM from $\alpha = 0000$ to $\alpha = 0001$.

In the state $\alpha = 0001$, the global free energy minimum is at $(n_a, n_f) \approx (55, 144)$. By Ca^{2+} binding to EF_4, the whole C-domain and the C-terminal half of the linker are stabilized to have A structures at $(n_a, n_f) \approx (55, 144)$, while the I structures still remain at the N domain (Fig.13(b)). This structure, then, promotes Ca^{2+} binding at EF_3, transforming CaM to the $\alpha = 0011$ state.

In the state $\alpha = 0011$ at $(n_a, n_f) \approx (55, 144)$, the N domain still has the I structural features but with the thermal activation of $\approx 8k_BT$ to the state $(n_a, n_f) \approx (75, 144)$ (iii in Fig.12), the N-terminal half of the linker and the binding site at EF_2 develop the A structures as shown in Fig.13(c). This pre-existing state should promote Ca^{2+} binding to EF_2, leading CaM to the state $\alpha = 0111$. In the state $\alpha = 0111$, although the free energy minimum remains at $n_a \approx 57$, the A structures develop in the whole N-domain when CaM is thermally excited to surpass the free energy barrier of a few k_BT. Appearance of A structures through excitation over this small free energy barrier then promotes Ca^{2+} binding to EF_1, which stabilizes the A structure of CaM extending over the entire protein chain from C- to N-domains.

As shown in Fig.13, residues with $\Theta_{bk} = 1$ and $\Theta_{sk} = 0$ are relatively densely populated in regions of EF_2 and EF_3. We should note, therefore, that side-chain configurational fluctuations play important roles in pre-existence fluctuations to generate the $\alpha = 0011$ state from the $\alpha = 0001$ state and those to generate the $\alpha = 0111$ state from the $\alpha = 0011$ state. In this way the pre-existing side-chain fluctuations at EF_2 and EF_3 and backbone fluctuations at EF_1, EF_4, and the linker region bring about the sequential development of A structures from C terminus through the linker to the N terminus, which leads to the sequential binding of Ca^{2+} to CaM. The dominant pathway of the allosteric transition obtained by this calculation is summarized in Fig.14.

By the Ca^{2+} binding to the N-(C-)domain, entropy is reduced and $|S_{e\alpha}|$ is increased in the course of conformational transition of the N-(C-)domain at around $n_a \approx 70$ ($n_a \approx 40$) as shown in Fig.15. Thus, the Ca^{2+} binding reduces the number of transition pathways but lowers the transition energy to activate the pre-existing fluctuations which further promote Ca^{2+} binding to guide CaM toward the A conformation.

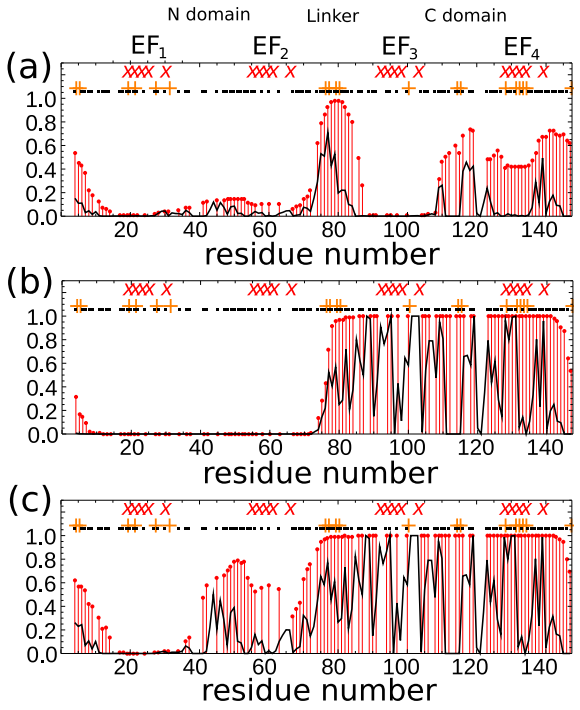

FIG. 13. Order parameter of formation of the A structure at the ith residue, $\rho^{\rm A}_{\alpha i}(n_a, N)$, and order parameter of formation of contacts specific to the A structure at around the ith residue, $\xi^{\rm A}_{\alpha i}(n_a, N)$, of calmodulin. $\rho^{\rm A}_{\alpha i}$ (points with vertical lines) and $\xi^{\rm A}_{\alpha i}$ (connected lines), in the states (a) $\alpha = 0000$ at $n_a = 35$, (b) $\alpha = 0001$ at $n_a = 55$, (c)$\alpha = 0011$ at $n_a = 75$. "+" and dot "." presented at top of each panel indicate residues with $\Theta_{k\rm b} = 0$ and $\Theta_{k\rm s} = 0$, respectively. "X" indicates the Ca^{2+}-binding residues. A trucated sequence of $4 \le i \le 147$ is used. $\varepsilon/k_{\rm B}T = -0.08$.

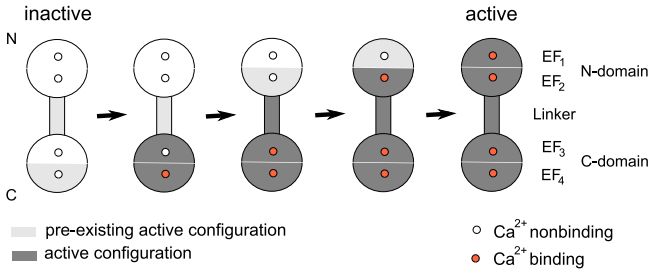

FIG. 14. Picture of sequential Ca^{2+} binding driven by a series of pre-existing fluctuations of calmodulin.

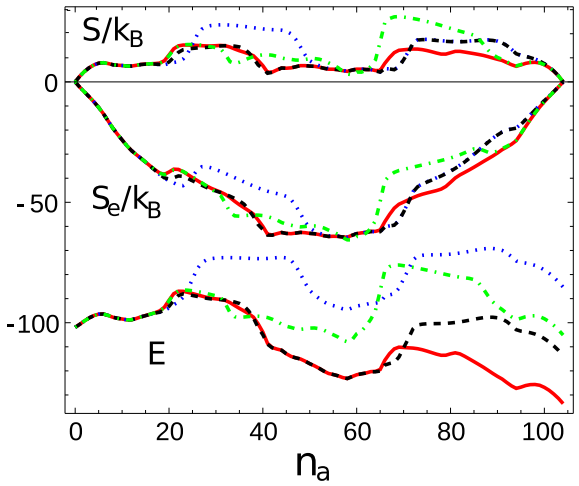

FIG. 15. Energy and entropy of calmodulin fixed in the fully folded state at $n_f = N$ are plotted along the reaction coordinate of allosteric transformation, n_a. Energy, entropy, and the Hamiltonian dependent part of entropy $S_{{\rm e}\alpha}$ are represented in the states $\alpha = 0000$ (dotted blue lines), $\alpha = 1100$ (dot-dashed green lines), $\alpha = 0011$ (dashed black lines), and $\alpha = 1111$ (solid red lines). $\varepsilon/k_{\rm B}T = -0.08$. The values of energy are scaled by $k_{\rm B}T$.

It is interesting to see that in Fig.13(b) for $\alpha = 0001$ at $n_a = 55$, N-domain is apo-like, C-domain is holo-like, and the linker is fluctuating between apo- and holo-structures. In the Ca^{2+}-saturated $\alpha = 1111$ state, as shown in Fig.12, free-energy difference between the state at around $n_a = 55$ and the state around A conformation at $n_a = 104$ is several $k_{\rm B}T$, so that conformations of $n_a \approx 50$ and those of $n_a \approx 100$ should coexist in equilibrium. This results is consistent with the observed NMR data for solution structures of Ca^{2+}-saturated CaM, which has shown that C-domain is similar to the crystal structure of A conformation but N-domain is considerably less open [46], and that the linker helix is highly flexible [47].

3. *CheY*

CheY is a prototypical response regulator in two-component signal transduction systems [48, 49]. Its inactive conformation (I) is switched to the active conformation (A) by phosphorylation of the residue, Asp57 (Fig.4(c)) [50, 51]. The backbone configurations do not significantly differ between I and A structures except at the $\beta4$-$\alpha4$ loop region. Changes in the side-chain configurations and the contact pairs, on the other hand, are more significant and spreading around the phosphorylation site as shown in Fig.4(d) and Fig.16. Asp57 is the allosteric site $\mathcal{L}$ and the set of modulated pairs $\mathcal{B}$ are the collection of $\langle i, j\rangle$ in which i or j is Asp57 or the residue that contacts Asp57 (we consider the truncated sequence $2 \le i, j \le 129$). See Appendix B for more details.

As shown in Fig.17, the transition involving the I conformation in the dephosphorylated state at $(n_a, n_f) = (1, 128)$ and the A conformation in the phosphorylated state at $(n_a, n_f) = (29, 128)$ occurs in the native basin $120 < n_f \le N$. For $\varepsilon/k_{\rm B}T < -0.054$, the A conformation at the native basin is stable in the phosphorylated state and the I conformation in the native basin is stable in the dephosphorylated state. It should be noted that free energy basin is shallow in both phosphorylated and dephosphorylated states, so that large fluctuations at the loop region and in side-chain rotamers are expected

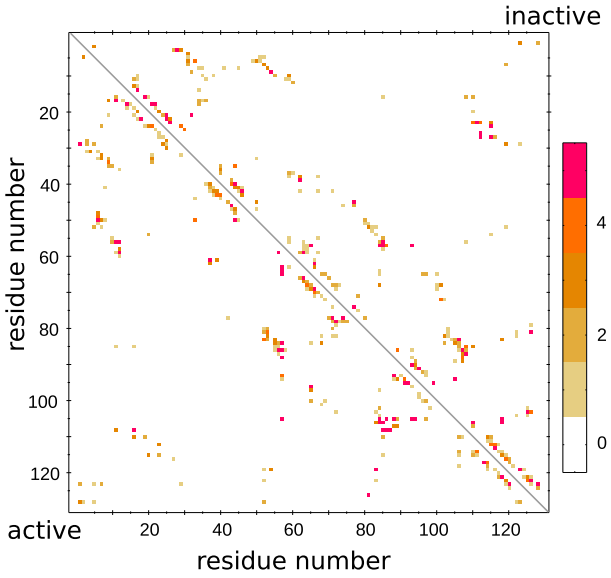

FIG. 16. Contact maps $q^{\rm I}_{i,j}$ (upper triangle) and $q^{\rm A}_{i,j}$ (lower triangle) of CheY.

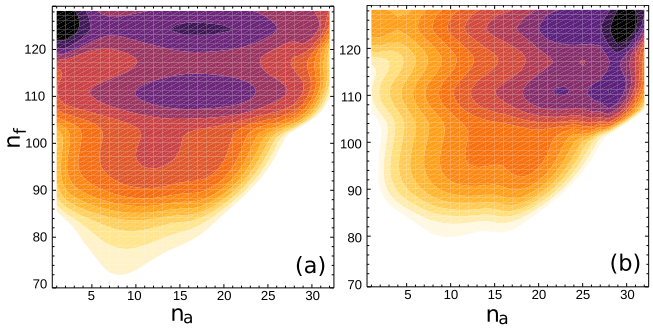

FIG. 17. Free energy surfaces in the space of the reaction coordinate (n_a, n_f) of CheY. Here, n_f is the number of residues that take the native (A, I, or C) configuration, i.e., the coordinate of folding process, and n_a is the number of residues that take the structure specific to the A conformation, i.e., the coordinate of allosteric transformation. $F_\alpha(n_a, n_f)$ in the dephosphorylated (a) and phosphorylated (b) states are presented. $\varepsilon/k_{\rm B}T = -0.057$. Contour is drawn in every $k_{\rm B}T$.

in both states, but the amplitude of fluctuations should be larger in the dephosphorylated state with the shallower free energy surface, which leads to the pre-existing fluctuations (appearance of A structures before phosphorylation) as discussed below. In the dephosphorylated state, there is a shallow basin of an intermediate state at around $(n_a, n_f) = (15, 125)$. The free energy barrier between the I state and the intermediate is about $3k_{\rm B}T$ at $\varepsilon/k_{\rm B}T = -0.057$. In the phosphorylated state, after surpassing the small barrier near the I state, the free energy surface toward the A conformation is almost downhill-like as shown in Fig.17(b). These features are approximately described along the coordinate fixed at $n_f = N$ as shown in Fig.18.

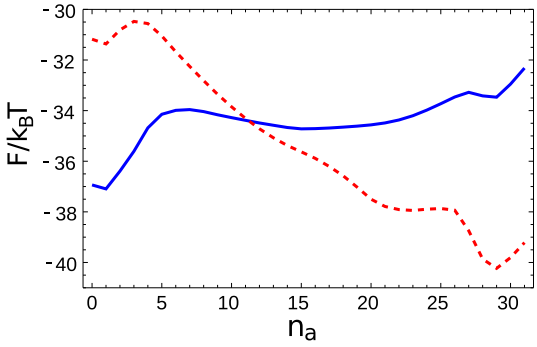

FIG. 18. Free energy surfaces fixed at $n_f = N$ of CheY in the fully folded state (the coordinate of folding is fixed at $n_f = N$) are shown as functions of the coordinate of allosteric transformation, n_a. $F_\alpha(n_a, n_f = N)$ in the dephosphorylated (solid blue line) and phosphorylated (dashed red line) states at $\varepsilon/k_{\rm B}T = -0.057$ are shown.

The free energy surface of the dephosphorylated CheY in the space of the reaction coordinate (n_b, n_s) is shown in Fig.19. In this coordinate, $(n_b, n_s) = (0, 0)$ for the I conformation and $(n_b, n_s) = (N_b, N_s) = (5, 31)$ for the A conformation. In structural change between the I state and the intermediate state, the backbone configurations at residues 58 and 86-91 ($\beta4$-$\alpha4$ loop) are fluctuating between I and A configurations, associated with the side-chain configurational changes of about 7 residues. The backbone fluctuations between the A and I configurations in the intermediate conformations are coupled with the side-chain fluctuation between A and I.

Fig.20 shows the order parameters $\rho^\sigma_{\alpha i}(n_a, N)$ and $\xi^\sigma_{\alpha i}(n_a, N)$ (σ = A, I) in the dephosphorylated state ($\alpha = 0$) at the intermediate state of $(n_a, n_f) = (15, 128)$. In addition to the $\beta4$-$\alpha4$ loop and its vicinity, residues i = 47, 62, 66-67, 75, 94, 100, and 106 have medium to high values of $\rho^{\rm A}_{0i}$, showing that side-chain configurations are changed from I to A at these residues by shifting CheY from the I state to this intermediate state, which characterizes the pre-existing fluctuation in this protein. In the N-terminal region ($i \le 47$) and the C-terminal helix ($i > 115$), on the other hand, the I structure appears more frequently than the A structure at each residue as $\rho^{\rm I}_{0i} > \rho^{\rm A}_{0i}$, but contacts which are specific to the A structure are beginning to develop throughout these regions. Such coexistence of I and A structures, in which local configurations and contact pairs are dynamically fluctuating between I and A, provides entropy of $\approx 13k_{\rm B}$ as shown in Fig.22. We should note that the side-chain configurational changes contribute more significantly to this entropy than the backbone structural changes which are only localized in the protein as shown in Fig.21.

Fig.20 also shows that the native contacts specific to the A structure develop at $i = 67, 70-75, 78-80, 83$ with $\xi^{\rm A}_{0i} \ge 0.6$ and $\xi^{\rm I}_{0i} \le 0.2$ in the intermediate state. The native structures of most of these residues are common to the A and I structures, i.e., most residues are type C residues, and hence the contact pairs specific to the

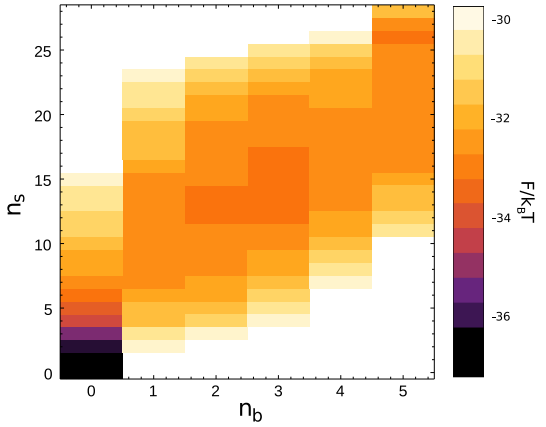

FIG. 19. A free energy surface in the space of the reaction coordinate (n_b, n_s) of CheY. Here, $n_{b(s)}$ is the number of residues that take the backbone (side-chain) configuration specific to the A conformation. $F_\alpha(n_b, n_s)$ fixed at $n_f = N$ in the dephosphorylated state ($\alpha = 0$) is presented. $\varepsilon/k_{\rm B}T = -0.057$.

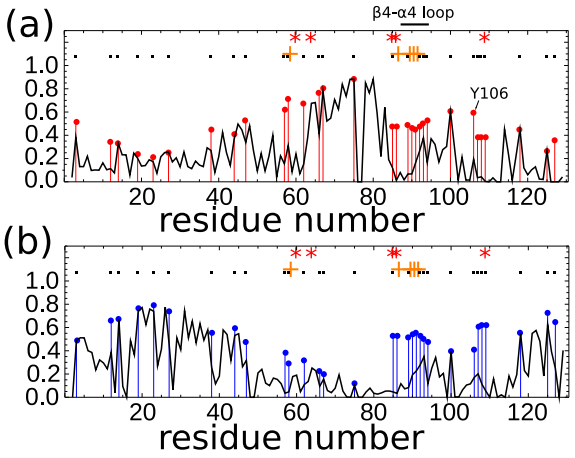

FIG. 20. Order parameter of formation of the σ (= A or I) structure at the ith residue, $\rho^\sigma_{\alpha i}(n_a, N)$, and order parameter of formation of contacts specific to the σ structure at around the ith residue, $\xi^\sigma_{\alpha i}(n_a, N)$, of CheY in the dephosphorylated state $\alpha = 0$. ρ^σ_{0i} (points with vertical lines) and ξ^σ_{0i} (connected lines) for σ = A (a) and σ = I (b) at $n_a = 16$. "+" and dot "." presented at top of each panel indicate residues with $\Theta_{k\rm b} = 0$ and $\Theta_{k\rm s} = 0$, respectively. "∗" indicates the neighbor residue of the phosphorylation site. A truncated sequence of $2 \le i \le 129$ is used. $\varepsilon/k_{\rm B}T = -0.057$.

A structure are stabilized when neighbor residues have features of A structures in the pre-existing fluctuation. Therefore, the backbone structural change of the β4-α4 loop from I to A configurations is coupled with the stabilization of contacts in these Type C residues, which contributes to spreading the A structure over the entire protein (Fig.21).

The activity of CheY to bind the partner protein is intrinsically related to the configuration of Tyr106 which is located on the binding surface. Therefore, the mechanism how the phosphorylation of Asp57 leads to the side-chain rotation of Tyr106 has been attracted much interest [52–54]. A proposed mechanism was the Y-T coupling mechanism, in which phosphorylation of Asp57 displaces Thr87 through formation of a hydrogen bond between them, which creates the space to allow Tyr106 to rotate [52]. In the present calculation, however, Tyr106 rotates to take the active side-chain configuration even in the dephosphorylated state with the relatively large probability at the intermediate state. This rotation is associated with the backbone configuration change at the β4-α4 loop, which includes Thr87, and such pre-existing Y-T coupling fluctuation is further stabilized upon phosphorylation. This population-shift mechanism of rotation of Tyr106 is consistent with the result obtained by a molecular dynamics calculation [54].

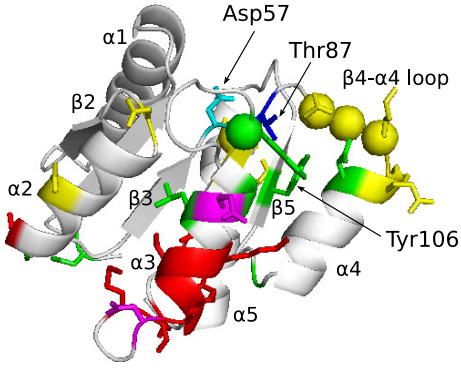

FIG. 21. Residues developing the active configuraions and contacts in the intermediate state at $n_a = 15$ in the dephosphorylated CheY ($\alpha = 0$). Green colored indicate residues exhibiting the large value of order parameter of formation of the A structure with $\rho^{\rm A}_{0i} \ge 0.5$, and yellow colored indicate residues indicating the intermediate level of formation of the A structure with $0.4 \le \rho^{\rm A}_{0i} < 0.5$. Red colored indicate residues around which the interactions specific to the A structure are developed with $\xi^{\rm A}_{0i} \ge 0.6$ and $\xi^{\rm I}_{0i} \le 0.2$. Purple colored indicate residues with $\rho^{\rm A}_{0i} \ge 0.5$, $\xi^{\rm A}_{0i} \ge 0.6$, and $\xi^{\rm I}_{0i} \le 0.2$. Spheres indicate residues that change backbone configurations largely ($\Theta_{k\rm b} = 0$).

VII. DISCUSSION

In this paper, a statistical mechanical model of allosteric transition was developed and applied to example proteins. The model is based on the native interactions defined in two native conformations, active (A) and inactive (I) conformations, and explains the allosteric conformational change by showing the balance among various energetic and entropic effects. A remarkable feature in this balance is the importance of large entropy arising from the combinatorially large number of mosaic patterns of A and I structures along the protein chain. Together with this large positive contribution to entropy, the net entropy is calculated by suitably taking into account the

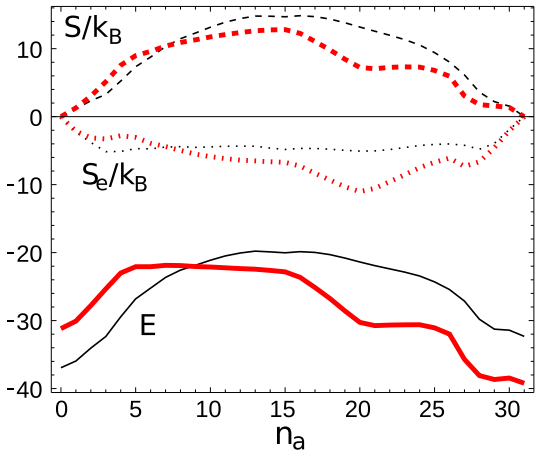

FIG. 22. Energy and entropy of CheY in the dephosphorylated (thin black lines) and phosphorylated (thick red lines) states. Energy (solid lines), entropy (dashed lines), and the Hamiltonian dependent part of entropy $S_{e\alpha}$ (dotted lines) fixed in the fully folded state at $n_f = N$, are represented along the reaction coordinate of allosteric transformation, n_a. $\varepsilon/k_BT = -0.057$. The values of energy are scaled by k_BT.

negative contribution of the Hamiltonian dependent part of entropy. As shown in Eq.45, this Hamiltonian dependent part of entropy reflects the many-residue correlation. Since the present model treats many-body correlations among residues by describing "stretches", or contiguous segments of residues having the same structural features, the model was able to describe the folding processes of many proteins in a quantitative way [22, 25, 27–29], and hence we expect that the many-body treatment in the present model also allows us to evaluate the allosteric transition processes in a suitable way.

In the present model, three-dimensional coordinates of atoms are used to define native interactions, allosteric sites, and modulated interactions around them, to distinguish sites which show the large backbone configuration change and sites which show the large side-chain structural change, and to distinguish type-C and type-$\bar{\text{C}}$ sites. In other conformations than A or I, the three-dimensional chain structure is not explicitly calculated in the model but is represented by four letter alphabet of A, I, C, and D, so that one might consider that the configurational entropy might be overestimated in those conformations by counting the unrealizable chain structure in which atoms collide each other in more realistic atomic models. Though this possibility of overcounting should be carefully checked by examining the atomistic models in future studies, we here note that the type -$\bar{\text{C}}$ sites with the change in the backbone structure are often found around loop motifs and the type -$\bar{\text{C}}$ sites with the change in the side-chain configuration are distributed over the protein surface. Those sites should be flexible in configuration to take various backbone structures or side-chain rotamer states, so that the configurational entropy should be indeed important for these sites, not to be underestimated. In our example proteins, side-chain rotamer variation is evident in the surface region of Ras at which Ras binds target proteins and the backbone change is found in Switch I and II regions. The side-chain rotamer variation without large backbone changes is clear around the 2nd and 3rd calcium binding sites in calmodulin (CaM), while the backbone configulational change is found in other calcium binding sites and at the linker region. In CheY, the allosteric change is dominated by the side-chain rotamer variations at the protein surface. Thus, in these examples, the mosaic patterns of different configurations would naturally take place in such structurally flexible regions to give rise to smooth free energy surfaces of allosteric conformational change.

The model explained important features of example proteins which are consistent with the experimental data. In Ras, the I conformation in the GDP-bound state is stabilized enough to prevent unnecessary signaling from A-like conformations, but the A-conformation in the GTP-bound state bears larger fluctuation. This fluctuation is consistent with the observations [37, 38] and should be needed to facilitate binding multiple target proteins. There are multiple intermediates found in the calculation, and fluctuation among the A conformation ($n_a \approx 80$), the intermediate at $n_a \approx 70$, and the intermediate at $n_a \approx 45$ is dominated by fluctuation in the side-chain configurations, while the fluctuation between intermediates at $n_a \approx 45$ and at $n_a \approx 20$ are coupled fluctuation of side-chain configuration and backbone structure in Switch I and II regions. The backbone structural change at Switch I and II, therefore, separates the I-like conformations from conformations exhibiting the large fluctuation around the A conformation.

In CaM, the calculated results suggested the scenario of sequential Ca^{2+} binding and the associated conformational change. How the four binding sites of CaM bind Ca^{2+} was distinguished by the 4-bit symbol of α in this paper, and the model showed that with the thermal excitation of several k_BT from a most stable free-energy minimum of the state α, the structure suitable to bind Ca^{2+} is prepared as a pre-existing fluctuation at the unbound site, which leads CaM to the next bound state α', and the sequential binding process is selected though such sequential activation of pre-existing fluctuations. This sequence of Ca^{2+} binding is consistent with the observed ordering of Ca^{2+} binding events in CaM [55, 56]. It should be noted that the pre-existing fluctuation from $\alpha = 0001$ to $\alpha = 0011$ and that from $\alpha = 0011$ to $\alpha = 0111$ in this sequence are dominated by fluctuations in the side-chain configuration, and other pre-existing fluctuations include both backbone structural fluctuations and side-chain variations.

In CheY, the side-chain rotamer variation dominates fluctuations between the I conformation in the dephosphorylated state to the A conformation in the phosphorylated state. In the dephosphorylated state, the A structure appears around the phosphorylation site, Asp57, as a pre-existing fluctuation and this structure enhances

interactions between residues around Asp57 and those around Thr87 through the change in the backbone structure of $\beta 4-\alpha 4$ loop, and these interactions enhance the A-structural features at Tyr106, which can explain the pre-existing Y-T coupling mechanism in CheY functioning [52].

The model provides information on detailed structural features of intermediates and transition states, which should give further opportunities to check the model prediction experimentally. The exact solution of the model can be a starting point for analyzing approximate treatments with more complex models, especially in studying allosteric transitions in multimeric protein complexes. By extending the local equilibrium approach of Zamparo *et al.* [26] for the present model, it is possible to develop the theory to calculate kinetics of allosteric transition and fluctuations, which should further contribute in comparing the model results with the experimental data to resolve the mechanism of allosteric transitions.

ACKNOWLEDGMENTS

This work was supported by Grants-in-aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology, Japan, and from Japan Society for the Promotion of Science.

Appendix A: Calculation of the generating function

We calculate the generating function by rewriting Eqs. (32) and (35) to recurrent equations, respectively. The generating function is given by

$$\mathcal{G}(\mathbf{X}) = \Omega^{\mathrm{D}} Z_0^N(\mathbf{X}), \tag{A1}$$

where $Z_0^N(\mathbf{X})$ is obtained by solving the recurrent equations:

$$Z_k^0(\mathbf{X}) = W_{N+1-k,N}(\mathbf{X}), \quad k=0,1,\cdots,N, \tag{A2a}$$

$$Z_k^l(\mathbf{X}) = W_{N+1-k-l,N-l}(\mathbf{X}) Z_0^{l-1}(\mathbf{X}) + Z_{k+1}^{l-1}(\mathbf{X}), \quad k=0,1,\cdots,N-l, \quad l=1,2,\cdots,N, \tag{A2b}$$

where $W_{r,s}(\mathbf{X}) = w_{r,s}^{\mathrm{N}}(\mathbf{X})/w_{r,s}^{\mathrm{D}}$ and $W_{r+1,r}(\mathbf{X}) \equiv 1$. $W_{r,s}^N(\mathbf{X})$ is obtained by simultaneous equations as

$$W_{r,s}^{\mathrm{N}}(\mathbf{X}) = w_{r,s}^{\mathrm{N}}(\bar{\mathbf{X}}) \prod_{k=d_1+1}^{d} X_k^{n_k(r,s)} \tag{A3}$$

$$w_{r,s}^{\mathrm{N}}(\bar{\mathbf{X}}) = w_{r,s}^0 [A_0^{L_{r,s}-1}(\bar{\mathbf{X}}) + I_0^{L_{r,s}-1}(\bar{\mathbf{X}})]. \tag{A4}$$

The recurrent equations are given by

$$A_k^0(\bar{\mathbf{X}}) = W^{(r.s)\mathrm{A}}_{L_{r,s}-k,L_{r,s}}(\bar{\mathbf{X}}), \quad k=0,\cdots,L_{r,s}-1, \tag{A5a}$$

$$I_k^0(\bar{\mathbf{X}}) = W^{(r.s)\mathrm{I}}_{L_{r,s}-k,L_{r,s}}(\bar{\mathbf{X}}), \quad k=0,\cdots,L_{r,s}-1, \tag{A5b}$$

$$A_k^l(\bar{\mathbf{X}}) = W^{(r.s)\mathrm{A}}_{L_{r,s}-l-k,L_{r,s}-l}(\bar{\mathbf{X}}) I_0^{l-1}(\bar{\mathbf{X}}) + A_{k+1}^{l-1}(\bar{\mathbf{X}}), \tag{A5c}$$

$$I_k^l(\bar{\mathbf{X}}) = W^{(r.s)\mathrm{I}}_{L_{r,s}-l-k,L_{r,s}-l}(\bar{\mathbf{X}}) A_0^{l-1}(\bar{\mathbf{X}}) + I_{k+1}^{l-1}(\bar{\mathbf{X}}), \tag{A5d}$$

$$k=0,\cdots,L_{r,s}-l-1, \quad l=1,\cdots,L_{r,s}-1,$$

We calculate the recurrent equations algebraically and obtain the free energy and mean values at $(n_1,\cdots,n_d)$ by extracting the coefficients of $\prod_{k=1}^d X_k^{n_k}$ in the expansion of the generating function using *Mathematica.*.

Appendix B: Allosteric sites and modulated pairs

1. Ras

The allosteric sites $\mathcal{L}$ of Ras are residues which directly interact with GDP or GTP [57]. $\mathcal{L}$ depends on whether $\alpha = t$ or $\alpha = d$, so that we use the notation of $\mathcal{L}_\alpha$; $i =$ 12-18, 28-31, 34-35, 60, 116-117, 119-120, and 145-147 for $i \in \mathcal{L}_t$, and $i =$ 11-18, 28-30, 32, 116-117, 119-120, and 145-147 for $i \in \mathcal{L}_d$. The interaction energies between the nucleotide and these allosteric sites are assumed to be $\varepsilon_{Lti}^{\mathrm{A}} = 1k_{\mathrm{B}}T$ and $\varepsilon_{Lti}^{\mathrm{I}} = 0$ in the GTP-bound state and $\varepsilon_{Ldi}^{\mathrm{A}} = 0$ and $\varepsilon_{Ldi}^{\mathrm{I}} = 1k_{\mathrm{B}}T$ in the GDP-bound state. We set $\varepsilon_{L\alpha i}^{\mathrm{N}} = 0$ for both $\alpha = t$ and d.

Interactions around $\mathcal{L}_\alpha$ should depend on α, so that we write interacting pairs around $\mathcal{L}_\alpha$ as $\mathcal{B}_\alpha$: $\mathcal{B}_{d(t)} = \{\langle i,j\rangle\}$ is a set of contact pairs in which at least either of i or j belongs to $\mathcal{L}_{d(t)}$, and the common set of $\mathcal{B}_t$ and $\mathcal{B}_d$ is denoted by $\mathcal{B}_c$. In the GDP-bound state, we set interaction strength of contacts as $\bar{\gamma}_{di,j}^{\mathrm{I}} = \bar{\gamma}_{di,j}^{\mathrm{N}} = 1$ for the pair $\langle i,j\rangle \in \mathcal{B}_d$ or $\langle i,j\rangle \in \mathcal{B}_t$, and we set $\bar{\gamma}_{di,j}^{\mathrm{A}} = 0$ for the pair $\langle i,j\rangle \in \mathcal{B}_t$ and $\bar{\gamma}_{di,j}^{\mathrm{A}} = 1$ for $\langle i,j\rangle \in \mathcal{B}_d - \mathcal{B}_c$. In the GTP-bound state, we set $\bar{\gamma}_{ti,j}^{\mathrm{A}} = \bar{\gamma}_{ti,j}^{\mathrm{N}} = 1$ for the pair $\langle i,j\rangle \in \mathcal{B}_d$ or $\langle i,j\rangle \in \mathcal{B}_t$, and we set $\bar{\gamma}_{ti,j}^{\mathrm{I}} = 0$ for $\langle i,j\rangle \in \mathcal{B}_d$ and $\bar{\gamma}_{ti,j}^{\mathrm{I}} = 1$ for $\langle i,j\rangle \in \mathcal{B}_t - \mathcal{B}_c$.

2. Calmodulin

A set of Ca^{2+}-binding residues in the kth EF-hand motif are denoted by EF_k $(k=1,2,3,4)$ with $\mathcal{L} = \cup_{k=1}^4 \mathrm{EF}_k$, where k is numbered from the N- to C-termini along the protein chain. See Fig.13 for the location of EF_k. We express the Ca^{2+}-binding states of $\mathcal{L}$ as $\alpha =$ "$\alpha_1\alpha_2\alpha_3\alpha_4$" with $\alpha_k = 0(1)$ representing the Ca^{2+}-free(binding) state of EF_k.

$\mathcal{B}_k$ $(k=1,2,3,4)$ is a set of pairs $\langle i,j\rangle$, where at least either i or j is the Ca^{2+}-binding site of EF_k [57]. We set

the parameters of interactions around the Ca^{2+}-binding sites as follows: When the pair $\langle i,j\rangle$ is included in $\mathcal{B}_k$, but not in $\mathcal{B}_l$ with $l \neq k$, $\bar{\gamma}^{\sigma}_{\alpha i,j}$ depends solely on α_k, $\bar{\gamma}^{\sigma}_{\alpha i,j} = \bar{\gamma}^{\sigma}_{\alpha_k i,j}$: For the state of $\alpha_k = 0$, $\bar{\gamma}^{\rm A}_{\alpha_k i,j} = 0$, $\bar{\gamma}^{\rm I}_{\alpha_k i,j} = \bar{\gamma}^{\rm N}_{\alpha_k i,j} = 1$, and for the state $\alpha_k = 1$, $\bar{\gamma}^{\rm A}_{\alpha_k i,j} = \bar{\gamma}^{\rm I}_{\alpha_k i,j} = \bar{\gamma}^{\rm N}_{\alpha_k i,j} = 1$. When the pair $\langle i,j\rangle$ is included in both of $\mathcal{B}_k$ and $\mathcal{B}_l$ ($k \neq l$), for the state $\alpha_k = 0$ and/or $\alpha_l = 0$, $\bar{\gamma}^{\rm A}_{\alpha i,j} = 0$, $\bar{\gamma}^{\rm I}_{\alpha i,j} = \bar{\gamma}^{\rm N}_{\alpha i,j} = 1$, and for the state $\alpha_k = \alpha_l = 1$, $\bar{\gamma}^{\rm A}_{\alpha i,j} = \bar{\gamma}^{\rm I}_{\alpha i,j} = \bar{\gamma}^{\rm N}_{\alpha i,j} = 1$. We use $\varepsilon^{\rm A}_{L\alpha i} = \varepsilon^{\rm I}_{L\alpha i} = \varepsilon^{\rm N}_{L\alpha i} = 0$ for all states of α.

3. CheY

Asp57 is the allosteric site $\mathcal{L}$ and modulated pairs $\mathcal{B}$ are a set of $\langle i,j\rangle$ in which i or j is Asp57 or the residue that contacts Asp57 (we consider the sequence of $2 \leq i,j \leq 129$). For such modulated pair $\langle i,j\rangle$, we set the parameters as: $\bar{\gamma}^{\rm A}_{0i,j} = 0, \bar{\gamma}^{\rm I}_{0i,j} = \bar{\gamma}^{\rm N}_{0i,j} = 1$ in the dephosphorylated state $\alpha = 0$, and $\bar{\gamma}^{\rm A}_{1i,j} = \bar{\gamma}^{\rm N}_{1i,j} = 1, \bar{\gamma}^{\rm I}_{1i,j} = 0$ in the phosphorylated state $\alpha = 1$. We use $\varepsilon^{\rm A}_{L\alpha i} = \varepsilon^{\rm I}_{L\alpha i} = \varepsilon^{\rm N}_{L\alpha i} = 0$ for $\alpha = 0$ or 1.

[1] N. M. Goodey and S. J. Benkovic, Nat. Chem. Biol. **4** , 474 (2008).
[2] J. F. Swain and L. M. Gierasch, Curr. Opin. Struct. Biol. **16**, 102 (2006).
[3] D. Kern and E. R. P. Zuiderweg, Curr. Opin. Struct. Biol. **13**, 748 (2003).
[4] K. Gunasekaran, B. Y. Ma, and R. Nussinov, Proteins **57**, 433 (2004).
[5] K. A. Henzler-Wildman and D. Kern, Nature **450**, 964 (2007).
[6] A. del Sol, C. J. Tsai, B. Ma, and R. Nussinov, Structure **17**, 1042 (2009).
[7] J. Vreede, J. Juraszek, and P. G.Bolhuis, Proc Natl Acad Sci USA **107**, 2397 (2010).
[8] A. K. Gardino, J. Villali, A. Kivenson, M. Lei, C. F. Liu, P. Steindel, E. Z. Eisenmesser, W. Labeikovsky, M. Wolf-Watz, M. W. Clarkson, and D. Kern, Cell **139**, 1109 (2009).
[9] Lei M, Velos J, Gardino A, Kivenson A, Karplus M, Kern D J. Mol. Biol. **392**, 823 (2009).
[10] K. Arora, and C. L. Brooks, III, Proc. Natl. Acad. Sci. U.S.A. **104**, 18496 (2007).
[11] G. Hummer, Proc. Natl. Acad. Sci. USA **107**, 2381 (2010).
[12] M. Ikeguchi, J. Ueno, M. Sato, and A. Kidera, Phys. Rev. Lett. **94**, 078102 (2005).
[13] W. Zheng, B. R. Brooks, and D. Thirumalai, Proc. Natl. Acad. Sci. USA **103**, 7664 (2006).
[14] F. Tama and C. L. Brooks, III, J. Mol. Biol. **318**, 733 (2002).
[15] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[16] O. Miyashita, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 12570 (2003).
[17] P. C. Whitford, J. N. Onuchic, and P. G. Wolynes, HFSP Journal **2**, 61 (2008).
[18] K. Itoh and M. Sasai, Proc. Natl. Acad. Sci. U.S.A. **107**, 7775 (2010).
[19] H. Wako and N. Saito, J. Phys. Soc. Jpn. **44**, 1931(1978).
[20] H. Wako and N. Saito, J. Phys. Soc. Jpn. **44**, 1939 (1978).
[21] N. Go and H. Abe, Biopolymers **20**, 991 (1981).
[22] V. Munõz and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. **96**, 11311 (1999).
[23] P. Bruscolini and A. Pelizzola, Phys. Rev. Lett. **88**, 258101 (2002).
[24] K. Itoh and M. Sasai, Proc. Natl. Acad. Sci. U.S.A. **101**, 14736 (2004).
[25] E. R. Henry and W. A. Eaton, Chem. Phys. **307**, 163 (2004).
[26] M. Zamparo and A. Pelizzola, Phys. Rev. Lett. **97**, 068106 (2006).
[27] K. Itoh and M. Sasai, Proc. Natl. Acad. Sci. U.S.A. **103**, 7298 (2006).
[28] K. Itoh and M. Sasai, Proc. Natl. Acad. Sci. U.S.A. **105**, 13865 (2008).
[29] K. Itoh and M. Sasai, J. Chem. Phys, **130**, 145104 (2009).
[30] B. F. Volkman, D. Lipson, D. E. Wemmer, and D. Kern, Science **291**, 2429 (2001).
[31] C. Hyeon, P. A. Jennings, J. A. Adams, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A.**106**, 3023 (2009).
[32] S. Wu, P. I. Zhuravlev, G. A. Papoian, Biophys J **95**,5524 (2008).
[33] C. J. Tsai, A. del Sol, and R. Nussinov, J. Mol. Biol. **378**, 1 (2008).
[34] A. E. Karnoub and R. A. Weinberg, Nat. Rev. Mol. Cell Biol. **9**,517 (2008).
[35] A. A. Gorfe, B. J. Grant, and J. A. McCammon, Structure **16**, 885 (2008).
[36] K. D. Corbett and T. Alber, Trends in Biochem. Sci. **26**, 710 (2001).
[37] Y. Ito et al. Biochemistry **36**, 9109 (1997).
[38] Y. Arai et al., Biochem. Biophys. Res. Comm. **343**, 809 (2006).
[39] C. O'Connor and E. L. Kovrigin, Biochemistry **47**,10244 (2008).
[40] A. P. Loh, W. Guo, L. K. Nicholson, and R. E. Oswald, Biochemistry **38**, 12547 (1999).
[41] A. Crivici and M. Ikura, Annu. Rev. Biophys. Biomol. Struct. **24**, 85, (1995).
[42] H. Kuboniwa et al., Nat. Struct. Biol. **2**, 768 (1995).
[43] S. Tripathi and J. J. Portman, J. Chem. Phys. **128**, 205104 (2008).
[44] S. Tripathi and J. J. Portman, Proc. Natl. Acad. Sci. U.S.A. **106**, 2104 (2009).
[45] A. Malmendal et al., J. Mol. Biol. **293**, 883 (1999).
[46] J. J. Chou et al., Nat. Str. Biol. **8**, 990 (2001).
[47] G. Barbato et al., Biochemistry **31**, 5269 (1992).
[48] R. Barak and M. Eisenbach. Biochemistry **31**, 1821 (1992).
[49] A. M. Stock, V. L. Robinson, and P. N. Goudreau, Annu. Rev. Biochem. **69**, 183 (2000)
[50] F. J. Moy et al., Biochemistry **33**, 10731 (1994).
[51] M. Simonovic and K. Volz, J. Biol. Chem. **276**, 28637 (2001).
[52] H. S. Cho et al., J. Mol. Biol. **297**, 543 (2000).

[53] M. H. Knaggs et al., Biophy. J. **92**, 2062 (2007).
[54] L. Ma and Q. Cui, J. Am. Chem. Soc.**129**, 10261 (2007).
[55] M. Yazawa, E. Kawamura, O. Minowa, K. Yagi, M. Ikura, and K. Hikichi, J. Biochem. **95**, 443 (1983).
[56] J. Evenäs, A. Malmendal, E. Thulin, G. Carlström, and S. Forsén, Biochemistry **37**, 13744 (1998).
[57] We use the data in *PDBsum* Web site (URL: `http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/`) to define residues which bind GDP/GTP in Ras and Ca ions in calmodulin.