Published in final edited form as:

J Chem Theory Comput. 2017 May 09; 13(5): 2230–2244. doi:10.1021/acs.jctc.6b01204.

Systematic Dissociation Pathway Searches Guided by Principal Component Modes

Zhiye Tang and Chia-en A. Chang*

Department of Chemistry, University of California, Riverside, California 92521, United States

Abstract

We introduce a novel method, Pathway Search guided by Internal Motions (PSIM), that efficiently finds molecular dissociation pathways of a ligand–receptor system with guidance from principal component (PC) modes obtained from molecular dynamics (MD) simulations. Modeling ligand–receptor dissociation pathways can provide insights into molecular recognition and has practical applications, including understanding kinetic mechanisms and barriers to binding/unbinding as well as design of drugs with desired kinetic properties. PSIM uses PC modes in multilayer internal coordinates to identify natural molecular motions that guide the search for conformational switches and unbinding pathways. The new multilayer internal coordinates overcome problems with Cartesian and classical internal coordinates that fail to smoothly present dihedral rotation or generate nonphysical distortions. We used HIV-1 protease, which has large-scale flap motions, as an example protein to demonstrate use of the multilayer internal coordinates. We provide examples of algorithms and implementation of PSIM with alanine dipeptide and chemical host–guest systems, 2-naphthyl ethanol– β -cyclodextrin and tetramethylammonium–cryptophane complexes. Tetramethylammonium–cryptophane has slow binding/unbinding kinetics. Its residence time, the length to dissociate tetramethylammonium from the host, is ~ 14 s from experiments, and PSIM revealed 4 dissociation pathways in approximately 150 CPU h. We also searched the releasing pathways for the product glyceraldehyde-3-phosphate from tryptophan synthase, and one complete dissociation pathway was constructed after running multiple search iterations in approximately 300 CPU h. With guidance by internal PC modes from MD simulations, the PSIM method has advantages over simulation-based methods to search for dissociation pathways of molecular systems with slow noncovalent kinetic behavior.

Graphical Abstract

*Corresponding Author: Phone: (951) 827-7263. chiaenc@ucr.edu.

ORCID

Zhiye Tang: 0000-0002-5092-1575

Notes

The authors declare no competing financial interest.

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b01204.

Flowchart of principal component analysis; fragment linker in multilayer internal coordinate definition; construction of multilayer internal coordinate definition; periodicity of dihedral and covariance matrix; search parameter and default values; alanine dipeptide search pathways; structure of β -cyclodextrin and definition of fingerprint angle; representative conformations of cryptophane–tetramethylammonium dissociation pathway; umbrella sampling of the cryptophane–tetramethylammonium complex; the second dissociation pathways for product G3P leaving its protein binding site (PDF)

1. INTRODUCTION

Modeling molecular association or dissociation reveals binding mechanisms and noncovalent binding kinetics, which bring insights into and understanding of molecular recognition and protein function.^{1–5} It also has various applications in chemical and pharmaceutical industries. In fact, binding kinetics is of crucial importance in drug development.^{6–10} Protein folding and protein function are often coupled with large-scale conformational switches. Molecular modeling that samples large-scale protein conformational transitions aids investigation in protein folding and their function as well. During ligand and protein binding/unbinding processes, numerous conformations appear that determine fast/slow binding kinetics during these temporary states. However, experiments usually have limitations in providing detailed descriptions of these transitions. All-atom molecular dynamics (MD) simulations that provide dynamic information with atomistic resolution is an alternative approach. The molecular energy surface is not smooth but contains many local energy minima and energy barriers with various heights between the minima. Therefore, the time for large-scale protein motions typically occurs in microseconds (μ s) to seconds, and ligand unbinding from its protein binding pocket may take even longer than seconds. Recent advances in computation resources that used graphic processor units (GPU)-accelerated techniques or specialized Anton machines allowed for performing traditional brute-force MD simulations to up to a few milliseconds (ms).^{11–15} However, performing ms MD simulations is far from reality for most researchers. Enhanced sampling methods such as the accelerated MD techniques and basin-hopping method successfully sample conformational changes and locate local energy minima.^{16–18} Regardless, existing methods are limited in sampling ligand–protein binding/unbinding pathways. We need methods that can efficiently and accurately model ligand–protein association and dissociation processes.

The MD-based sampling methods may be the most straightforward and popular methods for sampling molecular unbinding pathways. Increasing the temperature during MD simulations directly enhances sampling; the temperature accelerated MD uses artificially high temperature for selected collective variables to achieve efficient sampling.¹⁹ Many variants of the parallel tempering or replica-exchange (REMD) methodology have been developed to reduce the computational demand for conformational sampling.^{20–25} Another common strategy is selecting significantly deviated structures from existing MD trajectories to initiate multiple MD simulations.^{17,26,27} Enhanced sampling methods that alter the potential energy surface or provide additional forces to speed up conformational sampling include metadynamics, random accelerated MD, and self-guided MD (SGMD).^{2,28–37} Other methods use soft potentials to smooth the potential energy surface.^{38,39} Several methods such as steered MD and target MD require the start and end structures and additional forces to accelerate transitions between two structures.^{40–42} To facilitate conformational escaping from local energy minima and crossing large energy barriers, Monte Carlo (MC) sampling techniques are widely used in conformational search.^{43–47} Because sampling a long and continuous trajectory for investigating complex conformational changes and binding kinetics may still be excessively time-consuming or not practical even when using enhanced MD or MC techniques, methods have been developed to run many short time-length MD simulations for several microstates and using a Markovian or non-Markovian model to describe the transitions among these microstates.^{48–51} However, it is not always easy to define microstates and accurately model a continuous long binding/unbinding trajectory from multiple short MD simulations.

In addition to MD- and MC-based sampling methods, advanced conformational search methods include LMOD and Tork, which use eigenvectors (also called “normal modes” of vibration) computed from the Hessian matrix to bring one conformation to another.^{52–54} Recent methods also combine normal-mode analysis with MD simulations or other techniques to sample protein conformational transitions.^{55–58} Soft modes, also termed low-frequency modes, from normal-mode analysis typically describe natural motions of a molecular system. Therefore, distortion of a molecular system along the soft modes followed by energy minimization is often used to search for new low-energy conformations. These conformational search methods yield distinct local-energy minima instead of a continuous trajectory for conformational transitions. Methods such as RPHSA and Hopping Minima have been developed to connect multiple found local-energy minima to model ligand–receptor binding or unbinding pathways.^{59,60} However, these methods need to locate all local-energy minima, which can be impractical for modeling association/dissociation processes for complex molecular systems that feature a huge number of local-energy minima during the transient process.

This article introduces Pathway Search guided by Internal Motions (PSIM), a pathway search method that uses principal component (PC) modes with dihedral degrees of freedom to guide the search. It is particularly efficient to model ligand–protein dissociation pathways and overcomes the limitations discussed previously. The method can also be used to sample large-scale protein conformation transitions. Principal component analysis (PCA), also termed quasi-harmonic analysis or essential dynamics for the protein research community, is a covariance-matrix-based mathematical technique used to analyze MD trajectories for

molecular motions.^{61–64} Diagonalizing the covariance matrix constructed from MD frames yields PC modes that can be expressed by a set of eigenvectors and eigenvalues. For most popular PCA for protein motions, Cartesian coordinates of the α -carbon of each protein residue are considered in the covariance matrix, and the eigenvectors reveal essential motions of the protein backbone modeled by the MD simulation. Different from normal-mode analysis, which focuses on only harmonic motions near the bottom of one local-energy minimum, PC modes can express protein motions that cover multiple local-energy wells.

Use of Cartesian coordinates is convenient in PCA, and a PC mode is a linear combination of Cartesian displacements. However, considerable dihedral rotations cannot be accurately described by Cartesian displacements. Distorting a protein system following a Cartesian PC mode produces artificial deformation of the protein, and therefore dihedral degrees of freedom are used in constructing the covariance matrix, termed dihedral PCA (dPCA). The dihedral PC modes show smooth rotations for rotatable bonds and may provide more informative results than Cartesian PC modes.^{65–68} Nevertheless, using dihedral degrees of freedom or internal bond-angle- torsion coordinates to describe large protein motions is a daunting task because of the definition of the internal coordinates, and computation issues also arise from the periodicity of dihedral angles.

To take advantage of dihedral PC modes for guiding the conformational searches, we developed a multilayer internal coordinate definition and applied trigonometric functions to correctly present protein motions using dihedral rotations. We used HIV-1 protease (HIVp) as an example to illustrate the multilayer internal coordinate definition. With PSIM, a molecular system is distorted along the dihedral PC modes and their systematic combination to generate new conformations, followed by structure examination to reject unrealistic molecular distortion. For a molecular system with N atoms, PSIM does not use the entire $3N \times 3N$ covariance matrix but instead focuses on a small set of degrees of freedom that determine the essential conformations and ligand dissociation. Tests detailed here show that PSIM is efficient to sample ligand dissociation for chemical host–guest and ligand–protein complexes.

2. METHODS

Method Overview

PSIM uses internal PC modes computed from a MD trajectory to guide conformational searches. The procedure of computing internal PC modes is similar to performing classical PC analysis in Cartesian coordinates as illustrated in Figure SI 1. PSIM uses a unique multilayer internal coordinate definition and trigonometric functions to correctly present protein motions by using dihedral rotations, as detailed in later sections. The main module of the PSIM method reranks internal PC modes by the ease of moving the ligand in a molecular complex and distorts a bound state molecular system using these internal PC modes to sample new conformations in a systematic manner, accompanied by several auxiliary modules to ensure reasonable conformations are sampled along the pathways. The systematic search tree is illustrated in Figure 1, and the overall workflow controlled by the auxiliary modules is indicated by black solid lines. A systematic search starting from one input conformation along all allowed directions of PC modes is termed a *search branch* (one

set of colored tree branches in Figure 1). The workflow of a search branch is illustrated in Figure 2. The auxiliary modules *Acceptance Test*, *Motion Test*, and *Conformation Correction*, which are used to ensure reasonable conformations, are also indicated in Figure 2 and detailed in later sections. The search starts in the root branch, and the initial conformation (the input conformation of the root branch) is distorted stepwise along both positive and negative directions of each PC mode. After each distortion step, the *Acceptance Test* is applied to each distorted conformation to determine whether a distortion along this mode/direction should be continued or not. After the distortion along one mode/direction is done, the search enters one child branch. In each child branch, the excluded PC modes are excluded based on rules detailed in the later section (see dotted lines in Figure 1). A search branch can exit upon several events (Figures 1 and 2). First, the algorithm exits from the current branch and continues from the next PC mode in its parent branch, when the searches are completed in all of its child branches from conformations distorted along nonexcluded modes. Second, when a distortion along one direction/mode for k steps ($k \geq 2$) fails the k th *Acceptance Test*, the search in the current branch is postponed, and the search enters the child branch of the postponed branch. Until the searches on the child branch, and all possible grandchild branches, are done, the search then continues from the next direction/mode in the postponed branch. Third, when the PC modes in a branch are excluded or fail in the i th *Acceptance Test* after only one step of distortion (see red dots in Figure 1), the branch is considered a dead end, and the search continues from the next PC mode in its parent branch (see the solid black line above *Dead End* in Figure 1). Fourthly, when the algorithm reaches a user-defined step number, the *output step number* (see the solid black line after *Output* in Figure 1), it outputs a trajectory of the saved conformations, exits from the current branch, and returns to the next PC mode in the root branch and continues the search. Fifthly, the *Motion Test* is performed after a user defined numbers of search to ensure reasonable conformations. If the Motion Test is passed, the search will continue as normal into the child branch. Otherwise the search will exit from the current branch and continue from the next distorted conformation in the branch that passes a previous *Motion Test* (see the right branch in Figure 1). Finally, the search is completed, and the root branch exits when the systematic search exhausts the search tree. No solvent is explicitly or implicitly included in the distortion, because the solvent effect is implicitly included in the PC modes which are built from MD simulations in either explicit or implicit solvent models. Notably, the frames saved during the search are not energy-optimized, and users may perform a quick energy minimization in implicit solvent for a trajectory to yield a postoptimized trajectory with smoother conformational transition.

Computational Details

Multilayer Internal Coordinate Definition—Because biomolecules are larger than small molecules, we developed a multilayer definition of internal coordinates to avoid unrealistic motions produced by using classical internal bond-angle-torsion coordinates (Z-matrix). The classical definition uses a set of root atoms to begin converting Cartesian coordinates to a Z-matrix, and the position of atom i depends on other atoms that are bonded in sequence (Figure 3).⁶⁹ This situation creates artifacts in that small rotations of some torsion angles close to the root atoms may result in unrealistic motions on the far end of a molecule, as shown in Figure 3.

Therefore, we divided a molecule into N fragments, shown in Figure 4, with each fragment presented by the internal coordinates, and the fragments are connected to present an entire molecule. The fragments are chosen so that usually an α -helix or β -sheet is treated as one or two fragments. A loop may be divided into multiple fragments, with 3–4 residues in each fragment. Each fragment is presented as a small molecule, shown in Figure 3, with a designated root dihedral angle in the fragment. In systems with multiple molecules, each molecule is considered an individual group(s). A large molecule may also be represented by several groups, and each group has its own root fragment. For example, HIV-1 protease (HIVp) has two monomers, and each monomer is considered one group. As a result, the protein has two groups (see Figure 4), and typical kinases may have two groups, one for the C-lobe and another for the N-lobe. The root fragment of a group is used to connect other fragments through the six pseudodegrees of freedom shown in Figure SI 2. If more than one group exists in a molecule, the groups are again connected via the six pseudodegrees of freedom.

Here we use HIVp as an example. HIVp is a homodimer; therefore, the internal coordinates are defined identically for the two monomers. Each monomer is treated as one group, and the group is divided into 21 fragments, with the root fragment located on the N-terminus. As illustrated in Figure 4, three fragments are circled, and they are connected to the root fragment within the same group via six pseudodegrees of freedom (see details in Figure SI 2). Then the root fragments of the two monomers are connected via the six pseudodegrees of freedom.

Because a molecular system is divided into several fragments and each fragment is directly connected to the root fragment of its group by six pseudodegrees of freedom, atoms at the boundaries of two fragments may not be connected by physical bonds in our multilayer internal coordinates (Figure SI 2). This situation may produce unrealistic distortion in a PC mode. Therefore, one needs to assign fragments carefully and validate the PC motions beforehand. In addition, we recommend that the root atoms be located on the α -carbon, not atoms in flexible side-chains. A practical protocol for constructing the multilayer internal coordinate definition is provided in the SI.

Selection of Degrees of Freedom for the Covariance Matrix and PCA—As

shown in Figure 3, multiple torsion angles such as t_8 and t_9 may be used to present a bond rotation. Some bond rotations, such as methyl rotation and rotating torsion angles in rigid rings, are not important conformation determinants. In addition, peptide bonds in the protein backbone are not considered flexible. Therefore, in our internal PCA, only degrees of freedom important in conformational changes or ligand unbinding are selected into the covariance matrix. Notably, in addition to selected backbone and side-chain dihedral degrees of freedom, the pseudodegrees of freedom used to connect fragments are automatically included in the covariance matrix. When calculating the covariance matrix, the periodicity of dihedrals is considered, and the details are given in the SI.

Standard statistical procedures are used to compute the eigenvalues and eigenvectors for a given covariance matrix for PCA.⁷⁰ The resulting eigenvalues and eigenvectors are automatically ranked by the eigenvalues, and each eigenvector corresponds to an internal PC

mode for the molecular system. The PC modes are normalized before their use in the conformational search algorithm.

Conformational Search Algorithm. Re-Ranking Internal PC Modes—Before the search begins, we rerank the PCs computed from the covariance matrix to enhance the efficiency for ligand dissociation from a receptor. PCs with the greatest eigenvalues usually identify major motions of a molecule. However, for ligand–receptor motions, large relative motions of the two molecules may not appear in these low mode PCs. A complex conformation is distorted along the positive and negative directions of all PC modes, and the PC mode that moves a ligand the farthest is reranked as the first PC mode. As a result, the search can guide the ligand dissociation more efficiently.

Systematic Search Using PC Modes—Users provide an initial conformation to start the PSIM search, and the program performs a systematic conformational search using all the PC modes, as shown in Figures 1 and 2. A conformation is distorted along the positive and negative direction of each mode, except for a few PC modes/directions to be excluded, termed excluded modes (see Figure 1 dotted line). The excluded modes are those used in the parent branch to find the input conformation in the current branch and the opposite direction of any modes used to find the current conformation along the entire pathway; therefore, one can avoid distorting the molecular system to the same direction or conformation(s) that PSIM already visited. The acceptance test detailed later is applied to each distorted conformation to determine whether a distortion along this mode/direction should be continued or not. If the first distorted conformation cannot pass the acceptance test, this mode/direction is discarded in the current branch. If a frame (k th, $k - 2$) cannot pass the acceptance test, distortion along this mode/direction is terminated, and the last frame saved ($k - 1$ th) from this mode/direction is then used to initiate a new distortion by using a new mode/direction in the child branch. Each frame that passed the acceptance test is saved in the output trajectory. The procedure continues until the search tree is exhausted or reaches a user-defined maximum number of steps (frames to be saved), *output step number*, which are 2000 to 4000 steps in this work. Although the acceptance test avoids distorting a molecular system to a significantly unrealistic conformation, we developed two additional steps, *Motion Test* and *Conformation Correction*, to ensure that the protein does not denature and the ligand is not stuck in the binding site or moving back and forth in one location, as detailed in the following subsections.

Conformational Distortion—Before distorting a conformation to perform the search, PSIM converts the Cartesian coordinates (x, y, z) to internal coordinates (Q) by using the same multilayer internal coordinate definition used to compute the internal PC modes. The eigenvector (dQ) of an internal PC mode is applied to yield distorted internal coordinates ($Q' = Q + k_s dQ$), where k_s is the user-defined *step size*. To avoid clashes, small k_s such as 0.02 or 0.05 is recommended to gradually distort a molecule. To enhance the motion of the ligand, a number such as 1, 2 or 20, called *boost*, is multiplied to the degrees of freedom which correspond to ligand motion relevant to the receptor. Because dQ includes only selected degrees of freedom, the elements of the matrix not considered in dQ remain the same. After distortion, the program converts Q' back to Cartesian coordinates (x', y', z') by

using an existing algorithm⁷¹ for the acceptance test. If the conformation passes the test, a frame is saved in Cartesian coordinates and is converted to internal coordinates Q again to continue conformational distortion ($Q' = Q + k_s dQ$).

Acceptance Test—The acceptance test contains three tests: 1) collision test, 2) bond stretching test, and 3) bond angle bending test. The collision test detects whether any two nonbonded atoms i and j clash with each other. The distance between two atoms, r_{ij} , is measured and compared with their van der Waal radii, r_i and r_j , assigned in the force field for the molecular system. If $r_{ij} > k_{ds}(r_i + r_j)$, where k_{ds} is the *pair distance scaling* factor with default value 0.5, the collision test is passed. The program uses bond stretching and bond angle bending tests to ensure that the bond and bond angle in the distorted conformation do not considerably deviate from their equilibrium position. In the two tests, every bond, b_i , and angle, a_i , are calculated and compared to their equilibrium values assigned by the force field, b_0 and a_0 . A distorted conformation can pass the test when every bond b_i is within a desirable length, $b_0(1 - k_{bs}) < b_i < b_0(1 + k_{bs})$, where k_{bs} is the *bond stretch* scaling factor with a default value 0.4. The same test is performed for every bond angle bending with the k_{as} (*bond angle bending* scaling factor) as well. Table SI 1 lists all the values of scaling factors used for each molecular system in this study and the default values. Of note, no energy calculations are needed, so the tests are very fast.

Motion Test and Conformation Correction—Even if a frame passes the acceptance test, the protein may be denaturing and the ligand may get stuck in the binding site. Therefore, the program uses a root-mean-square deviation (RMSD)-based motion test to ensure that the search leads to proper ligand motions and protein conformations. When the acceptance rejects the k th step distortion along one PC mode, the k -th conformation will be used as input conformation in the child branch. Before the child-branch search starts, if the search passes the *motion test step number*, which has a default value 200, the motion test is carried out for this frame. This is because *motion test* is after a series of searches, and it is not efficient to stop the search to do *motion test* exactly at every 200 steps. For example, if the search has *output step number* = 2000, the program will perform the motion test for approximately every 200 frames (e.g., frame #200, frame #400, frame #600). For example, if a search already finished 197 steps along one search branch and the search needs to continue using another branch (Figures 1 and 2), because 197 steps have not reached #200 yet, no motion test is performed. The search continues but stops after moving nine more steps (197 + 9), which brings the accumulated steps to more than 200. Then the motion test will be carried out for the saved frame #205 (197 + 9–1). If the motion test passes, the distortion can continue using another mode for the search.

Before carrying out the motion test, the receptor is aligned to the initial structure, frame #1. Notably, if the molecular system is a complex, only the protein or chemical host is used in the alignment. Here we use frame #400 as an example, in which the second motion test is performed, and the RMSD for the protein and the ligand with respect to frame #1 is measured. If the RMSD for the protein is smaller than a cutoff value R_{RMSD} (default is 5 Å) and the RMSD for the ligand is larger than another cutoff value $L_{\text{RMSD1}} \times \text{number of motion test}$ (default of L_{RMSD1} is 0.1 Å, number of motion tests is 2 at frame #400), then

the motion test continues. Frame #400 is then aligned with frame #205, the previous frame that passed the motion test, to perform another ligand motion test. If the RMSD for the ligand is larger than L_{RMSD2} (default is 0.5 Å), then the motion test passes. The second ligand motion test is for ensuring that the ligand is indeed moving. Although this study considered all-atoms in the alignment and RMSD calculations, users may consider only the protein backbone for the alignment.

The frame that passes the motion test will be followed by a quick minimization and then saved in the trajectory. This step is termed *Conformation Correction*. Because the search continues distorting the molecular system, *Conformation Correction* prevents the program from using unrealistically deformed conformations during the search. To speed up the calculation, PSIM minimizes only bond stretching, bond angle bending, and the improper term using less than 100 minimization steps. However, if the motion test fails, the search will be backed up to the conformation in the previous motion test and continued using the next mode in that search branch.

Molecular Systems

This subsection includes simulation configurations such as MD setup and parameters used in PSIM and umbrella sampling. All MD simulations were run with the Amber 14 package⁷² and a 2 fs time step. The SHAKE algorithm was used to constrain hydrogen atoms in water molecules during the simulations. Unless otherwise stated, all systems were heated gradually and equilibrated before the production runs. Parameters used in the PSIM search are detailed in Table SI 1. All PSIM calculations in this work were performed by using Intel Xeon E5-2620 v2 CPUs, and computational times were single CPU clock time.

Alanine Dipeptide—We used Amber FF14SB force field for alanine dipeptide.⁷³ The molecule was heated from 200 to 298 K, followed by a 10 ns MD simulation in implicit water by using the generalized Born (GB) model.⁷⁴ The simulation was performed with the Amber 14 package⁷² and a frame was saved every 1 ps. We obtained the internal PC modes by selecting the ϕ and ψ angles in the molecule with all 10,000 frames from MD. Because this molecule is small enough, the whole molecule was treated as one fragment. We randomly chose one conformation from the MD trajectory and minimized the frame, termed Minimum A, to serve as a starting conformation. The motion test was disabled for this small molecule. The *step size* was set to 0.05. The *pair distance scaling* factor was set to 0.8. The *bond stretch* and *bond angle bending* factors were set to 0.2. The *output step number* was set to 100 steps. Both PC modes were used in the conformational search. We also saved trajectories that were postoptimized by Broyden–Fletcher–Goldfarb–Shanno (BFGS) minimization with GB for 50 steps.

The potential energy surface (PES) of the alanine dipeptide in aqueous solution was obtained by rotating the ϕ and ψ angles stepwise (0.5°) and calculating the energy of resulting conformer with the GB model.⁷⁴ The original and minimized transition pathways obtained by PSIM were projected onto the PES. Corresponding conformational transitions were also illustrated.

HIVp—A trajectory from a 250 ns MD simulation saved every 100 ps of HIVp was used in the internal PC calculation. The Amber FF99SB force field was used for the protein. Each of the two identical monomers was divided into exactly the same 21 fragments according to the secondary structures followed by a trial-and-error process and refinement (Figure 4). Only ϕ and ψ angles on the backbone were considered in the internal covariance matrix. First, three low mode motions were investigated. For comparison, the first three PC modes from a dihedral PCA based on a classical internal coordinate definition (Z-matrix) and an α -carbon PCA in Cartesian coordinates were computed. The root-mean-square fluctuations (RMSFs) of the internal PCA with a multilayer internal definition, classical dihedral PCA, and α -carbon PCA were calculated and compared.

β -Cyclodextrin-2-naphthyl ethanol—The structure of β -cyclodextrin was from the Cambridge Crystallographic Data Centre (Refcode: WEWTOJ), and the structure of 2-naphthyl ethanol was drawn manually by using Vega ZZ.⁷⁵ The ligand was docked into the binding site with use of Autodock 4.2.⁷⁶ The initial structure from docking was solvated with 1736 TIP3P⁷⁷ water molecules, and then the Amber general force field was applied to it (GAFF).⁷⁸ MD simulations were performed with the Amber 14 package.⁷² A water equilibrium was run at 298 K in the NPT ensemble for 1 ns. Then the system was heated to 298 K. A trajectory containing only 500 ns MD simulation length and only a ligand–cyclodextrin bound complex was used for internal PCA calculation. For internal PCA, the β -cyclodextrin was treated as a whole fragment. All 72 PC modes obtained from the internal PCA were used in the conformational search. The *step size* used in this example was 0.02, different from all other examples (Table SI 1). We collected 20 dissociation pathways from PSIM in approximately 60 CPU h.

As compared with brute force MD simulations, we also had a production run of total 11.5 μ s, run at 298 K in the NPT ensemble. Overall, 15 dissociation events were identified from the trajectory. For the 15 dissociation events, we evaluated the conformations of cyclodextrin. We first smoothed the trajectories by averaging 100 forward and 100 backward frames on the concurrent frame throughout the whole trajectory to remove the noise. Then along the smoothed trajectory, we calculated the fingerprint angles of β -cyclodextrin, as defined in Figure SI 3. For the 20 dissociation trajectories from PSIM, we calculated the same fingerprints without smoothing the trajectory. We analyzed the conformational changes of the host molecule by comparing the MD and PSIM trajectories using the conformational fingerprints.

Cryptophane–Me₄N⁺ Complex—The bound structure of cryptophane–Me₄N⁺ was taken from a previous work,⁵⁹ and the CHARMM22 force field was used for this system.^{79–81} MD simulation was performed with a similar procedure as for alanine dipeptide, except that the dielectric constant in GB⁷⁴ of the solvent was set to 8.42, corresponding to the value of tetrachloroethane. The MD production run was 20 ns. In the MD run, no dissociation event occurred. We performed internal PCA by dividing the cryptophane into six fragments and leaving the ligand as one group/fragment. In the conformational search, the *output step number* was set to 3000. The *step size* was set to 0.05. *Pair distance scaling* factor was set to 0.5. All 63 modes were used for the conformational search. Four ligand dissociation

pathways, saved as four trajectories, were found by PSIM in approximately 150 CPU h. The trajectories were then postoptimized by the BFGS minimizer for 100 steps to remove unphysical clashes, bond stretching and bond angle bending. The first trajectory was used as an example for illustrating the search result. For comparison, one dissociation pathway was created by using steered MD by pulling the ligand outside the binding cavity of cryptophane for 200 steps. The velocity of the ligand was held as a constant at 0.1 Å per step. We also performed umbrella sampling by using conformations along the found dissociation pathway to construct a potential of mean force (PMF) as detailed in the SI.

Tryptophan Synthase—The crystal structure of tryptophan synthase (TRPS) in complex with the aminoacrylate E(A-A) intermediate in the β -subunit and the F9 inhibitor in the α -subunit was obtained from a PDB code 4HN4.⁸² We replaced the F9 inhibitor by the natural product glyceraldehyde-3-phosphate (G3P) in the α -subunit (Figure 5) because we aimed to model product dissociation. A 10 ns MD simulation for the dimeric α - and β -subunit complex was performed, with a frame saved every 1 ps. The protein and substrates were modeled by using Amber FF99SB and GAFF,⁷⁸ respectively. For the internal PC analysis, we considered only coordinates for the α -ligand G3P and the α -subunit of TRPS. The α -subunit has 266 residues and was divided into 42 fragments by using the procedure described for HIVp. We considered only backbone and side-chain dihedrals in residues 56 to 67, 174 to 193, 210 to 217, and 232 to 242 and dihedrals in G3P when constructing the dihedral covariance matrix. All 254 PC modes were used for conformational search. Any randomly chosen frames from the MD simulation could serve as an initial conformation for our search, and we used the initial structure of MD simulation in this case. The *output step number* was set to 2000 steps, with the *step size* equal to 0.05 and the *pair distance scaling* equal to 0.5.

Starting from the initial structure, we performed the PSIM search and repeated for multiple iterations. We chose conformations from the outputted trajectories as new input conformations for further conformational search based on four criteria: 1) TRPS remained folded in its native structure, 2) loop 6 is more open and not in the fully closed form, 3) conformations of G3P were not flipping over in the binding site, and 4) G3P was not moving to the intersubunit tunnel, which should be used for diffusing the other product, indole, to the β -subunit. We manually chose conformations from an existing search to start a new iteration until G3P dissociated from the binding site. We repeated this procedure and performed nine iterations to obtain two dissociation pathways for this system, and the searches can be continued if users prefer. Of note, because of the large protein systems, a dissociation pathway is a collection of multiple trajectories obtained from each iteration. When the search reaches the *output step number*, 2000, PSIM outputs one trajectory. As a result, each iteration output multiple trajectories, but none brought G3P directly from the binding site to the solvent. Thus, users need to manually select frames from the current iteration to continue a new PSIM search until G3P successfully leaves the binding pocket. Multiple trajectories from different search iterations are then combined for a complete dissociation pathway. Therefore, two pathways may share the same trajectories in the beginning of dissociation and deviate afterward.

3. RESULTS AND DISCUSSION

In this section, we first illustrate the basic algorithm of the PSIM method by using a classical model molecule, alanine dipeptide. We used HIVp, which has large-scale conformational switches, as an example to demonstrate the multilayer internal coordinates. Large-scale flap motions presented by the PC modes generated from the multilayer internal coordinates, α -carbon in Cartesian space, and the classical internal coordinates (Z-matrix) were compared. To show that the dissociation pathway from PSIM is similar to those found with MD simulations, the search trajectories are compared to conformations obtained from MD runs for β -cyclodextrin-2-naphthyl ethanol, with dissociation pathways that can be sampled by using MD simulations for this fast-binding/unbinding system. PSIM was also used to search for dissociation pathways for a slow ligand dissociation system, a cryptophane-Me₄N⁺ complex. Finally, we used PSIM to find pathways for the product G3P releasing from the active site of TRPS.

Performance of the Multilayer Internal Coordinate Definition

As illustrated in Figure 3, using the classical internal coordinates (Z-matrix) to present a large molecule can create artifacts in that small rotations of some torsion angles may result in unrealistic motions on the far end of the molecule. However, the multilayer internal coordinate definition avoids this accumulated dependence problem by breaking the dependence by using fragment and reconnecting fragments with pseudodegrees of freedom. Figure 6 illustrates, for HIVp, motions identified by the first PC mode computed by the covariance matrix by using the α -carbon in Cartesian coordinates, dihedrals in multilayer internal coordinates and dihedrals in classical internal coordinates. The flap motions are well studied and show the nearly symmetric open/closed switches that have been reported.⁸³ The flap motions identified by the PC modes using the first two coordinates reproduce the known flap opening/closed conformations, whereas PC modes using the classical definition for internal coordinates show unrealistic motions because of the accumulation of dependence from their covalent-bonded predecessors. The RMSFs of the first PC mode computed by the three coordinates were calculated to quantify the fluctuation of each residue for detailed comparison (Figure 7). The calculated RMSF values clearly demonstrate the agreement between motions revealed by the first PC mode computed from the covariance matrix by using the multilayer internal coordinate and the α -carbon in Cartesian coordinates. A small deviation arises from use of fragments in the loop region, but the difference in overall motion is indistinguishable. However, use of the classical internal coordinate is unambiguously not suitable for presenting this large-scale motion (Figure 7).

Although PC modes that use the α -carbon in Cartesian space yield the same motion as those that use multilayer PC modes, the internal coordinate has advantages in more accurately presenting motions for side-chains. Selecting dihedral rotations is challenging when using Cartesian coordinates. We examined the connection between each fragment to ensure that the motion is continuous and rigorous. We suggest visualizing the motions of first internal PC modes to confirm the selection of fragments. Although in general, one α -helix or β -sheet is assigned one fragment, it may be divided into multiple fragments depending on the length and flexibility of the region. When aggressively distorting the molecule along a PC mode,

artifacts in the connection area (bond and angles) between fragments can be observed because of use of pseudobondangle- torsion in the multilayer internal coordinates to connect the fragments. As a result, PSIM includes the motion test and conformation correction during the search, as detailed in the Methods section.

Pathways of Conformation Transitions for Alanine Dipeptide

Here we first used a model system, alanine dipeptide, to illustrate how PSIM searches for new conformations and conformational transition pathways by using internal PC modes obtained by MD simulations. We obtained four conformational transition pathways with the PSIM method shown in Figure 8, 1–4. The search used Minimum A to start distorting the dihedrals along four directions, the positive and negative of the two PC modes, and stepwise rotations of dihedrals show straight lines when plotting the dihedral rotations. For example, in Figure 8 1, by distorting along the positive direction of mode #1 for 18 steps (green line in the middle figure), the search reached a high energy region (15 kcal/mol). The acceptance test detected the energy barrier and rejected further distortion along this PC mode/direction. Notably, the acceptance test is based on a geometrical cutoff rather than an energetic cutoff to save computation time; however, the test is sensitive to identify energy barriers. The search then started to distort the molecule by using mode #2. Because distorting the molecule along the positive direction of mode #2 brought the conformation to the same energy barrier, the search immediately abandoned this direction but used the negative direction of mode #2 for distortion. After distorting for another 82 steps (yellow line in the middle figure), the accumulated step number reached the user-defined *output step number* 100, so the search was terminated.

The contour plot in the right of Figure 8 shows the energy landscape for different dihedral rotations. The first found pathway was able to cross the energy barrier in cyan (~7–8 kcal/mol) and lead to minimum C. Because alanine dipeptide has intramolecular interactions in addition to dihedral energy, we performed quick optimization to postprocess the pathway. The minimized path shows a smoother transition in the energy landscape, and before reaching the final destination minimum C, it visited the vicinity of minimum B, searching for a low energy route to regions near B, and then ended at C. Similarly, in the identified pathway #2 shown in Figure 8 2, the search started from Minimum A along the negative direction of mode #1 (green line in the middle figure). The distortion stopped when encountering a high-energy barrier (the red region in the right figure), and another mode was then used for continuing the distortion, leading to the conformation transitions from A to E. The details of searching modes and number of steps are summarized in Table SI 2.

Dissociation of 2-Naphthyl Ethanol from β -Cyclodextrin: Pathways Found by PSIM vs MD Simulations

Our first ligand–receptor system, β -cyclodextrin and the 2-naphthyl ethanol complex, has fast kinetics, with experimentally measured k_{on} and k_{off} values $2.9 \pm 1.6 \times 10^8$ (1/s·M) and $1.8 \pm 0.7 \times 10^5$ (1/s), respectively.⁸⁴ Therefore, the dissociation pathways found by our search method can be compared by the pathways sampled with the unbiased classical MD simulations. The search is efficient for this system with short ligand-binding residence time. By using the reranked PC modes, PSIM could identify 20 pathways in approximately 60

CPU h. For comparison, 15 dissociation pathways were observed during a total of 11.5- μ s MD simulations. Notably, in the crystal structure of free β -cyclodextrin, the cavity of the host is wide open, without any ring flipping (Figure SI 3). The dissociation pathways obtained by PSIM and MD simulations both suggest that the glucose ring flips in the ligand dissociation process. The ring-flipping behaviors in the dissociation pathways were examined by using the fingerprint angles (Figure SI 3) to classify the dissociation events into three groups: a) no glucose ring flipping, b) flipping of one glucose ring, and c) flipping of multiple glucose rings. Representative dissociation trajectories from MD and PSIM and their plots of fingerprint angles are in Figure 9.

Both PSIM and MD simulations revealed the dissociation pathways classified in the above groups. With approximately 60 CPU h, PSIM could reproduce 20 dissociation pathways that were similar to the 15 sampled by 11.5- μ s MD simulations. The test demonstrated the feasibility of using PC modes computed from a MD trajectory that included only the bound state to guide the search. Even when the covariance matrix included only information from the bound complex, natural motions suggested by the PC modes with the multilayer coordinates could guide ligand dissociation. The similarity between PSIM and MD trajectories validates the guided distortion utilized by PSIM. In Group a, both MD and PSIM could find dissociation pathways in which the glucose rings did not flip significantly. In Group b, one glucose ring was parallel to the regression plane and others are perpendicular to it, and when dissociation occurred, that glucose ring (arrowed in gray) flipped during ligand unbinding, as shown in the trajectory found by MD and PSIM. In Group c, multiple glucose rings flip in the dissociation pathways found by both MD and PSIM. Although the dissociation paths and concerted motions between the ligand and cyclodextrin are similar during unbinding, MD trajectories showed more complicated molecular fluctuations. For example, in MD simulations, typically, the ligand moved back and forth when leaving the host, but the PSIM trajectories showed ligand movement clearly toward one direction. MD simulations follow molecular mechanics and consider explicit water molecules, whereas PSIM relies on PC modes in which solvent effects and comprehensive interactions can only be captured implicitly. As a result, PSIM cannot realistically completely reproduce MD simulations. In addition, the PC modes may not be able to entirely reproduce molecular motions; thus, not all different glucose flipping events are captured by using PSIM. However, the found pathways serve as a decent initial pathway for further refinement and postprocessing.

Tetramethylammonium Unbinding from Cryptophane- E. Host: Finding the Pathways and the Corresponding Free Energy Barriers

The cryptophane- Me_4N^+ complex has a slow dissociation rate constant ($k_{\text{off}} = 4.81 \times 10^{-2}$ 1/s; residence time = 14 s),⁸⁵ which is impractical when using classical MD simulations to sample its dissociation pathways. With 63 PC modes constructed from a 20 ns MD simulation for the bound complex, PSIM successfully found 4 dissociation pathways in approximately 150 CPU h. The trajectories are similar because of the symmetry of the host molecule, and one representative dissociation pathway is shown in Figure 10. The two methoxybenzyl groups serve as a gate for ligand binding/unbinding, and the gate is closed when the cryptophane host is in a free state and bound to a guest such as

tetramethylammonium. Our search found that the methoxy groups need to rotate away to open the gate during ligand dissociation processes (Figure SI 4). We also compared one pathway found by PSIM with two other methods: the hopping minima method and steered MD (Figure 10). The hopping minima method first performed thorough conformation search to obtain numerous local energy minima, and the minima were connected by using normal modes to build the association pathways. When performing steered MD, we assumed no *a priori* knowledge for the dissociation pathway, and tetramethylammonium was pulled from the binding pocket directly toward the solvent. All three methods showed the gate opening for ligand unbinding. After tetramethylammonium left the pocket, it immediately retained contact with the surface of the host molecule, as illustrated in the pathways modeled by hopping minima and PSIM (Figure 10). However, because of the biased pulling, a pathway generated by the steered MD showed that the ligand moved straight toward the solvent, which is less likely because the pulling direction did not consider possible intermolecular attractions when tetramethylammonium left the pocket.

To reveal the free energy barrier for ligand unbinding, umbrella sampling together with the weighted histogram analysis method (WHAM), was used to plot a potential of mean force (PMF) profile (Figure SI 5). The choice of initial conformations was based on a dissociation pathway found by PSIM, as detailed in the Methods section. The energy barrier of dissociation $\Delta G_{d,calc}^{\ddagger}$ is 16.8 kcal/mol, and the association barrier $\Delta G_{a,calc}^{\ddagger}$ is 9.1 kcal/mol. The binding affinity computed from this pathway yields -7.8 kcal/mol, which agrees well with the experimental measurement, $G^{\circ} = -7.3$ kcal/mol.⁸⁵ For protein–ligand complex systems, we anticipate finding significantly more complicated dissociation pathways. In addition to classical umbrella sampling, other methods⁸⁶ may be needed for computing a free energy profile along a suggested dissociation pathway.

Search for Dissociation Pathways for Product G3P Leaving Its Protein Binding Site

This discussion illustrates the use of PSIM for finding the dissociation pathways in protein systems. The α -subunit of TRPS cleaves indole-3-glycerol phosphate (IGP) to yield indole and G3P; indole then diffuses to a tunnel that leads to the β -subunit for completing tryptophan biosynthesis. The protein subsequently undergoes conformational changes to release the other product, G3P. This slow process is difficult to be modeled by using MD simulations. Therefore, we used PSIM to sample G3P dissociation pathways with a short MD run for the bound complex; G3P remained in the bound state during a 10 ns MD simulation.

In the two dissociation pathways obtained by PSIM, G3P was released from the binding site of the α -subunit of TRPS. For demonstration purposes, we stopped the search when finding two pathways instead of exploring multiple possible dissociation pathways. During ligand dissociation, the protein maintained a nice folded conformation because we distorted only regions relevant to opening/closed conformations of the α -subunit of TRPS. During the search, our *motion test* and *conformation correction* steps successfully prevented the protein from unreasonably huge deformation.

Figure 11 illustrates the first dissociation pathway obtained by PSIM. Loop 6 changed the conformation significantly to switch the protein from a closed to an open form for releasing G3P. Our search also showed that loops 6 and 2 moved in concert with several other α -helices near the binding site (Figure 5). A closer look reveals that residue Thr182 located near the center of loop 6 began to rotate, breaking the nonpolar contacts with G3P, which resulted in a more flexible loop 6. More key interactions such as nonpolar attraction between Phe211 and G3P and the hydrogen bond between Ser234 and G3P, then lost, provided more space for G3P to move. The search showed that after loop 6 opened fully, G3P stayed in the binding site shortly before it broke all intermolecular hydrogen bonds and then left the pocket.

Of note, because the system is large, even exploring all combinations of modes as shown in Figure 1, one PSIM iteration could not successfully bring G3P to the solvent. Therefore, multiple iterations are needed. We used the default value of the *output step number* set to 2000. Thus, a trajectory was saved when the search successfully moved on for 2000 steps, and we examined the trajectories found by PSIM. Multiple frames from the iteration were selected to continue another iteration. Because the system has more than 200 PC modes, finishing the distortion guided by all the combined modes may be unnecessary. Users should check the output trajectories during the search and stop the iteration when reasonable numbers of unbinding pathways are sampled or when the search is stuck. Users can restart a new search iteration by using a frame from any of the found pathways. For this system, 2000 frames were saved in one trajectory, and 620 trajectories were found in nine search iterations in approximately 300 CPU h in total. Most trajectories did not bring the motions of interest, and only six trajectories were used to construct the dissociation pathway shown here. Another 12 trajectories were used to build another found pathway illustrated in Figure SI 6.

General Comments—The PSIM method uses internal PC modes computed from short MD simulations. It shares some similarity with the normal mode-based conformational search methods.^{52–55} The method is more suitable to search possible dissociation pathways and protein motions, rather than sampling various conformations of small molecules for docking. Unlike other methods using classical normal modes, which present motions near the bottom of an energy well, PSIM takes advantage of various snapshots obtained from MD runs to construct PC modes that can cover multiple energy wells. Hence, the PC modes are more flexible, and the search uses these modes to guide the distortion of the molecule or the complex. Using dihedral degrees of freedom allows for expressing bond rotations accurately and for easy selection of dihedral rotations important in changing molecular conformations. A dihedral rotation presented by Cartesian coordinates is a linear combination of Cartesian displacements, so selection of specific dihedral rotations is not easy when calculating the covariance matrix and PC modes, which yields a larger matrix and significantly more modes. Use of internal PC modes produces a more efficient search by providing easier conformational changes and fewer modes to consider. In addition, typical conformational search methods involve full minimization, which leads to distinct local-energy minima. PSIM uses only quick but not full minimization (conformation correction) to eliminate unrealistic bond and angle stretch, so the output trajectories show continuous conformational transition and ligand unbinding. The strategy can save computation time. For large protein

systems, one search may not be enough, and multiple iteration is necessary. Any frames from the output trajectories can be used as new conformations to begin another iteration that can result in other trajectories smoothly connecting to the previous ones. No artificially defined states are needed during the search or connecting the search results, which are required for methods involving the Markov State Model.^{87,88}

When the molecular system moves far from its original conformations used to construct the internal PC modes, existing modes may less accurately present its motions. One can perform another brief MD run and construct another set of internal PC modes to continue the search. Because PSIM is based on molecular distortion, some unbinding pathways or conformational transitions inevitable slightly deviate from their natural motions. One can use the PSIM trajectories to launch other sampling techniques such as umbrella sampling, nudged elastic band, or other methods^{89–91} to refine the pathways and/or compute the free energy profile along a given coordinate. Moreover, we propose a method to refine the pathway from PSIM by using an elastic network model for large protein systems. The averages and standard deviations of the distances between each residue pair in the native protein calculated from a MD simulation are used to implement soft harmonic functions on the distances of residue pairs in addition to existing molecular-mechanics force field. This quick minimization process in our postanalysis can efficiently bring the protein conformations from considerably distorted structures back to conformations more similar to their native states. Although small molecular systems do not need this additional elastic network model for postanalysis, it is available in the PSIM package.

The PSIM method can be accelerated by using parallel computing techniques and using random search instead of systematic search. In the current version, we use multithread computing for the distortion modules, but for minimization, which is time-consuming for protein systems, we used a single-thread BFGS minimizer. Therefore, a multithread minimizer can be used to further increase the parallel computing efficiency. A systematic search algorithm is used to combine all PC modes for dissociation pathway search in the current version. Based on our preliminary tests (data not shown), a random combination of PC modes will keep the performance of the search but increase the efficiency. As a result, future work for PSIM will include selections for random or user preferred PC modes for the search.

4. CONCLUSIONS

Our PSIM method models conformational transitions as well as ligand dissociation pathways for chemical host–guest and biomolecular systems. Sampling molecular motions and ligand unbinding reveals insights into fundamental mechanisms about molecular recognition, conformational transition, and binding kinetics. The method requires a MD trajectory to construct internal PC modes for guiding the molecular distortion for conformational searches. Classical internal coordinates such as the Z-matrix are widely used with small molecules but create unrealistic motions when directly applying the coordinates to large protein systems. We introduced the multilayer internal coordinate definition and applied trigonometric functions to correctly present protein motions by using dihedral

rotations. The internal PC modes constructed by selected dihedral degrees of freedom can efficiently and accurately guide the conformational searches.

We have developed multiple strategies to select and rank the internal PC modes and to check and correct unrealistic molecular distortion during the search to speed up the calculations. The alanine dipeptide was used as an example to elucidate the PSIM method and to provide a validation. The method demonstrated the accuracy and efficacy of sampling multiple unbinding pathways with tetramethylammonium– cryptophane, 2-naphthyl ethanol– cyclodextrin, and G3P– TRPS complexes as example systems and deepened our understanding of noncovalent binding kinetics. We also used HIVp, a widely studied system with large flap motions, to explain how to correctly describe large-scale protein conformational changes by using multilayer internal coordinates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This study was supported by the US National Institutes of Health (GM-109045), US National Science Foundation (MCB-1350401), and NSF national supercomputer centers (TG-CHE130009).

We thank Dr. Wei Chen for useful comments on technical details and Mr. Mark Raymundo for testing the novel internal coordinate definition on HIVp.

ABBREVIATIONS

PSIM	Pathway Search guided by Internal Motions
MD	molecular dynamics
PC	principal component
PCA	principal component analysis
HIVp	HIV-1 protease
RMSD	root-mean- square deviation
PES	potential energy surface
GB	generalized Born
RMSF	root-mean-square fluctuations
TRPS	tryptophan synthase
G3P	glyceraldehyde-3-phosphate
PMF	potential of mean force
IGP	indole-3-glycerol phosphate

References

1. Klenin K, Strodel B, Wales DJ, Wenzel W. *Biochim Biophys Acta, Proteins Proteomics*. 2011; 1814(8):977–1000.
2. Andersen OJ, Grouleff J, Needham P, Walker RC, Jensen F. *J Phys Chem B*. 2015; 119(46):14594–14603. [PubMed: 26482713]
3. Boehr DD, Nussinov R, Wright PE. *Nat Chem Biol*. 2009; 5(11):789–796. [PubMed: 19841628]
4. Stank A, Kokh DB, Fuller JC, Wade RC. *Acc Chem Res*. 2016; 49(5):809–815. [PubMed: 27110726]
5. Changeux JP. *Biophys Rev*. 2014; 6(3–4):311–21. [PubMed: 25505495]
6. Copeland RA. *Nat Rev Drug Discovery*. 2016; 15(2):87–95. [PubMed: 26678621]
7. Cusack KP, Wang Y, Hoemann MZ, Marjanovic J, Heym RG, Vasudevan A. *Bioorg Med Chem Lett*. 2015; 25(10):2019–2027. [PubMed: 25782745]
8. Lu H, Tonge PJ. *Curr Opin Chem Biol*. 2010; 14(4):467–474. [PubMed: 20663707]
9. Guo D, Heitman LH, Ijzerman AP. *ACS Med Chem Lett*. 2016; 7(9):819–821. [PubMed: 27660682]
10. Meyer-Almes FJ. *Drug Discovery Today: Technol*. 2015; 17:1–8.
11. Harvey MJ, De Fabritiis G. *Drug Discovery Today*. 2012; 17(19–20):1059–1062. [PubMed: 22504137]
12. Shaw DE, Grossman JP, Bank JA, Batson B, Butts JA, Chao JC, Deneroff MM, Dror RO, Even A, Fenton CH, Forte A, Gagliardo J, Gill G, Greskamp B, Ho CR, Ierardi DJ, Iserovich L, Kuskin JS, Larson RH, Layman T, Li-Siang L, Lerer AK, Li C, Killebrew D, Mackenzie KM, Mok SYH, Moraes MA, Mueller R, Nociolo LJ, Peticolas JL, Quan T, Ramot D, Salmon JK, Scarpazza DP, Schafer UB, Siddique N, Snyder CW, Spengler J, Tang PTP, Theobald M, Toma H, Towles B, Vitale B, Wang SC, Young C. SC14: International Conference for High Performance Computing, Networking, Storage and Analysis. 2014:41–53.
13. Arkhipov A, Shan Y, Das R, Endres NF, Eastwood MP, Wemmer DE, Kuriyan J, Shaw DE. *Cell*. 2013; 152(3):557–569. [PubMed: 23374350]
14. Decherchi S, Berteotti A, Bottegoni G, Rocchia W, Cavalli A. *Nat Commun*. 2015; 6:6155. [PubMed: 25625196]
15. Pierce LCT, Salomon-Ferrer R, de Oliveira CAF, McCammon JA, Walker RC. *J Chem Theory Comput*. 2012; 8(9):2997–3002. [PubMed: 22984356]
16. Hamelberg D, Mongan J, McCammon JA. *J Chem Phys*. 2004; 120(24):11919–11929. [PubMed: 15268227]
17. Harada R, Kitao A. *J Chem Theory Comput*. 2015; 11(11):5493–5502. [PubMed: 26574337]
18. Wales DJ, Doye JPK. *J Phys Chem A*. 1997; 101(28):5111–5116.
19. Maragliano L, Vanden-Eijnden E. *Chem Phys Lett*. 2006; 426(1–3):168–175.
20. Berg BA, Neuhaus T. *Phys Lett B*. 1991; 267(2):249–253.
21. Doshi U, Hamelberg D. *J Phys Chem Lett*. 2014; 5(7):1217–1224. [PubMed: 26274474]
22. Yamamori Y, Kitao A. *J Chem Phys*. 2013; 139(14):145105. [PubMed: 24116651]
23. Swenson DWH, Bolhuis PG. *J Chem Phys*. 2014; 141(4):044101. [PubMed: 25084875]
24. Liu P, Kim B, Friesner RA, Berne BJ. *Proc Natl Acad Sci U S A*. 2005; 102(39):13749–13754. [PubMed: 16172406]
25. Ostermeir K, Zacharias M. *Biochim Biophys Acta, Proteins Proteomics*. 2013; 1834(5):847–853.
26. Bacci M, Vitalis A, Caflisch A. *Biochim Biophys Acta, Gen Subj*. 2015; 1850(5):889–902.
27. Harada R, Takano Y, Shigeta Y. *J Chem Phys*. 2014; 140(12):125103. [PubMed: 24697482]
28. Laio A, Rodriguez-Forte A, Gervasio FL, Ceccarelli M, Parrinello M. *J Phys Chem B*. 2005; 109(14):6714–6721. [PubMed: 16851755]
29. Ensing B, De Vivo M, Liu ZW, Moore P, Klein ML. *Acc Chem Res*. 2006; 39(2):73–81. [PubMed: 16489726]
30. Marinelli F, Faraldo-Gomez JD. *Biophys J*. 2015; 108(12):2779–2782. [PubMed: 26083917]
31. Geyer T. *BMC Biophys*. 2011; 4:7. [PubMed: 21596002]

32. Hornak V, Simmerling C. *J Mol Graphics Modell.* 2004; 22(5):405–413.
33. Miao Y, Feher VA, McCammon JA. *J Chem Theory Comput.* 2015; 11(8):3584–3595. [PubMed: 26300708]
34. Doshi U, Hamelberg D. *Biochim Biophys Acta, Gen Subj.* 2015; 1850(5):878–888.
35. Bernardi RC, Melo MCR, Schulten K. *Biochim Biophys Acta, Gen Subj.* 2015; 1850(5):872–877.
36. Ludemann SK, Lounnas V, Wade RC. *J Mol Biol.* 2000; 303(5):797–811. [PubMed: 11061976]
37. Wu XW, Brooks BR. *Chem Phys Lett.* 2003; 381(3–4):512–518.
38. Mollica L, Decherchi S, Zia SR, Gaspari R, Cavalli A, Rocchia W. *Sci Rep.* 2015; 5:11539. [PubMed: 26103621]
39. Sinko W, Miao YL, de Oliveira CAF, McCammon JA. *J Phys Chem B.* 2013; 117(42):12759–12768. [PubMed: 23721224]
40. Schlitter J, Engels M, Kruger P. *J Mol Graphics.* 1994; 12(2):84–89.
41. Leech J, Prins JF, Hermans J. *IEEE Comput Sci Eng.* 1996; 3(4):38–45.
42. Isralewitz B, Gao M, Schulten K. *Curr Opin Struct Biol.* 2001; 11(2):224–230. [PubMed: 11297932]
43. Cafilisch RE. *Acta Numerica.* 1998; 7:1–49.
44. Andricioaei I, Straub JE, Voter AF. *J Chem Phys.* 2001; 114(16):6994–7000.
45. Brown S, Head-Gordon T. *J Comput Chem.* 2003; 24(1):68–76. [PubMed: 12483676]
46. Li H, Li G, Berg BA, Yang W. *J Chem Phys.* 2006; 125(14):144902. [PubMed: 17042645]
47. Hayward S, Kitao A. *J Chem Theory Comput.* 2015; 11(8):3895–3905. [PubMed: 26574470]
48. Bello-Rivas JM, Elber R. *J Chem Phys.* 2015; 142(9):094102. [PubMed: 25747056]
49. Shukla D, Hernandez CX, Weber JK, Pande VS. *Acc Chem Res.* 2015; 48(2):414–422. [PubMed: 25625937]
50. Chodera JD, Noe F. *Curr Opin Struct Biol.* 2014; 25:135–144. [PubMed: 24836551]
51. Zuckerman DM. *Annu Rev Biophys.* 2011; 40:41–62. [PubMed: 21370970]
52. Kolossvary I, Guida WC. *J Am Chem Soc.* 1996; 118(21):5011–5019.
53. Kolossvary I, Keseru GM. *J Comput Chem.* 2001; 22(1):21–30.
54. Chang CE, Gilson MK. *J Comput Chem.* 2003; 24(16):1987–1998. [PubMed: 14531053]
55. Labute P. *J Chem Inf Model.* 2010; 50(5):792–800. [PubMed: 20429574]
56. Marechal JD, Perahia D. *Eur Biophys J.* 2008; 37(7):1157–1165. [PubMed: 18286277]
57. Kantarci-Carsibasi N, Haliloglu T, Doruker P. *Biophys J.* 2008; 95(12):5862–5873. [PubMed: 18676657]
58. Zheng WJ, Brooks BR. *Biophys J.* 2006; 90(12):4327–4336. [PubMed: 16565046]
59. Roberts CC, Chang C-eA. *J Chem Theory Comput.* 2013; 9(4):2010–2019. [PubMed: 26583550]
60. Strodel B, Wales DJ. *Chem Phys Lett.* 2008; 466(4–6):105–115.
61. Case DA. *Curr Opin Struct Biol.* 1994; 4(2):285–290.
62. Brooks BR, Janezic D, Karplus M. *J Comput Chem.* 1995; 16(12):1522–1542.
63. Levy RM, Karplus M, Kushick J, Perahia D. *Macromolecules.* 1984; 17(7):1370–1374.
64. David CC, Jacobs DJ. *Methods Mol Biol.* 2014; 1084:193–226. [PubMed: 24061923]
65. Mu YG, Nguyen PH, Stock G. *Proteins: Struct, Funct, Genet.* 2005; 58(1):45–52. [PubMed: 15521057]
66. Maisuradze GG, Leitner DM. *Proteins: Struct, Funct, Genet.* 2007; 67(3):569–578. [PubMed: 17348026]
67. Maisuradze GG, Liwo A, Scheraga HA. *J Mol Biol.* 2009; 385(1):312–329. [PubMed: 18952103]
68. Riccardi L, Nguyen PH, Stock G. *J Phys Chem B.* 2009; 113(52):16660–16668. [PubMed: 20028141]
69. Miyazawa T, Pitzer KS. *J Chem Phys.* 1959; 30(4):1076–1086.
70. Hotelling H. *J Educ Psychol.* 1933; 24:417–441.
71. Parsons J, Holmes JB, Rojas JM, Tsai J, Strauss CEM. *J Comput Chem.* 2005; 26(10):1063–1068. [PubMed: 15898109]

72. Case, DA., Berryman, JT., Betz, RM., Cerutti, DS., Cheatham, TE., Darden, TA., Duke, RE., Giese, TJ., Gohlke, H., Goetz, AW., Homeyer, N., Izadi, S., Janowski, P., Kaus, J., Kovalenko, A., Lee, TS., LeGrand, S., Li, P., Luchko, T., Luo, R., Madej, B., Merz, KM., Monard, G., Needham, P., Nguyen, H., Nguyen, HT., Omelyan, I., Onufriev, A., Roe, DR., Roitberg, A., Salomon-Ferrer, R., Simmerling, CL., Smith, W., Swails, J., Walker, RC., Wang, J., Wolf, RM., Wu, X., York, DM., Kollman, PA. Amber 2015. University of California; San Francisco: 2015.
73. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. *J Chem Theory Comput.* 2015; 11(8):3696–3713. [PubMed: 26574453]
74. Still WC, Tempczyk A, Hawley RC, Hendrickson T. *J Am Chem Soc.* 1990; 112(16):6127–6129.
75. Pedretti A, Villa L, Vistoli G. *J Comput-Aided Mol Des.* 2004; 18(3):167–173. [PubMed: 15368917]
76. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. *J Comput Chem.* 2009; 30(16):2785–2791. [PubMed: 19399780]
77. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. *J Chem Phys.* 1983; 79(2): 926–935.
78. Wang JM, Wolf RM, Caldwell JW, Kollman PA, Case DA. *J Comput Chem.* 2004; 25(9):1157–1174. [PubMed: 15116359]
79. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. *J Comput Chem.* 1983; 4(2):187–217.
80. Mackerell AD, Wiorkiewicz-Kuczera J, Karplus M. *J Am Chem Soc.* 1995; 117(48):11946–11975.
81. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. *J Phys Chem B.* 1998; 102(18):3586–3616. [PubMed: 24889800]
82. Niks D, Hilario E, Dierkers A, Ngo H, Borchardt D, Neubauer TJ, Fan L, Mueller LJ, Dunn MF. *Biochemistry.* 2013; 52(37):6396–6411. [PubMed: 23952479]
83. Hornak V, Simmerling C. *Drug Discovery Today.* 2007; 12(3–4):132–138. [PubMed: 17275733]
84. Barros TC, Stefaniak K, Holzwarth JF, Bohne C. *J Phys Chem A.* 1998; 102(28):5639–5651.
85. Garcia C, Humiliere D, Riva N, Collet A, Dutasta JP. *Org Biomol Chem.* 2003; 1(12):2207–2216. [PubMed: 12945915]
86. Jo S, Suh D, He Z, Chipot C, Roux B. *J Phys Chem B.* 2016; 120(33):8733–42. [PubMed: 27409349]
87. Swope WC, Pitera JW, Suits F. *J Phys Chem B.* 2004; 108(21):6571–6581.
88. Swope WC, Pitera JW, Suits F, Pitman M, Eleftheriou M, Fitch BG, Germain RS, Rayshubski A, Ward TJC, Zhestkov Y, Zhou R. *J Phys Chem B.* 2004; 108(21):6582–6594.
89. Weinan E, Ren WQ, Vanden-Eijnden E. *Phys Rev B.* 2002; 66(5):052301.
90. Torrie GM, Valleau JP. *J Comput Phys.* 1977; 23(2):187–199.
91. Henkelman G, Uberuaga BP, Jonsson H. *J Chem Phys.* 2000; 113(22):9901–9904.

Figure 1.

A representative systematic search tree from 4 PC modes. The reranked PC modes are rendered in dark red (Mode #1 +), red (Mode #1 -), orange (Mode #2 +), yellow (Mode #2 -), green (Mode #3 +), cyan (Mode #3 -), blue (Mode #4 +), and purple (Mode #4 -). Note that PSIM reranks the modes as detailed in the text, and this step is not shown in the figure. The search performs in a systematic fashion so that the overall direction in the figure is from top to bottom and from left to right. One set of colored tree branches is called a search branch (gray dashed boxes). In a branch, PC modes are either excluded (dashed) or in use (solid). The black solid/dashed lines indicate the workflow in the PSIM method where dashed lines mean that some branches in between are omitted for simplicity. After the distortion along one mode/direction in a branch, the search starts in the child branch from the current distorted conformation. If a branch is a *Dead End*, the search continues from its parent branch. After *Output*, the search continues from the root branch. After *Motion Test*, the search either continues normally (passed) or continues from the branch after the previous passed *Motion Test* (failed). For clearer illustration, not all branches are drawn, and some child branches are represented by gray dashed arrows to indicate searches continue along this branch. Modes end with a red dot if the acceptance test fails immediately after only one step distortion. Note that modes in a branch after an output or a failed motion test are not considered in the future search.

Figure 2.

Overview of the algorithm in a search branch used in PSIM. The tree diagram presenting search and termination criteria for branches is in Figure 1. The two parallelograms show the Parent and Child branches to this current branch. Rounded rectangles represent the conformations, which are new conformations generated by distortions or the input from the parent branch. Rectangles represent the operation modules. Ellipses represent a decision or a test module. PSIM enters a search branch represented by the flowchart from an initial conformation (in the case of the root branch) or a conformation produced by its Parent Branch (in the case of any child branch). It starts the loop over PC modes (over i) and postpones the PC mode loop when a direction of a PC mode is done, and it enters the child branch. After its child branch (which follows the same flowchart) is completed, it continues the loop over i in the current branch through " $i = i+1$ ". The current branch exits when " i is larger than Number of Modes". After outputting a trajectory, "Output and Back to Root", the search returns to the root branch to continue the search from the next mode (see Figure 1 for action after *Output*). If the motion test fails, the search returns to the " $i = i+1$ " mode in the branch which last passed the motion test to continue the search (see Figure 1 for action after *Motion Test Failed*). A systematic search run is accomplished when all modes/branches have been visited (the root branch is done).

Figure 3.

Classical internal coordinate representation (Z-matrix) for a small molecule. Atom 1 is at the origin; atom 2 is defined by a bond length (b_2) and placed on the z -axis; and atom 3 is defined by a bond length (b_3), bond angle (a_3) and placed on the x - z plane. Six external coordinates are eliminated, and a molecule of N atoms has $3N-6$ internal coordinates. Atoms 1–3 are termed root atoms for this molecule. Atoms $i > 3$ are defined by (b_i, a_i, t_i), where t are torsion angles. For example, atom 4 is presented by (b_4, a_4, t_4). The dashed light gray line shows the molecule after rotating the bond between Atoms 2 and 3 (t_4) 180 deg.

Figure 4.

An example of fragmentation of HIVp for multilayer internal coordinate definition. Each monomer is considered one group and divided into 21 fragments. The root fragment on monomer 1 is in blue. The other fragments on monomer 1 are in dark and light green, and these fragments are defined as the root fragments. In monomer 2, the root fragment is in red, and the other fragments connected to the root fragment are in light and dark magenta. The root fragments of the two monomers are connected to each other. All fragments including the two root fragments are connected through six degrees of freedom shown in Figure SI 2. For illustration purposes, the root fragment and three other fragments on monomer 1 are circled. The connections between them are represented by thick gray lines.

Figure 5.

Structure of tryptophan synthase (TRPS) in the complex with the α -subunit of glyceraldehyde-3-phosphate (G3P). Loops 6 and 2 are in orange and red, respectively. Bound G3P is in green. Three residues in close contact with G3P are labeled: S234 forms a hydrogen bond with the phosphate group of the ligand, and nonpolar interactions between F211, T182, and G3P hold G3P in the binding site.

Figure 6.

Motions of HIVp identified by the first PC mode. The PC mode was computed by the covariance matrix using a) the α -carbon in Cartesian coordinates; b) dihedrals in multilayer internal coordinates; and c) dihedrals in classical internal coordinates (Z-matrix). For visualization purposes, the amplitudes of the motions were scaled by 2.67 in a), 1.0 in b), and 0.2 in c). In c), the artifact introduced by the accumulated dependence on their covalently bonded predecessors for defining dihedrals in the classical internal coordinates results in a rigid left monomer, whereas the right monomer moves too much.

Figure 7.

A structure of HIVp and RMSF of Ca atoms computed by a PC motion. Top: the structure of HIVp. Bottom: the comparison of RMSF of motions presented by the first PC mode computed by using the covariance matrix with multilayer internal coordinate definition (blue), classical internal coordinates (red), and Cartesian coordinates (green). The colors on the x axis of the bottom figure are identical to structural regions in the top figure. Residue numbers shown in different structural/color are as follows: the dimer interface in blue/yellow: residues 1–4 (100–103) and 96–99 (195–198); fulcrums in red: residues 10–23 (109–122); flap elbows in orange: residues 37–42 (136–141); flaps in green: residues 43–58, (142–157); flap tips in ochre: residues 49–52 (148–151); cantilevers in cyan: residues 59–75 (158–174); loop near flaps in purple: residues 76–84 (175–183). Notably, HIVp is a homodimer, and the residues in the other monomer are shown in the bracket.

Figure 8.

Four conformational transition paths of alanine dipeptide found by PSIM. The search started from Minimum A, and four distortions using the positive and negative directions of PC modes #1 and #2 led the transitions to different minima shown in 1) to 4). The left column of each path shows the conformational change of the original path using PC modes #1 (Green) and #2 (Yellow). The middle column shows the search path guided by PC mode #1 in green and PC mode #2 in yellow. The right column shows the postoptimized paths (white bordered dots) on the potential energy surface (PES) of alanine dipeptide from PSIM. The original paths are shown in black bordered dots for comparison.

Figure 9.

Comparison between dissociation events of the β -cyclodextrin-2-naphthyl ethanol complex obtained from MD and PSIM. Trajectories are classified by number of glucose ring flippings in β -cyclodextrin: a) no glucose ring flipping, b) flipping of one ring, and c) flipping of multiple rings. The bound and free conformations of β -cyclodextrin are on the left and middle of the figure, with the ligand removed from the bound complexes for clearer visualization. The plots of 7 β -cyclodextrin conformation fingerprint angles along trajectories are on the right. Corresponding glucose rings are labeled on the conformations with the same color as in the plots. The bound state conformation (rectangle), free state conformation (hexagon), and moment of dissociation (triangle) are indicated on the plots.

Figure 10.

One dissociation pathway of the cryptophane–tetramethylammonium complex found by PSIM, Hopping Minima, and steered MD. The dissociation paths are represented by yellow traces. In PSIM and steered MD, the dissociation paths start from the bound state and end with the ligand outside the binding pocket. Hopping Minima connected four local energy minima to build a binding/unbinding path.

Figure 11.

Snapshots of a G3P dissociation pathway from TRPS found by PSIM. Six frames, #1, #2001, #4001, #6001, #8001, and #10001, are shown here from a trajectory of 11,472 frames. G3P is presented as an enlarged green sphere, and the tryptophan α -subunit is shown in gray with loop 6 highlighted in orange. Important residues illustrated in Figure 5 are also shown in this figure. In the dissociation process, the protein maintains its native conformation except for opening loop 6 to release G3P.