TOPICAL REVIEW — Modeling and simulations for the structures and functions of proteins and nucleic acids

Find slow dynamic modes via analyzing molecular dynamics simulation trajectories*

Chuanbiao Zhang(张传彪)[1] and Xin Zhou(周昕)[2,†]

[1] *College of Physics and Electronic Engineering, Heze University, Heze* 274015*, China*
[2] *School of Physical Sciences, University of Chinese Academy of Sciences, Beijing* 100190*, China*

(Received 18 June 2020; revised manuscript received 27 July 2020; accepted manuscript online 7 August 2020)

It is a central issue to find the slow dynamic modes of biological macromolecules via analyzing the large-scale data of molecular dynamics simulation (MD). While the MD data are high-dimensional time-successive series involving all-atomic details and sub-picosecond time resolution, a few collective variables which characterizing the motions in longer than nanoseconds are needed to be chosen for an intuitive understanding of the dynamics of the system. The trajectory map (TM) was presented in our previous works to provide an efficient method to find the low-dimensional slow dynamic collective-motion modes from high-dimensional time series. In this paper, we present a more straight understanding about the principle of TM via the slow-mode linear space of the conformational probability distribution functions of MD trajectories and more clearly discuss the relation between the TM and the current other similar methods in finding slow modes.

Keywords: molecular dynamics simulation, slow modes, trajectory map

PACS: 87.15.A–, 02.70.Ns, 82.20.Wt **DOI:** 10.1088/1674-1056/abad24

1. Introduction

Molecular dynamics (MD) simulation has been widely used in recent a few decades in very various fields, such as detecting the relationship between chemical structures and functions of biological macromolecules.[1,2] The raw data of MD simulations are high-dimensional time series with the total time length usually in the order of microsecond or even longer and containing the femtosecond resolution and all-atomic coordinates and velocities. The data are too complicated to provide an intuitive understanding of the simulated system; thus we usually need to find much less (collective) variables to characterize the motion of the system in a larger timescale to understand its main features.[3] For simpler systems, there may exist some direct ways to select some collective variables to (approximately) represent the main features via our experiences or intuitions. For example, for small proteins or peptides, the dihedral angles in the backbone can describe well the large (and usually slow) motions of molecules. Thus they can be used as the key collective variables. However, in more complex systems, it is usually needed to develop more systematical data analysis methods to capture the main features from big MD data.

Many methods were presented to simplify large data to take out the main features. For example, lots of clustering algorithms[4–6] and reduction dimensionality techniques[7–11] focused on finding the low-dimensional structure of data point set, and were applied in the analysis of MD simulation data of biomolecules.[12–23] The methods are based on the geometric distances between data points (conformations of molecules) can get the static structure of the whole data set without applying any dynamics information. A recent technique, the diffusion map, constructed an artificial diffusion dynamics among the static data points based on the distances then tried to find the slow modes of the diffusion dynamics, which achieved the static structure robustly under the help of the artificial dynamics information.[11] On the other hand, some methods, such as the Markov states model (MSM),[24–31] and its improvement, tICA,[32,33] the trajectory map (TM),[34–40] applied the dynamics information of MD trajectories to find the slow-dynamic structure of data. The TM gives a simple and efficient way to take into account the time successiveness along the MD trajectory to construct the structure of slow dynamics robustly. It promises a general method in analyzing various complicated time series. The implementation and application of TM were presented in previous works.[34,37–40] Here we revisit the mathematical principle behind the TM and the relationship between the TM and other related techniques; thus we can more clearly discuss the advantages and disadvantages of these methods.

2. Theory and method

2.1. Slow dynamics modes

Generally, let us consider a stochastic process, as the molecular dynamics (MD) simulation, *e.g.*, following the

*Project supported by the National Natural Science Foundation of China (Grant No. 11904086).

†Corresponding author. E-mail: xzhou@ucas.ac.cn

(overdamped) Langevin equation

$$\gamma \mathrm{d}\boldsymbol{q} = -\frac{\partial U}{\partial \boldsymbol{q}}\mathrm{d}t + \sqrt{2k_{\mathrm{B}}T\gamma \mathrm{d}t}\,\mathrm{d}\boldsymbol{G}, \tag{1}$$

where γ is the friction coefficient, q is a simple denotation of the (high dimensional) conformational coordinate, $U(q)$ is the potential energy surface, k_{B} is the Boltzmann constant, T is the temperature, and $\mathrm{d}\boldsymbol{G}$ is the normal Gaussian white noise. Equivalently, we can described the process by the Fokker–Planck equation based on the probability density function,

$$\frac{\partial}{\partial t}P(q,t) = \boldsymbol{L}_{FP}P(q,t), \tag{2}$$

with the operator

$$\boldsymbol{L}_{\mathrm{FP}} = \frac{\partial}{\partial \boldsymbol{q}} \cdot \left[\gamma^{-1}\left(\frac{\partial U}{\partial \boldsymbol{q}}\right) + (k_{\mathrm{B}}T)\,\gamma^{-1}\frac{\partial}{\partial \boldsymbol{q}}\right].$$

Simply, we set γ to be a constant, (the unit if resetting time $\mathrm{d}\tau = \mathrm{d}t/\gamma$). The equilibrium solution is the Boltzmann distribution, $P_{\mathrm{eq}}(q) = (1/Z)\,\mathrm{e}^{-\beta U(q)}$. Here $Z = \int \mathrm{d}q\mathrm{e}^{-\beta U(q)}$, named as the partition function, is the center quality in statistical physics, and $\beta = 1/k_{\mathrm{B}}T$. In this paper, we will set $k_{\mathrm{B}}T$ as unit without explicit mention again.

Defining $P(q,t) = [P_{\mathrm{eq}}(q)]^{1/2}\Psi(q,t)$, we have

$$-\frac{\partial}{\partial t}\Psi(q,t) = \boldsymbol{H}\Psi(q,t), \tag{3}$$

the imaginary time Schödinger equation, where $\boldsymbol{H} = -\partial^2/\partial q^2 + V(q)$ is a real symmetric operator, the same as the Hamilton operator in quantum mechanism. The effective potential

$$V(q) = \frac{1}{4}\left[\frac{\partial}{\partial q}U(q)\right]^2 - \frac{1}{2}\frac{\partial^2}{\partial q^2}U(q).$$

Therefore, we have the spectrum expansion

$$P(q,t) = P_{\mathrm{eq}}(q)\sum_{n=0}^{\infty} a_n(0)\,\mathrm{e}^{-\lambda_n t}\hat{\Phi}_n(q), \tag{4}$$

where $\hat{\Phi}_n(q) = \Phi_n(q)/\Phi_0(q)$, and $\Phi_n(q)$ is the eigenfunction of the operator $\boldsymbol{H}$ with the eigenvalue λ_n, *i.e.*, $\boldsymbol{H}\Phi_n(q) = \lambda_n\Phi_n(q)$. We have $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_n < \cdots$, and $\Phi_0(q) = [P_{\mathrm{eq}}(q)]^{1/2}$. The orthogonal condition $\langle\hat{\Phi}_n(q)\hat{\Phi}_m(q)\rangle_{\mathrm{eq}} = \int \Phi_n(q)\Phi_m(q)\,\mathrm{d}q = \delta_{nm}$, and the completeness $P_{\mathrm{eq}}(q)\sum_{n=0}^{\infty}\hat{\Phi}_n(q)\hat{\Phi}_n(q') = \delta(q-q')$. Here $\langle\cdot\rangle_{\mathrm{eq}}$ means the expectation under the distribution $P_{\mathrm{eq}}(q)$. The expansion coefficient $a_n(0) = \langle\hat{\Phi}_n(q)\rangle_0 = \int \mathrm{d}q\hat{\Phi}_n(q)P(q,0)$, the expectation under the initial distribution $P(q,t=0)$.

Due to the fast decay of motion modes with large eigenvalues as time, we can consider only the equilibrium mode $\Phi_0(q)$ and the first N eigenfunctions, $\{\Phi_1(q),\ldots,\Phi_N(q)\}$, named as slow modes, to describe the main features of the dynamics of system at long time scale. For example, if $U(q) = \frac{1}{2}kq^2$ in one dimension space, we have $V(q) = \frac{1}{4}k^2q^2 - k/2$. It is easy to know the eigenvalues and eigenfunctions of the operator $\hat{H}$ from that of the quantum harmonic oscillator, $\lambda_n = nk$, and the $\hat{\Phi}_n(q)$ is the n-order Hermite polynomial. In this case, $\Phi_0(q)$ gives the equilibrium information, and $\lambda_1 = k$ corresponds to the time scale to reach the equilibrium.

2.2. Linear space of slow modes and metastable states

For any probability function $P(q,t)$, we can suppose it already evolved a not-short time, thus

$$\frac{P(q,t)}{P_{\mathrm{eq}}(q)} \approx 1 + \sum_{n=1}^{N} a_n(t)\hat{\Phi}_n(q), \tag{5}$$

i.e., it approximately belongs to the linear space spanned by the slow modes. For example, considering an MD trajectory $q(t), t \in [0,\tau]$, without losing generality, the corresponding probability function $P(q) = (1/\tau)\int_0^{\tau}\mathrm{d}t\delta(q-q(t))$ is a linear combination of the slow modes (if not, reset a later time as $t = 0$ and discarding the earlier conformations). Therefore, we can generate lots of MD trajectories from different initial conformations, the corresponding probability functions (or more exactly, the finite-size samples), denoted as $\{P_i(q)\}$, $i = 1,\ldots,m$, can be applied to linearly combine to span the slow dynamic modes of the system

$$\hat{\Phi}_n(q) = \frac{\sum_i b_{n,i}P_i(q)}{P_{\mathrm{eq}}(q)}, \tag{6}$$

where a linear uncorrelated subset of $\{P_i(q)\}$ is sufficient to expand any slow mode.

Usually, we can split the whole conformational space into some conformational regions which are basins or super-basins of the potential energy surface and separated by high energy barriers; each is named as a metastable state in slow dynamics. In other words, each metastable state is a conformational region where the system often reaches local equilibrium inside before leaving out. The slow modes correspond to transitions among metastable states, and the inner local equilibrium within each (small) metastable state is corresponds to faster modes. Therefore, all the slow-mode functions $\{\hat{\Phi}_n(q)\}$ provide a splitting of the whole conformational space into metastable states, which are approximately constants inside each state but vary obviously only at boundaries of metastable states. We have

$$\hat{\Phi}_n(q) = \sum_{\alpha} c_{n\alpha}\Theta_{\alpha}(q). \tag{7}$$

Here $\Theta_{\alpha}(q)$ is the characteristic function of the α metastable state, whose value is unit inside the state but zero otherwise. It means that conformations inside the same state are equal in the viewpoint of slow dynamics. As involving more (shorter) dynamic modes, more metastable states are split and more refined description about the conformational space are obtained.

MD trajectories similarly give the splitting of metastable states, since

$$\frac{P_i(q)}{P_{\text{eq}}(q)} = \sum_{\alpha} c'_{i\alpha} \Theta_\alpha(q), \tag{8}$$

i.e., the local equilibrium distribution is proportional to the global one inside each metastable state. It means that there is the same linear structure among slow-mode functions, the characteristic functions of metastable states, and the conformational probability functions from MD trajectories. The central idea of the trajectory map (TM)[34,38] is to achieve slow modes or metastable states from MD trajectories.

2.3. Trajectory map

We apply a set of known conformational functions, denoted as $\{A^\mu(q)\}$, $\mu = 1, 2, \ldots, N_0$, as basis functions to coarsely represent the linear space spanned by the slow-mode functions,

$$\frac{P_i(q)}{P_{\text{eq}}(q)} \approx \sum_{\mu} p_{i,\mu} A^\mu(q). \tag{9}$$

Here, the approximation which comes from the incompleteness of the set of basis functions less affects the construction of the slow-mode space. The coefficient $p_{i,\mu} = \langle A_\mu(q) \rangle_i$ is estimated from the finite-size conformational sample corresponded to $P_i(q)$ (a segment of MD trajectory length of τ), where the conjugated basis function $A_\mu(q)$ satisfying

$$\langle A_\mu(q) A^\nu(q) \rangle_{\text{eq}} = \delta_{\mu\nu}, \tag{10}$$

is estimated in the conformational sample corresponded to $P_{\text{eq}}(q)$. Usually, the equilibrium sample is absence, we can apply a reference distribution $P_{\text{ref}}(q)$, for example, the mean of all MD trajectories, to replace $P_{\text{eq}}(q)$, which less affects the construction of linear space of slow modes, since $P_{\text{ref}}(q)$ is also proportional to $P_{\text{eq}}(q)$ in each metastable state, thus $P_i(q)/P_{\text{ref}}(q)$ still belongs to the slow-mode linear space.

In this paper, we do not distinguish the conformation sample of an MD trajectory and its corresponding probability function $P_i(q)$, unless avoiding some possible confusions. In addition, for simplification, we linearly combine these selected physical variables $A^\nu(q)$ to get a set of orthonormalize basis functions $\{\hat{A}^\mu(q) = \hat{A}_\mu(q)\}$. Thus, except $\hat{A}^0(q) \equiv 1$, we have $\langle \hat{A}^\mu(q) \rangle_{\text{ref}} = 0$. We use the Einstein summation convention of repeat subscript and superscript indexes below without explicit mentioning.

Thus, each MD trajectory or segment $P_i(q)$ is mapped as a vector

$$\boldsymbol{p}_i = p_{i,\mu} \hat{A}^\mu(q), \tag{11}$$

and all these vectors $\{\boldsymbol{p}_i\}$ span the slow-mode linear space (more exactly, its projection in the applied basis functions $\{\hat{A}^\mu(q)\}$). A larger set of basis functionss is helpful for providing more complete information about slow modes, but even when the basis functions are not sufficient so that some of the slow modes are not able to distinguish completely, the linear projection does not bring any additional biased results. In practice, the size of sample corresponded to $P_{\text{ref}}(q)$ also limits the number of linear uncorrelated basis functions $\{\hat{A}^\mu(q)\}$. More discussion about the basis functions were shown in the previous literatures about TM.[34,37–40]

We apply the principle component analysis (PCA) to the trajectory-mapped vectors $\{\boldsymbol{p}_i\}$ to get the first a few principle components (PCs), denoted as $\{\hat{B}_\alpha(q)\}, \alpha = 1, \ldots, m$, as a set of orthonormalized basis functions of the linear space of slow dynamics modes,

$$\hat{B}_\alpha(q) = b_{\alpha\mu} \hat{A}^\mu(q), \tag{12}$$

here $b_{\alpha\mu}$ is the α-th eigenvector of the variance–covariance matrix of these mapped points, $\sum_i p_{i\mu} p_{i\nu}$. We have $\langle \hat{B}_\alpha(q) \hat{B}_\beta(q) \rangle_{\text{ref}} = \delta_{\alpha\beta}$. A hint for choosing the number of principal components is provided by the plot of eigenvalues sorted in decreasing order. The first few eigenvalues which are significantly greater than zero are usually correspond to slow processes.

In the low-dimensional space of $\{\hat{B}_\alpha(q)\}$, many common analyzing or visualizing techniques can be applied to achieve the slow dynamics modes or identify metastable states and transition among states. For example, we can project the original MD trajectory $q(t)$ to get a smoothed slow-dynamic-dominate trajectory $\boldsymbol{B}(t)$ with the components

$$\hat{B}_\alpha(t) = \frac{1}{\Delta t} \int_t^{t+\Delta t} \hat{B}_\alpha(q(t'))\, \mathrm{d}t'. \tag{13}$$

Here we applied a time-window smoothing (with length Δt) to further filter fast dynamics modes of MD trajectory. Usually, Δt is set about two to three orders of magnitude smaller than the length of trajectory segment (τ), and the results of TM is insensitive to the specific value of Δt. Then we can identify the obvious change along the trajectories $\boldsymbol{B}(t)$ as transition events between metastable states, or calculate the two-point similar matrix in the B space, $C(t_2, t_1) = \boldsymbol{B}(t_2) \cdot \boldsymbol{B}(t_1)$ to get the transition events.[38]

3. Application

In this section, we use the Trp-cage protein to illustrate the basic application of the TM. The Trp-cage contains 20 residues, includes three secondary structures in its native structure: an α helix, a 3–10 helix, and a polyproline II segment.[41–43] This protein can fold in microseconds, and the stability of native structure originates from the hydrophobic core around the Trp residue.[43,44] Due to its small size and various meta-stable states, Trp-cage becomes an ideal protein for testing both sampling algorithms and force fields of simulations, thus it has been extensively studied by MD and experiments.[45–54] However, its folding kinetics and folding pathways are still not fully understood.[55] Recently, Lindorff–Larsen *et al.* have performed a 208-μs equilibrium MD simulation of the Trp-cage at 290 K.[43] They applied the CHARMM22* force field[56] and the modified TIP3P water model compatible with the CHARMM force field.[57,58] The generated MD trajectory involves about 2.08×10^5 snapshots, with a time interval of 200 ps. We downloaded the trajectory file from D. E. Shaw Research.

We choose the dihedral angles of protein backbone, ϕ_i and ψ_i with $i = 1,\ldots,18$, as collective variables to describe the large-scale motions. Due to the periodicity of dihedral angles, we transform the angles into their cosine and sine functions, to get 72 basis functions in total,[59] $\{A^\mu(q)\}$, $\mu = 1,\ldots,72$ in the TM. The time evolution of the root-mean-square deviation (RMSD) clearly distinguishes the folded state (native structure) and unfolded structure, but more details inside the unfolded structure is not so clear (Fig. 1(a)).

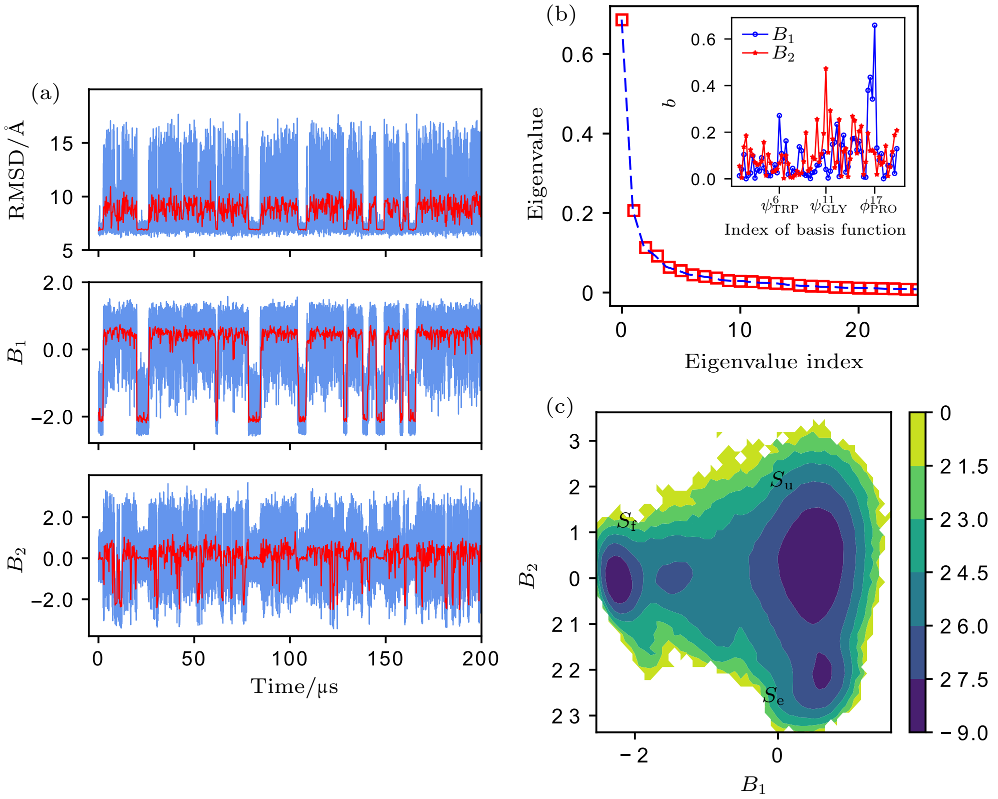

Fig. 1. (a) Time evolution of the RMSD of the Trp-cage to its native structure, the slow variables B_1 and B_2 obtained in the TM. Red line is time-window-smoothed one ($\Delta t = 200$ ns). (b) Eigenvalue of the variance–covariance matrix of the trajectory-mapped points. The inset is the contribution of each basis function to the slow variables B_1 and B_2. (c) The free-energy landscape (in units of $k_{\rm B}T$) in the slow-variable space (B_1, B_2).

We apply the TM with the free parameter $\tau = 2$ μs to detect the slow modes in the τ time scale, which hiding in the total 208-μs-long all-atomic MD trajectory. As shown in Fig. 1(b), two eigenvalues of variance-covariance matrix are significantly greater than zero, which indicates that this system involves two slow dynamics processes. These two slow variables (denoted as B_1 and B_2) are shown in Fig. 1(a). We show the free energy surface in the (B_1, B_2) space (Fig. 1(c)), $\Delta G = -k_{\rm B}T \ln P$, where P is the probability distribution of the molecular system along the slow variables. In this two-dimensional space, the entire region includes three minima that correspond to three metastable states, *i.e.*, the folded state ($S_{\rm f}$), unfolded state ($S_{\rm u}$), and extended state ($S_{\rm e}$). The B_1 primarily distinguishes state $S_{\rm f}$ from the other two states. In other words, the B_1 which combined by 72 angle-based basis functions has similar physical meaning with the RMSD which is common used to identify folded state of protein. The B_2 distinguishes state $S_{\rm e}$ from the other states.

Figure 2(a) shows the time-ordered similarity matrix of the MD trajectory after projecting in the slow dynamics space, $C(t_2,t_1) = \boldsymbol{B}(t_2) \cdot \boldsymbol{B}(t_1)$. The matrix is found to divide into some blocks. Inside each block (each time segment), the element of similarity matrix is almost equal to unity, it indicates that all the conformations in each time segment are almost identical in the $\boldsymbol{B}$ space, *i.e.*, they belong to the same metastable state without occuring transition in the time segment. At the time points which separated different blocks, the MD trajectory occurs transitions from one to another metastable states, thus the similarity of conformations in the B space is obvious. Thus, this matrix can help us to identify transition events.

From the similarity matrix, the PCCA+ algorithm[23] is implied to cluster the similar segments together. As shown in Fig. 2(b), the rearranged similarity matrix clearly show three blocks which correspond to three metastable states. A transition network was constructed from the similarity matrix (Fig. 2(c)). The folding path was $S_{\mathrm{e}} \longrightarrow S_{\mathrm{u}} \longrightarrow S_{\mathrm{f}}$, without the direct transition between S_{e} and the nature structure S_{f}. The typical protein structure of each state (Fig. 2(c)) and the components of slow variables help us to understand the main distinction between each state. As mentioned in Eq. (12), these variables are actually linear combinations of basis functions. The combination coefficient $b_{\alpha\mu}$ is illustrated in Fig. 1(b) which can used to obtain the understanding about the physical meaning of slow variables. The main components of B_1 is ψ^{6}_{TRP} (ψ angle of TRP6) and ϕ^{17}_{PRO}. Those two residues correspond to the hydrophobic core. The folding/unfolding process of the protein is closely related to the formation of hydrophobic core. The main components of B_2 is ψ^{17}_{GLY}, which correspond to the middle part of the protein. The twist of ψ^{17}_{GLY} make the structure of state S_{e} looser than S_{u} (Fig. 2(c)). The three-state model constructed by TM provides us with more details about the folding dynamics of the Trp-cage protein.

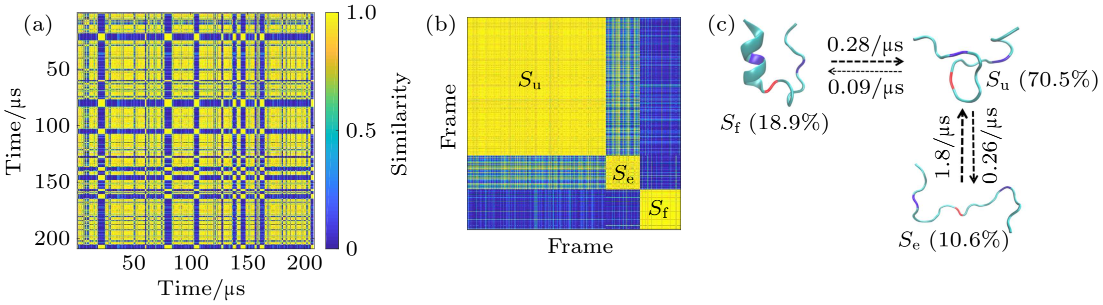

Fig. 2. Time-ordered similarity matrix of the MD trajectory. The similarity between two samples $C(t_2,t_1) = B(t_2)\cdot B(t_1)$. (b) The time-rearranged similarity matrix, suggesting three metastable states. (c) Kinetic transition network. Numbers near the arrows are the corresponding transition rates. The population of each state in the 208-μs MD trajectory is listed in bracket (which approaches to the equilibrium one, in consistent with the fact the folding and unfolding transitions occur more than ten times during the MD simulation). Residue TRP6 and PRO17 are shown in blue, GLY11 in red.

4. Discussion

The propagator, $G(q,t|q_0,0)$, also named as Green function, is the condition probability of the system to be q at time t while it was q_0 at t_0, which describes the dynamics of system. For large t, we have

$$G(q,t|q_0,0) = P_{\mathrm{eq}}(q)\left[1+\sum_{n=1}^{N} \mathrm{e}^{-\lambda_n t}\hat{\Phi}_n(q)\hat{\Phi}_n(q_0)\right], \quad (14)$$

and $P(q,t) = \int G(q,t|q_0,0)P(q_0,0)\,\mathrm{d}q_0$.

It is easy to know that the slow-mode functions are the first a few eigenfunctions of $G(q,t|q_0,0)$. In principle, we can represent the propagator as a time correlation matrix by using the basis functions $\{A^{\mu}(q)\}$,

$$\begin{aligned} G^{\mu}_{\nu}(t,0) &= \int \mathrm{d}q\mathrm{d}q' A^{\mu}(q)G(q,t|q',0)A_{\nu}(q')P(q',0) \\ &\equiv \langle A^{\mu}(q(t))A_{\nu}(q(0))\rangle_0, \end{aligned} \quad (15)$$

then diagonalize the matrix to get the slow-mode functions.

For well constructing the propagator to get the slow dynamics modes, it is natural to use a great number of (thousands or much larger) cell functions $\theta_{\mu}(q)$ as an approximately complete set of basis functions. Here $\{\theta_{\mu}(q)\}$ is the character function of the cells, whose value is unit while q is in the μ cell, otherwise zero. It means that $\{\theta_{\mu}(q)\}$ splits whole the conformational space into lots of small cells, then any conformational function can use these cell function to expand by neglecting the fast-spatial varying of function inside each of cell (In practice, we can only split the sample of all conformations and group the sample into many clusters according to their neighboring, then $\{\theta_{\mu}(q)\}$ which are the character functions of these clusters, describe the belonging of any sampled conformation q). These cells are required to so small that we can suppose the local equilibrium inside each cell is easy to reach in a short time. Thus the cells provide a coarse description of conformations, we do not distinguish conformations inside the same cell and think they are identity in slow-dynamics viewpoint. Thus, $\{\theta_{\mu}(q)\}$ is approximately complete in describing the slow dynamics modes. We can apply $\{\theta_{\mu}(q)\}$ to represent the detailed varying of slow-mode functions $\hat{\Phi}(q)$ in whole the conformational space (in whole the finite-size set of all sampled conformations in practice). In the cases, $A^{\mu}(q) = \theta_{\mu}(q)$, and $A_{\mu}(q) = (1/\hat{Z}_{\mu})\theta_{\mu}(q)$, where $\hat{Z}_{\mu} = \langle\theta_{\mu}(q)\rangle_{\mathrm{eq}}$ is the equilibrium probability inside the μ cell. The propagator is the transition probability matrix among these cells,

$$\begin{aligned} G^{\mu}_{\nu}(t) &= \frac{1}{\hat{Z}_{\mu}}\langle\theta_{\mu}(q(t))\theta_{\nu}(q(0))\rangle_{\mathrm{eq}} \\ &= \frac{n_{\mu\nu}(t,0)}{n_{\nu}(0)}, \end{aligned} \quad (16)$$

provides a very detailed representation of the propagator. Here $n_{\mu\nu}(t,0)$ is the probability that q locates in the μ cell at t and starts from the ν cell at $t=0$, and $n_{\nu}(0)$ is the probability in the ν cell at $t=0$. It is worth to mention, due to the character of the cell functions, we actually can use any initial probability $P(q,0)$ to calculate the matrix without altering results. It is easy to know that $G^{\mu}_{\nu}(t) \geq 0$, and $\sum_{\mu} G^{\mu}_{\nu}(t) \equiv 1$. Thus, the eigenvalues of the matrix are between 0 and 1 for any t, and the largest a few eigenvalues and eigenvectors give the slow modes of motion.

It is actually the main idea of MSM which was widely applied for analyzing metastable states of slow motions in biological molecular systems from ensemble dynamics simulations.[24–31] In MSM, the cells are defined in the sample set of sampled conformations. All sampled conformations are clustered into thousands of groups based on the pair distance of conformations, then each group of conformations corresponds into a cell function. Along the MD trajectories, the transition probability from one cell to another after a time interval t is estimated as the corresponding matrix element. Due to the completeness of basis functions and not requiring the initial distribution be equilibrium, MSM provides a complete construction of slow-dynamics modes in principle, and its eigenfunctions are one-to-one corresponding to the slow dynamics modes. However, in practice, very large data set of simulations is often required to get good estimate of each matrix element with sufficient statistical accuracy, which requiring sufficient events between each pair of cells.

Rather than applying a great number of cell functions as the basis functions, a recent improvement of the original MSM, named as tICA, was presented to identify slow dynamics by constructing the time correlation matrix of some physical variables, $G_{\mu\nu}(t,0) = \langle \hat{A}_{\mu}(q(t))\hat{A}_{\nu}(q(0))\rangle$. Here $\{\hat{A}^{\mu}(q)\}$ is same as the basis functions applied in the TM. As we already mentioned, the time correlation matrix is a finite-dimensional approximate representation of the dynamics propagator. Thus, in principle, the tICA gives slow dynamical modes by diagonalizing the time correlation matrix to achieving its first a few slowest eigen-modes, while sufficient basis functions are applied and the initial distribution $P(q,0)$ already reached the equilibrium one. Anyway, in practice, usually not too many basis functions can be applied, and the initial distribution $P(q,0)$ may be deviated obviously from $P_{\mathrm{eq}}(q)$, thus the calculated time correlation matrix may loss the character of the original propagator more or less. For example, since the time correlation matrix is not symmetric, some of its eigenvalues may be not real, thus do not correspond to the rates of dynamical modes then we cannot directly get slow modes from the eigenvalues directly.

As a comparison, TM calculates and diagonalizes the variance-covariance matrix of trajectory-mapped points,

$$\Sigma^{\mu\nu} = \frac{1}{N_0}\sum_{i=0}^{N_0}\langle \hat{A}_{\mu}(q)\rangle_i \langle \hat{A}_{\nu}(q)\rangle_i.$$

It is easy to know, $\Sigma^{\mu\nu}$ is symmetric and a kind of average of $G_{\mu\nu}(t_2,t_1)$,

$$\Sigma^{\mu\nu} = \frac{1}{\tau}\int_0^{\tau} \mathrm{d}t \left(1-\frac{t}{\tau}\right)[C_{\mu\nu}(t)+C_{\nu\mu}(t)], \tag{17}$$

where

$$C_{\mu\nu}(t) = \frac{1}{\tau - t}\int_0^{\tau - t} \mathrm{d}t_1 G_{\mu\nu}(t+t_1,t_1). \tag{18}$$

The TM focuses more on the difference of MD trajectories $\{P_i(q)\}$, but less on that of conformations in the same trajectory, since the latter is mainly related to the short-time correlation. Therefore, the average of the time correlation matrix in the TM provides a suitable way to more efficiently extract the slow dynamics modes by filtering fast motions.

5. Summary

The TM can extract the slow dynamic processes from time-series data. The key of TM is to make use of the time continuity between conformations, rather than only based on the geometric similarity of single conformation to build the slow variables of the system. Compared with the other methods, the TM is to apply the probability function of trajectories to combine the slow-dynamics functions directly linearly. It makes the TM robust and straightforward, less affecting by the incompleteness of basis functions in representing these slow dynamics functions. Besides, the TM gives slow-dynamics related analyzed collective variables, which not only provides a simple understanding of the slow dynamics of systems from MD trajectories but also is applied to extend the time scale of further MD simulations[39] by combining with some enhanced sampling techniques, such as metadynamics,[60] umbrella sampling,[61] and forward flux methods,[62] *etc*.

References

[1] Piana S, Lindorff-Larsen K and Shaw D E 2012 *Proc. Natl. Acad. Sci. USA* **109** 17845

[2] Lyulin S V, Gurtovenko A A, Larin S V, Nazarychev V M and Lyulin A V 2013 *Macromolecules* **46** 6357

[3] Lane T J, Shukla D, Kyle A B and Vijay S P 2013 *Curr. Opin. Struct. Biol.* **23** 58

[4] Jain A K 2008 *Machine Learning and Knowledge Discovery* (Berlin, Heidelberg: Springer) pp. 3–4

[5] Schubert E, Sander J, Ester M, Kriegel H P and Xu X 2017 *ACM Trans. Database Syst.* **42** 19

[6] Alex R and Laio A 2014 *Science* **344** 1492

[7] Hotelling H 1933 *J. Educ. Psychol.* **24** 417

[8] Hyvrinen A and Oja E 2000 *Neural Netw.* **13** 411

[9] Schwantes C R and Pande V S 2013 *J. Chem. Theory. Comput.* **9** 2000

[10] Tenenbaum J B, de Silva V and Langford J C 2000 *Science* **290** 2319

[11] Nadler B, Lafon S, Coifman R R and Kevrekidis I G 2006 *Appl. Comput. Harmon. Anal.* **21** 113

[12] Shea J E and Brooks C L 2001 *Ann. Rev. Phys. Chem.* **52** 499

[13] Mu Y G, Nguyen P H and Stock G 2005 *Proteins* **58** 45

[14] Sims G E, Choi I G and Kim S H 2005 *Proc. Natl. Acad. Sci. USA* **102** 618

[15] Rao F and Karplus M 2010 *Proc. Natl. Acad. Sci. USA* **107** 9152

[16] Das P, Moll M, Stamati H, Kavraki L E and Clementi C 2006 *Proc. Natl. Acad. Sci. USA* **103** 9885

[17] Nadler B, Lafon S, Coifman R R and Kevrekidis I G 2006 *Appl. Comput. Harmon. Anal.* **21** 113

[18] Krivov S V and Karplus M 2004 *Proc. Natl. Acad. Sci. USA* **101** 14766

[19] Maisuradze G G, Liwo A and Scheraga H A 2009 *Phys. Rev. Lett.* **102** 238102

[20] Torda A E and Gunsteren W F 1994 *J. Comput. Chem.* **15** 1331

[21] Shao J Y, Tanner S W, Thompson N and Cheatham T E 2007 *J. Chem. Theory. Comput.* **3** 2312

[22] Deuflhard P, Huisinga W, Fischer A and Schutte C 2000 *Linear Algebra Appl.* **315** 39

[23] Deuflhard P and Weber M 2005 *Numer Linear Algebra Appl.* **398** 161

[24] Gfeller D, De Los Rios P, Caflisch A and Rao F 2007 *Proc. Natl. Acad. Sci. USA* **104** 1817

[25] Noe F, Horenko I, Schutte C and Smith J C 2007 *J. Chem. Phys.* **126** 155102

[26] Chodera J D, Singhal N, Pande V S, Dill K A and Swope W C 2007 *J. Chem. Phys.* **126** 155101

[27] Bowman G R and Pande V S 2010 *Proc. Natl. Acad. Sci. USA* **107** 10890

[28] Bowman G R, Meng L and Huang X 2013 *J. Chem. Phys.* **139** 121905

[29] Weber J K, Jack R L and Pande V S 2013 *J. Am. Chem. Soc.* **135** 5501

[30] Pande V S, Beauchamp K and Bowman G R 2010 *Methods* **52** 99

[31] Deng N J, Dai W, and Levy R M 2013 *J. Phys. Chem. B* **117** 12787

[32] Naritomi Y and Fuchigami S 2011 *J. Chem. Phys.* **134** 065101

[33] Nuske F, Keller B G, Perez-Hernandez G, Mey A S J S and Noe F 2014 *J. Chem. Theory. Comput.* **10** 1739

[34] Gong L C and Zhou X 2010 *J. Phys. Chem. B* **114** 10266

[35] Gong L C and Zhou X 2009 *Phys. Rev. E* **80** 026707

[36] Zhang C B, Li M and Zhou X 2015 *Chin. Phys. B* **24** 120202

[37] Gong L C, Zhou X and Ouyang Z C 2015 *PloS One* **10** e0125932

[38] Zhang C B, Yu J and Zhou X 2017 *J. Phys. Chem. B* **121** 4678

[39] Zhang C B, Ye F F, Li M and Zhou X 2019 *Sci. China: Phys. Mech.* **62** 067012

[40] Zhang C B, Xu S and Zhou X 2019 *Phys. Rev. E* **100** 033301

[41] Neidigh J W, Fesinmeyer R M and Andersen N H 2002 *Nat. Struct. Biol.* **9** 425

[42] Bipasha B, Lin J C, Williams V D, Kummler P, Neidigh J W and Andersen N H 2008 *Protein Eng. Des. Sel.* **21** 171

[43] Lindorff-Larsen K, Piana S, Dror R O and Shaw D E 2011 *Science* **334** 517

[44] Day R, Paschek D and Garcia A E 2010 *Proteins* **78** 1889

[45] Spiwok V, Oborsky P, Pazurikova J, Krenek A and Kralova B 2015 *J. Chem. Phys.* **142** 115101

[46] Kim S B, Dsilva C J, Kevrekidis I G and Debenedetti P G 2015 *J. Chem. Phys.* **142** 085101

[47] Andryushchenko V A and Chekmarev S F 2016 *Eur. Biophys. J.* **45** 229

[48] Zang T W, Yu L L, Zhang C and Ma J P 2014 *J. Chem. Phys.* **141** 044113

[49] Zhan L X, Chen J Z Y and Liu W K 2007 *Proteins* **66** 436

[50] Huang X H, Hagen M, Kim B, Friesner R A, Zhou R H and Berne B J 2007 *J. Phys. Chem. B* **111** 5405

[51] Pitera J W and Swope W C 2003 *Proc. Natl. Acad. Sci. USA* **100** 7587

[52] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A and Simmerling C 2006 *Proteins* **65** 712

[53] Lai Z Z, Preketes N K, Mukamel S and Wang J 2013 *J. Phys. Chem. B* **117** 4661

[54] Abaskharon R M, Culik R M, Woolley G A and Gai F 2015 *J. Phys. Chem. Lett.* **6** 521

[55] Andryushchenko V A and Chekmarev S F 2016 *Eur. Biophys. J.* **45** 229

[56] Piana S, Lindorff-Larsen K and Shaw D E 2011 *Biophys. J.* **100** L47

[57] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 *J. Chem. Phys.* **79** 926

[58] MacKerell A D, Bashford D, Bellott M, *et al.* 1998 *J. Phys. Chem. B* **102** 3586

[59] Altis A, Otten M, Nguyen P H, Hegger R and Stock G 2008 *J. Chem. Phys.* **128** 245102

[60] Laio A and Gervasio F L 2008 *Rep. Prog. Phys.* **71** 126601

[61] Torrie G M and Valleau J P 1977 *J. Comput. Phys.* **23** 187

[62] Allen R J, Valeriani C and Wolde P R 2009 *J. Phys.: Condens. Matter* **21** 463102