Published in final edited form as:

J Comput Chem. 2011 January 30; 32(2): 196–209. doi:10.1002/jcc.21564.

Computing ensembles of transitions from stable states: Dynamic Importance Sampling

Juan R. Perilla*,

Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

Oliver Beckstein,

Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK and Department of Physiology, Johns Hopkins University, School of Medicine, Baltimore, Maryland 21205

Elizabeth J. Denning,

Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine Baltimore, Maryland 21205

Thomas B. Woolf

Department of Physiology and Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

Abstract

There is an increasing dataset of solved biomolecular structures in more than one conformation and increasing evidence that large scale conformational change is critical for biomolecular function. In this paper we present our implementation of a dynamic importance sampling (DIMS) algorithm that is directed towards improving our understanding of important intermediate states between experimentally defined starting and ending points. This complements traditional molecular dynamics methods where most of the sampling time is spent in the stable free energy wells defined by these initial and final points. As such, the algorithm creates a candidate set of transitions that provide insights for the much slower and probably most important, functionally relevant degrees of freedom. The method is implemented in the program CHARMM and is tested on six systems of growing size and complexity. These systems, the folding of Protein A and of Protein G, the conformational changes in the calcium sensor S100A6, the glucose-galactose binding protein, maltodextrin and lactoferrin, are also compared against other approaches that have been suggested in the literature. The results suggest good sampling on a diverse set of intermediates for all six systems with an ability to control the bias and thus to sample distributions of trajectories for the analysis of intermediate states.

I. INTRODUCTION

As the database of solved tertiary protein structures continues to grow, molecular understanding of function and our ability to predict functionally important regions of a

*.jrperillaj@jhu.edu.

protein structure increases. This knowledge ultimately will depend on an ability to predict and to understand conformational change between different functionally important states. Computational methods can greatly aid this research by providing testable routes for the molecular conformational change and by making estimates of any obligatory intermediates that must be passed through for the conformational change to occur.

Several methods, over many years, have been proposed for calculating transition pathways. Probably the earliest is the use of minimum energy, adiabatic, pathways. The work of Fisher[1, 2], and Elber et al. [3] may serve as examples of this research direction [4–9]. In their approach a starting point is first minimized to a low gradient RMS on the force and then minimum energy pathways are explored to move from one conformation to another. Similar ideas have also been proposed for a type of low-energy Brownian walk to find a minimum energy path[10]. Modifications to this basic idea have been explored by many groups. Perhaps the most currently well-known example is the transition path sampling methods developed by the Chandler group [11].

In transition path sampling, based on ideas developed by Pratt[12] and on work of Pratt and Chandler[13], an initial path estimate is made to connect the beginning and ending states. From this initial path a series of possible candidate improved paths are sampled in the Monte Carlo space generated by random moves from the initial path. The method has been explored and expanded from its initial foundations by the Bolhuis[14] and the Dellago groups[11]. In particular there has been considerable progress in identifying the class of perturbations from the original candidate path that will most likely lead to the starting and ending points rather than diverging.

Alternative pathway finding algorithms have worked with ideas developed by the Elber group[3]. In general these approaches use a transform of the problem to the space of finding spatial coordinates that fit the discretized space. Examples of these are the MaxFlux formalism [15], the string methods developed by the Vanden-Eijnden group[16], and the elastic rubber band approach first proposed by Jónsson et al[17, 18]. The class of mathematical optimization that is used depends on solutions to diffusion equations and asks for the most optimal arrangement of conformational steps along a fixed set of time points. In this way the problem is transformed from one of dynamics to one of sampling on intermediate conformations that sample well on the underlying energy surface. A problem with this approach is that entropy contributions to the free energy of the conformational change are not well sampled.

Another class of methods has been based on biased selection from the dynamics. This has been implemented in different ways, and a survey of some of the approaches and their relative advantages and disadvantages was assessed by the Post group [19]. For example, in biased molecular dynamics [20, 21] a one-sided potential is added to a standard MD simulation. This bias favors moves towards a target structure and penalizes moves away and so attempts to gently perturb the dynamics towards a transition. This is somewhat similar to meta-dynamics [22] in that the potential is adjusted during the simulation and that it aims at creating a flat distribution of states between the initial and final points. A related concept is that of weighted ensemble Brownian (WEB) dynamics [23] where a set of equally weighted

conformations is iterated forward in time to create a set of Brownian walkers that are eventually evenly distributed over the barrier crossing space. One advantage of the WEB dynamics is that the relative probability of a particular transition path is determined during the trajectory production. The Zuckerman group [24] has shown that this method can produce high quality transitions in their test systems.

In this work we employ dynamic importance sampling (DIMS) in combination with MD simulations to sample macromolecular transitions. DIMS goes back to ideas introduced by Wagner[25] for variance reduction on trajectories defined by stochastic differential equations. General importance sampling addresses the problem of how to simulate rare events such as transitions between two states x^A and x^B , of a system. In statistical mechanics the average $\langle g \rangle$ of an observable $g(x)$ determines many experimentally measurable events for the system. Here x characterizes the state of the system and is distributed according to a distribution $Q(x)$; if x is a point in configuration space and Q is the Boltzmann factor then $\langle g \rangle$ would be a canonical average. In importance sampling, states x are generated according to a new distribution $D(x)$, which typically includes a bias to sample rare (*important*) states preferentially compared to Q . When both Q and D are known, the desired average is

$$\langle g \rangle = \int dx g(x) Q(x) = \int [D(x) dx] \frac{g(x) Q(x)}{D(x)}. \quad (1)$$

$\langle g \rangle$ is computed by generating an ensemble S_D of x values and averaging gQ/D ; because sampling was performed under the biased distribution, the values x are now distributed according to D . When the biased distribution D improves the sampling of rare, but important events for gQ this becomes *importance sampling*. With well designed importance sampling distributions, the addition of the bias can dramatically improve the efficiency of sampling.

DIMS [26–28] applies a bias that depends on the current, evolving state of the system x . Application of a known, and time-evolving, bias to individual atoms (possibly a subset of the whole) helps to guide the evolution of a system starting in x^A towards trajectories that end in x^B . A probability of generating the biased path is computed along the trajectory. The general framework can be applied to the calculation of reaction rates and to simple quantitative descriptions of finite-temperature, average dynamic paths [27]. The most common way of creating conformational changes for a large biomolecule involves the determination of a reaction coordinate which is often not easy to define in a low dimensional subspace. One advantage of DIMS is that no reaction coordinate is needed for the transitions to be computed.

In this paper we present two algorithms that generate independent trajectories without previously defining a reaction coordinate. The *soft ratcheting* algorithm[29] generates transitions between states using a stochastic approach. A second method, the normal mode biasing algorithm generates transitions using information determined from the second derivative matrix. In the earlier papers, DIMS was introduced and successfully applied to a set of small systems that included butane [30, 31] and an all-atom model of the alanine dipeptide [28, 29]. We successfully implement and apply these algorithms to generate

intermediate states for systems of different sizes and complexities: the folding of Protein A and Protein G, calcium sensor S100A6, glucose-galactose binding protein, maltodextrin and lactoferrin.

II. SIMULATION METHODS

All simulations were performed using CHARMM[32, 33] with the analytical continuum electrostatics (ACE2)[34, 35] and the CHARMM22 force field[36, 37]. The Generalized Born implicit solvent model was used for efficiency, but we note that the same DIMS formulation also works with explicit solvents. Depending on the system different set of atoms may be selected for biasing e.g. the solvent, critical structural waters, or subsets of atoms from the protein. The structures were downloaded from the PDB data bank. The algorithms described here have been implemented in the c35b2 version of CHARMM.

A. Probability estimates

Dynamic Importance Sampling (DIMS) uses a biasing with correction approach to improve the sampling efficiency of rare events. To apply this approach to the calculation of rate constants requires an estimate of relative probabilities. Specifically, the probability score used in DIMS calculations is the ratio of two quantities[26–28]: the probability that the given trajectory would occur in an unbiased simulation; and the probability that the given trajectory was produced in the biased simulation. Therefore at each time step a single step probability score is computed as the ratio of two probabilities,

$$R_i = \frac{T_Q(x_i|x_{i-1})}{T_D(x'_i|x_{i-1})}, \quad (2)$$

where x'_i is the biased state, T_Q is the unbiased transition probability, and T_D is the biased transition probability which depends on the biasing scheme (such as the soft-ratcheting or normal-modes biasing scheme discussed below). The index i runs over the length of the transition L . The total score is computed for every trajectory by the product of individual transition probabilities

$$S = \prod_i^L R_i. \quad (3)$$

Another way to assess the likelihood of a transition, is to use the Onsager-Machlup action (or OM score). The OM score has the functional form[38, 39],

$$S[\mathbf{X}(t)] = \frac{1}{2\eta} \sum_{i=0}^L \left(\eta \mathbf{M} \left(\frac{\mathbf{X}_i - \mathbf{X}_{i-1}}{\Delta t} \right) - \mathbf{F}(\mathbf{X}_i) \right)^2 \cdot \Delta t, \quad (4)$$

where $\mathbf{X}_j$ are the spatial coordinates of the system, $\mathbf{M}$ is the masses matrix, $\mathbf{F}(\mathbf{X}_j) = -\nabla U(\mathbf{X}_j)$ is the force from the force field, η is a friction coefficient, t is the integration step, and the sum runs over the length of the trajectory L . In general equation (4) reflects the balance between the velocity along the transition and the force felt. Equation (4) can be interpreted as *the amount of noise* required to generate a particular transition, or the likelihood that a particular transition is generated by equation (5) (similar to the DIMS score[28]).

B. Generating dynamics and creating stochastic samples

In our description of the system, we use an implicit solvent and study the time evolution under a Brownian walk. We first describe the problem for a generic system under the influence of noise and an external force. That is, for a mass m particle with coordinates $\mathbf{x}$, under an external force, $\mathbf{f}(\mathbf{x}) = -\nabla U(\mathbf{x})$, the motion is governed by:

$$\frac{d\mathbf{x}}{dt} = \frac{1}{m\gamma}\mathbf{f} + \mathbf{R}(t), \quad (5)$$

$$\langle \mathbf{R}(t) \rangle = 0 \quad \langle \mathbf{R}(t)\mathbf{R}(t') \rangle = \left(2\frac{k_B T}{m\gamma}\right)\delta(t - t'). \quad (6)$$

where the term $\mathbf{R}$ is a family of zero-mean uncorrelated noise processes of intensity $2\frac{k_B T}{m\gamma}$, and γ is an arbitrary constant. In the case where γ is the collision frequency of a particle moving in a medium, equation (5) describes Langevin dynamics.

C. Types of Biases used in DIMS

In this work we are considering two biasing schemes: *soft ratcheting* and a normal-mode based bias. The soft-ratcheting approach showed success in small test systems[29]. In general outline, this approach uses a pre-defined progress variable and at each time step asks whether the motion is towards or away from the target state. If a trial molecular dynamics step is towards the target then the motion is accepted. A motion away from the target is only accepted with a certain probability, and this probability decreases with the size of the movement away from the target. This is similar to a Brownian ratchet system with the general direction of the random walk being towards the target[40]. Because the DIMS[28] method keeps track of both biased and unbiased moves at each time step, the size of the perturbation can be assessed and a relative probability score assigned to each sampled trajectory.

In addition to the simple and robust soft-ratcheting algorithm we also investigated a new biasing scheme based on normal modes. The basic idea is that the lowest frequency normal modes encode large scale protein movements. We can use these modes to guide the transition towards the target state. In practice the normal modes are calculated along a transition at intervals and a biasing force is constructed only based on those modes that

move the system towards the target, as indicated by the change in progress variable that is introduced below.

1. Progress variable—In order to *bias* the transitions we introduce a progress variable. Let $d(x, y)$ be a metric, which allows us to define a distance between any two states x^A and x^B of a protein. For example, we can use root mean square distance (RMSD), as defined by:

$$d_{RMSD}(x^A, x^B) = \frac{1}{M} \sqrt{\sum_{i \in \text{atoms}} m_i (x_i^A - x_i^B)^2}. \quad (7)$$

This allows us to define a one-dimensional change in distance by

$$\Delta\varphi(\mathbf{x}_{n-1} \rightarrow \mathbf{x}_n) = d(\mathbf{x}_n, \mathbf{x}^B) - d(\mathbf{x}_{n-1}, \mathbf{x}^B). \quad (8)$$

Equation (8) provides a metric for assessing whether a move is towards or away from a particular state. Note that this metric $d(x, x)$ is not restricted to the functional form in equation (7) as we discuss below.

2. Soft-ratcheting biasing—Stochastic ratchet dynamics have been widely studied in several transport phenomena problems [40–45]. In particular, thermal ratchets driven by noise have been used in the calculation of activation rates[44] and to the study of current in periodic structures [42]. Usually a ratchet is modeled as a periodic potential $V(x)$ with a broken parity symmetry [40] for example, a sawtooth potential, together with a Langevin equation similar to equation (5). In contrast, the soft-ratcheting algorithm (SRA) proposed by Zuckerman and Woolf[29], to generate transitions between known states of a protein, uses a ratchet-like probability acceptance filter (Metropolis-like criterion) p_{acc} . The dynamics is not limited to the Langevin equation, and can be used with explicit and coarse grained models. In this approach a trial molecular dynamics step is generated according to Equation (5). The MD step has a probability acceptance given by:

$$p_{acc}(\Delta\varphi) = \begin{cases} 1 & \text{if } \Delta\varphi \leq 0 \\ \exp[-|\Delta\varphi/\varphi_0|^2] & \text{if } \Delta\varphi > 0 \end{cases} \quad (9)$$

where φ_0 is a parameter which controls the width of the exponential function, and therefore the backwards decay. Here $\varphi \leq 0$ means that the proposed trial step is moving towards the target; conversely, $\varphi > 0$ indicates a motion away from the target. The *softness* of soft-ratcheting is due to this (backwards) decay: where backwards steps can be accepted as part of the longer term motion towards the target. If the MD step is not accepted, a new trial MD step is generated. In this way the system is biased, in a non-monotonical, ratchet-like manner. The net displacement is towards the target.

The probability density that SRA will generate a given φ increment, b_{φ} , is proportional to the product of the unbiased probability of generating the increment, p_{φ} and the acceptance probability, p_{acc} [29],

$$b_{\Delta\varphi}(\Delta\varphi) = \frac{1}{N} p_{\Delta\varphi}(\Delta\varphi) p_{acc}(\Delta\varphi), \quad (10)$$

where N is a normalization factor, given by the fraction of steps initiated at x_{n-1} and that would be accepted by the SRA. Then the single-step score, as in equation (2), for SRA is written:

$$R_i = \frac{p_{\Delta\varphi}(\Delta\varphi)}{b_{\Delta\varphi}(\Delta\varphi)} = \frac{N}{p_{acc}(\Delta\varphi)}. \quad (11)$$

and the total score is the product of individual steps, as in equation (3).

3. Normal Mode Biasing—High frequency modes are localized and describe local motions, whereas low frequency modes describe global collective motions and are associated with functionally relevant conformational changes [46]. This has motivated techniques such as Monte Carlo NM following [47] (MC-NMF) and the eigenvector following approach [48–51]. These techniques use the information from a single mode, the lowest frequency mode, in order to displace the structure from one state to another. There are also other techniques based on elastic network models[52] that generate transitions for macromolecular systems using information from the normal mode set of the system. Following or projecting multiple normal modes to generate a pathway has also been studied [53, 54]. These studies showed that although some systems presented single-mode dominance this was not true in general and some systems would, in fact, require multiple modes to generate a feasible transition. We designed an algorithm based on normal modes that generates transitions between a given initial and final state: $\mathbf{x}^A$ and $\mathbf{x}^B$, respectively. The algorithm projects multiple normal modes and uses all-atom molecular dynamics. The algorithm is as follows:

1. An unbiased step is computed $\mathbf{x}_{i+1} = \mathbf{x}_i + (\mathbf{f}_i/m\gamma) \Delta t + \mathbf{x}_R$.
2. The set of normal modes $\{\zeta(\mathbf{x}_i)\}$ is computed using the rotation-translation block method (RTB)[55].
3. A filter looks for the best N_{best} modes and tries combinations of $1, 2, \dots, N_{comb}$ modes. The set of modes $\{\eta\} \subset \zeta$ is the combination of modes -or a single mode- that *favors the most* the system evolution towards the target according to equation (7).
4. The modes selected $\{\eta\}$, are scaled by α and then applied as a bias to obtain the biased step $\mathbf{x}'$:

$$\mathbf{x}'_{i+1} = \mathbf{x}_{i+1} + \alpha_i \eta. \quad (12)$$

Although modes could be computed at every step, we recompute them every N_{skip} steps and apply the bias for N_{biased} steps, hence there is a number $N_{\text{unbiased}} = N_{\text{skip}} - N_{\text{biased}}$ of unbiased steps. In order to slowly perturb the system, the scaling factor α has the functional form,

$$\alpha_i = \frac{1}{2}\beta \left(1 + \tanh \left(\frac{24}{N_{\text{biased}}} (i \bmod N_{\text{biased}}) - 4 \right) \right), \quad (13)$$

where β is an adjustable constant.

In contrast to other methods such as elastic network models[52, 54] we include both the information from all-atom MD and a bias from multiple modes. In order to increase the variety of trajectories we use two strategies: 1) we use Langevin dynamics that adds a stochastic component to the simulation, and 2) we employ a “self avoidance” scheme. The latter ignores modes already used in previous trajectories within a N_{avoid} window and forces the algorithm to use different modes or combinations of modes.

The advantage of generating candidate moves based on normal modes is that we can select for more global, cooperative motions, and encourage the more complex displacements that most need to be sampled on a longer time-scale in order for the motion to occur towards the target. The construction of sample mode-based biasing enables a selection of a biased walk towards the target that is more likely to be similar to the low-probability events that underlie a transition.

The transition probability in equation (2) is chosen from the Boltzmann distribution of energy changes due to the NM bias. Thus, the single step probability score becomes the ratio of energy changes due to unbiased and biased movements of sets of atoms[28],

$$R_i \approx \frac{e^{-\Delta E_Q/kT}}{e^{-\Delta E_D/kT}}, \quad (14)$$

where E_Q and E_D are the unbiased and biased potential energies respectively, T is the temperature and k is the Boltzmann constant. The score associated with a trajectory is the product of single steps as in equation (3).

III. RESULTS AND DISCUSSION

For the testing of the CHARMM implementation of DIMS, we focused on protein A (60 residues), protein G (61 residues), human calcium sensor S100A6 (90 residues), glucose-galactose binding protein (GGBP) (305 residues), maltodextrin (370 residues), and

lactoferrin (691 residues). These six systems (Figure 1) span a range of sizes and complexities, from protein folding to hinge bending. For GGBP, S1006A, maltodextrin, and lactoferrin x-ray crystal structures are available in two different states that act as the basis for our choice of initial and final conformations in the DIMS calculations. For the Protein A and Protein G, representative states for the denatured protein were generated by high temperature molecular dynamics calculations in an implicit solvent.

The set of parameters used to generate the transitions using SRA and NMB are shown in Table I and II respectively. We show the generation of trajectories that do not evolve from one another and are statistically independent. For this we evaluate our algorithms and use different analysis to show trajectory diversity using different metrics: pairwise RMSD, native contacts analysis, hinge-opening angles, etc. We also study the effects of the bias on the dynamics of the systems.

A. Metric selection

For the metric in equation (8), we initially used the root mean square distance (RMSD) equation (7) as the *progress variable*. By construction coordinate space RMSD is dependent on the alignment between structures and poor structural alignment will lead to a RMSD that also includes contributions from the overall translational and rotational degrees of freedom. Therefore it is necessary to re-align the system as the transition is calculated; in order to not affect the dynamics one has to align the target structure to the structure at the current time step. The frequency at which the structure is re-aligned should be selected in such a way that no artifacts are introduced in the dynamics. Our tests showed that high frequencies can introduce high frequency jitter into trajectories whereas low frequencies lead to abrupt changes in bias. In particular if transitions are unable to reach the target for weak biases, it is an indicator that the alignment settings are not optimal. Selection of ρ_0 is a function of the system size and sample values are presented in Table I. All systems worked well with RMSD and overall it proved to be straightforward, robust, and flexible enough to allow different pathways to be sampled as we show below.

B. Protein-A and Protein-G Folding

Initial testing of the algorithm used two small peptides: Protein A (PDBID:1BDC[56]), and Protein G (PDBID:1IGD [57]). These consist respectively of a helix and a β -sheet (Figure 1a,b). For both systems we used a similar protocol. Starting with the crystal structure from the PDB Bank, we created a set of unfolded states by heating at 2000 K for 10 ns using Langevin dynamics in CHARMM. The transitions from *unfolded* to the crystal structure for protein A used the normal modes approach, and the transitions for protein G were performed with the soft ratcheting approach. For protein G we created a set of 60 SRA trajectories going from a single unfolded state towards the crystal structure(Figure 2), and also sets of ten SRA trajectories going from six different unfolded states towards the same crystal structure. It is worth noting that the normal mode approach takes longer to compute a single transition than SR, hence the larger sets of trajectories using SR compared to those from NMB.

A comparison to targeted molecular dynamics (TMD)[58] provides some insights into the nature of the sampling provided by DIMS. While the parameters for TMD were not extensively optimized, they were chosen to be representative of the types of transitions generated by TMD within the CHARMM program. Calculations in these two frameworks are shown in Figure 3 for protein A. Note that the RMS starting and ending points are identical for both methods. The TMD method has large energy changes as the transition is generated in moving from the unfolded to the folded state. In contrast, the DIMS method has much lower energy changes, while also creating a trajectory connecting the two states. This illustrates a real strength of the DIMS method: the steps along the transition are designed to be as close as possible to conformational changes that would be natively seen in a molecular dynamics trajectory with sufficient time to sample on the transitions. The transitions generated by DIMS are assigned probabilities (see later section below) that relate to their degree of likelihood for occurring without a biased stochastic walk. As in previous DIMS studies [28] we expect that with sufficient sampling the set of transitions generated by DIMS will provide a representative set of conformations that can be used to estimate kinetics, relative free energies and pathways between the states.

A way to further explore the ability to generate multiple transitions between states is by a histogram analysis shown in Figures 4 and 5 for protein A and protein G respectively. In this case the set of trajectories is projected onto the two dimensions of the number of hydrogen bonds and the percentage of native contacts. This analysis is similar to the approach frequently used in the protein folding simulation community. The plots show a diversity of sampled conformational regions between the two states. It suggests certain points in hydrogen bonding/native contact space that most trajectories are likely to pass through and other regions that are more briefly visited. The true nature of the projected surface will depend on the underlying relative free energy, but these two figures show that there is a diversity in the sampled trajectories between the two states.

These histograms clearly show that a distribution of states is being sampled. We used this set to further compare the diversity between trajectories sampled with different parameters. Using contact analysis we explored the difference between two strategies to generate diversity. This involved running several ($N=60$) trajectories between two fixed initial and final structures. Our analysis shows how this compares to generating fewer trajectories ($N=10$) between an ensemble of six different initial states all going to the same final target (Figure 5). For these two strategies, multiple and single initial states, the target structure is the same. The histograms show that the shapes of the distributions are different. The histogram for multiple initial states Figure 5a) is clearly broader than the one for a single initial state Figure 5b). This shows that the number of distinct intermediate states (states in between the origin and the target) is higher for multiple initial states than for the fixed starting and ending points. This suggests that a larger number of intermediate points is reached with sampled initial points from the starting point well. This type of framework, building on equilibrium molecular dynamics runs at both the initial and the final states and then sampling on these conformations for the runs, is our main framework for sampling in DIMS. In order to focus our results on the sampling efficiency of DIMS itself, the rest of our simulations use a single initial and final state.

Spatial differences can also be studied by computing atomic distances between structures. A comparison based on the RMS values (Figure 6(a)) between conformations is used to analyze the trajectories. The results suggest that each trajectory is sampling a different set of conformations. Using Dijkstra's shortest path algorithm [59], we obtain the set of points between the initial and the target structure with minimum RMS distance (shown in white in Figure 6(a)). For a set of eight trajectories, the RMS distances along the Dijkstra path on the RMSD surface are all high in the intermediate regions ($\approx 3\text{--}5\text{ \AA}$) and are zero for the initial state and low near the target ($< 1\text{ \AA}$) due to our cutoff of $1\text{ \AA}$. The distribution of OM-Scores in figure 7 also reinforces the finding that different sets of structures are being computed. This shows an important property created by DIMS trajectories: wide sampling of the conformational space between the initial and final states.

With the idea of increasing the number of escape routes from the initial structure that DIMS would sample we also tried using $RMSD(x) = d_{rmsd}(x_A, x) - d_{rmsd}(x_B, x)$ as the guiding metric. We observed transitions would move away from the initial structure and reach a mid-point where $RMSD(x) = 0$ and then proceed towards the target structure. Such conformations had a globular shape in which some secondary structure elements were formed but the overall protein structure was distorted. However, the ability to generate different pathways using this metric might provide a good starting point for path annealing techniques.

C. Calcium sensor S100A6

The function of the 90-residue S100A6 Calcium sensor involves a large conformational change between two known end states. It undergoes a transition that includes a major rotation of one of its helices (1.5rad) (Figure 1c). The RMS distance between the apo and the Ca^{2+} loaded structure is $5.0\text{ \AA}$ (PDBID: 1K9K and 1K9P)[60]. Although it is Ca^{2+} that induces the conformational change observed in S100A6 we do not include Ca^{2+} in our simulations. This approach is justified by the hypothesis that the apo protein can sample all conformations that are seen for the holo protein, albeit with differing probabilities. Such a population shift model for the binding events has been observed in other proteins that exhibit a similar range of motions on ligand binding [61–63]. In order to generate the transition we used the minimized crystal structures from the pdb. We generated a set of 100 transitions going from the apo to the Ca^{2+} loaded state using the normal mode bias together with the self avoidance algorithm and Langevin dynamics with parameters as specified in Table II.

Selecting different modes along the transition enhances sampling of different structures. Our method only selects those modes that favor displacement towards the target and that had not been used to generate previous transitions within a given window. Figure 8 shows how different modes are being selected along the transitions. Low frequency modes are being used while the RMS distance to the target is above $1.2\text{ \AA}$ (red colored line). This in turn suggests that with enough computing time eventually the set of modes and combinations that had been explored will span the entire spectra.

D. Glucose galactose Binding Protein

The structure of an open unbound glucose-galactose binding protein (GGBP) was crystallized by Borrock et al. (PDBID:2FW0)[64] at 1.55 Å. This system exhibits a 0.5 *rad* hinge opening motion from one state to the other (Figure 1d). The large change for the hinge domain between states and the existence of both open and closed structures without any sugar bound creates a good system for testing DIMS. This is a medium size protein (305 residues long) letting us test the algorithm's ability to generate an ensemble of trajectories in a reasonable amount of time. The RMS distance between the initial structure and the target is 9.9 Å. Using the soft ratcheting algorithm we generate an ensemble of 150 trajectories going from the open state towards the closed state, at 300K, using the parameters specified in Table I.

In order to assess the sampling diversity in this set we define an opening angle θ using the center of mass from the N- and C- terminal domains as well as the three-segment hinge that connects them. We compute the hinge opening angle θ for every structure in the set of trajectories, and build the probability density of finding a particular angle at a given time step (Figure 9(a)). The fact that at a given time step different values of θ have different probabilities is a clear indicator that different pathways are being explored.

Also we can explore the net force density that it is being applied on this angle by the bias. If a constant pulling force was applied we would expect to see a constant force or a distribution of forces centered at a positive value. However our results in Figure 9(b) show that for every time step the net force ranges from positive to negative and it is centered around zero. This suggests that the force applied along the angle is non-monotonical and the distribution is bell-shaped around zero. Since the average force applied along the angle is close to zero, the protein dynamics in the ensemble of DIMS transitions is close to the dynamics that would be observed in an ensemble of unbiased transitions.

E. Maltodextrin

Increasing in size (370 residues long) and complexity we apply DIMS to the maltodextrin binding protein. This system presents a hinge bending conformational change between the domains that regulates access to and from the binding site located between the two domains[65] (Figure 1e). It also has a second conformational change in which local rotameric changes of residues are observed [65]. Maltodextrin first binds the ligand producing a small conformational change, and then undergoes the large conformational change burying the sugar deep inside the two globular domains [65]. It also exists in equilibrium between the open and closed forms [66]. We use the structures with PDB IDs 1OMP [66] and 1ANF [65], which has an RMSD of 8.4 Å. We generated 50 trajectories using the soft-ratcheting algorithm going from the closed to the open state at 300 K. Maltose was removed from the structure and simulations performed in its absence. Once again we assumed a population shift model with the hypothesis that samples determined in the apo state are relevant to the ligand bound state [61–63].

Figure 10 shows sampling among opening angles during the transition. Given the complexity of the change there is a highly populated region at the end of the trajectory. With

the direction of our transitions, from closed to open, the residues involved in the second conformational change can re-arrange either during the transition or at the end (leading to several types of transitions). This also has the effect of making the trajectory look as if it had different kinetics, varying between the case where the re-arrangement occurs at the end (highly populated region near the open state) versus the large conformational change. Figure 11 shows the total root mean squared distance along the transition for every residue for the trajectories that populate the opening angle region in 10, in black and red for the initial and final part of the transition respectively. Residues in the domain involved with closing (labeled by green diamonds) show more displacement in the first part of the transition. Residues in the active site (labeled by blue bullets) show more displacement at the end of the transition. In particular Arg-66 and Tyr-341 show the largest conformational changes [65, 66], along with the largest distances to reach the end of the transition (Figure 11).

The probability of accepting a trial move (equation (9)) from a Gaussian distribution is also a good indicator for the quality of the transition. If too many trials are rejected this means that the strength of the bias is too strong, on the other hand if few trials are rejected it means that the transition could have occurred in an unbiased simulation. Figure 12 shows that although two transitions have poor p_{acc} values, most of the generated transitions have relatively high probability of acceptance ($p_{acc} \approx 0.8$). This in turn suggests that most transitions were weakly biased and likely to occur in an unbiased simulation.

F. Lactoferrin

Lactoferrin is the largest of the systems we studied using our method (691 residues long). It has a two-lobe, four-domain structure characteristic of the transferrin family, each lobe has an iron site [67]. Lactoferrin presents a large conformational change which is responsible for the ability to release iron under appropriate conditions (Figure 1f). It is thought that the closed apo state is accessible even without iron bound; hence the iron binding mechanism proposed suggests that iron binds first to one domain of the open form and then once the closed state is sampled (by thermal fluctuations) iron interacts with the site clamping the domains shut [68]. There are several structures available, from which we selected the structures with PDB ids 1LFH and 1LFG. The RMSD between the two structures is 9.0 Å and it shows a 0.52 rad hinge motion between the two domains of the N lobe. Using DIMS we generate 32 trajectories going from the closed to the open state at 300 K.

For further comparison with other methods we examine transitions from three other methods: The Database of Macromolecular Movements [69], the intermediate states from a coarsened grained Elastic Network Model method by Kim et al. [70][52], and three transitions using the TMD method implemented in CHARMM [71], with pulling constants 1×10^{-2} Å, 1×10^{-3} Å and 1×10^{-4} Å respectively. We define an opening angle θ using the center of mass from the N1- and N2- lobes of the N domain, as well as the two-segment hinge that connects them. We compute the opening angle θ for every structure in the set of trajectories for each method. The results for DIMS (Figure 13 and 14) show a distribution of angles, and as with the previous systems a net force along the opening angle centered about zero. This clearly indicates that different states are being sampled along the transition without forcing the conformation.

The lactoferrin simulations illustrate how the complexity of the transitions increases with the system size. Methods that generate intermediate states by spatially biasing the system tend to mix the time scales for certain events. For example, in TMD using a low value for the pulling constant makes the global motions too fast for an optimal adjustment of the local high frequency motions. This in turn leads to poor sampling of local motions and unphysical intermediate states (Figure 14). Also TMD tends to introduce strains along the transition that creates a series of unrealistic oscillations at the end of the trajectory (Figure 14). Furthermore given TMD's deterministic nature only one pathway is generated for each pair of initial and final states. The Molecular Motions Database generates transitions using a flat potential surface and transitions are purely geometrical. These intermediate states are quite similar to those from ENM due to the rigid body-like character of the transition. DIMS handles time scales in a different way, as it allows the local motions to be sampled along the transition at a distribution of times. Its stochastic formulation gives it the advantage of generating a range of transitions (Figure 14(a)). For the most optimized TMD set of parameters (Figure 14) we observe that the pulling velocity, and thus the progress along the trajectory, is similar to that of DIMS. Notice, however, that this does not mean that the methods are equivalent or that the pathways being sampled are similar.

G. Current Framework for the application of DIMS

This paper presents the current DIMS framework implemented in the c35b2 version of CHARMM. We believe that it will be an important tool for the generation and analysis of conformational intermediates in biomolecular transitions. At the same time we need to point out the limitations of the current framework and the working assumptions behind the approach. First, while our method can, in principle, be applied to both explicit solvent and implicit solvent settings, we have decided to present implicit solvent simulations in these results. There is nothing in the present implementation that prevents this from being used with explicit solvent.

It can be imagined that a series of DIMS calculations are applied to a particular set of transitions. In the most ideal case, these calculations would use a range of different starting points, differing amounts of explicit water and different order parameters. This would create the greatest degree of conformational sampling on the intermediate states. We found something similar in the current paper, where there is a greater diversity of states sampled on the intermediate surface with a diversity of starting points for the dynamics (Figure 5). Thus, the current best approach to using DIMS is a range of starting and ending states, a range of solvent conditions and a range of order parameters. In the best implementation of that framework, the data and the self-consistent equations would be solved under a variety of conditions to create the most likely estimate of the underlying free energy surface. A single calculation is not sufficient if the desired result is a strong connection to a physical measured event or a biomolecular function. The choice of progress variable, the conformations of the end points but also the treatment of the solvent, are likely to influence transitions in a system-specific manner as needs to be assessed on a case-by-case basis. As shown in the current work, the variability of DIMS transitions can be increased by varying parameters of the algorithms and especially by considering the equilibrium ensembles of the transition end points. To find optimal estimates for the intermediate states, the full set of candidate

conditions should be used and the full set of transitions evaluated collectively using the Onsager-Machlup score.

As a third framework point, it should be emphasized that the conformational sampling defined by the DIMS method can provide initial guidance for the underlying free energy surface, that defines the chemical physics of the transition. That said, it is expected that an ensemble of DIMS transitions will be less efficient than the use of method such as umbrella sampling so long as one or two degrees of freedom can be clearly defined as a reaction coordinate. It is also worth noting that a similar concept is in place with the use of the weighted histogram analysis method (WHAM) [72], where a range of windows are combined to create the best estimate for the underlying free energy surface. DIMS does then, provide some information on the higher-dimensional free energy surfaces that are not readily setup for explicit sampling. This certainly includes the surfaces defining realistic transition pathways and the identification of saddle points. For example, we found the initial transition paths useful to guide conformations for umbrella sampling in the case of AdK [73].

In our fourth framework point, we want to emphasize that kinetics is the long-term goal of these calculations. We believe that a converged estimate of kinetics can be obtained from the DIMS formulation [27]. This estimate requires knowledge of dwell times in both initial and final states, a determination of the transition state dividing surface, as well as sufficient sampling on the temporal distribution of barrier crossing events for the value to be meaningful. As we described in one of our first papers on DIMS, the rate is defined as the slope of the probability distribution relating a transition from a starting state to an ending state after different amounts of time. The choice of a correct dividing surface and the assumption of a smooth underlying free energy at the transition state are assumed in this definition, along with sufficient sampling for the slope to be meaningful [27]. Due to these sampling limitations, we have not emphasized the actual calculation of the kinetics, the relative free energies, or the average dynamical pathways defined by DIMS, though all of these can be defined by sufficient distributions of events for an ensemble set that may start to provide statistical confidence in conclusions related to experimental measurements. Ideally these kinetic connections would use DIMS calculations with a range of order parameters and a range of sampled starting and ending points.

We are currently working to further improve the range of order parameters that can be used for DIMS and examining the rate of convergence needed for a given level of statistical accuracy in a kinetics calculation. This will then help us to understand the connections between our biased sampling and the unbiased free energy surface. Thus, a final working assumption in our framework is that the dynamical states sampled on by DIMS are consistent with states that would have been obtained with a non-biased approach. The work presented here and elsewhere [73] suggests that indeed DIMS can give insights into the ensemble of transitions at a fraction of the computational cost of equilibrium simulations. In order to rigorously test the assumption that DIMS transitions and equilibrium transitions sample the same set of states one would generate a family of DIMS transitions, rank-order and weight them by Onsager-Machlup score and compare them to transitions that arise from equilibrium molecular dynamics calculations. Within the context of a much simpler system (butane [30]) the results of comparisons between unbiased and biased transitions are

encouraging – suggesting considerable improvement in sampling efficiency and no systematic biasing driving transitions away from unbiased, native-like, transitions. A full test of this assumption, will, however, require the collection of several thousand unbiased barrier crossing trajectories as well as an equal number of DIMS trajectories for a non-trivial biological system. We hope to compare DIMS in this way in the future, but such a comparison is beyond the scope of this paper.

IV. CONCLUSIONS

We have described our implementation of dynamic importance sampling (DIMS) within the program CHARMM. The approach enables enhanced sampling of intermediate conformations between two defined states. It does this by creating a set of biasing distributions that are used for improving the sampling of barrier crossing trajectories and then correcting for this bias. As such it is a dynamic implementation of importance sampling. With this new feature in CHARMM we are now able to sample on conformational transitions within a large range of biomolecules. This considerably extends the types of systems that can be approached with the DIMS formulation. In particular, the method is robust and applicable to a wide range of systems and problems with a simple progress variable such as RMSD. The bias does not apply large forces to the protein structure so that the transition can evolve naturally. In a related study we have also shown that DIMS transitions of the enzyme adenylate kinase sample biologically relevant conformations as verified by comparison to a large number of intermediate crystal structures [73]. By comparing our results to other existing methods we suggest that there are many advantages to the DIMS approach to sampling on conformational transitions. With the increasing size of structural databases involving more than one conformation of biomolecules and with the increasing importance of understanding large scale conformational change, the DIMS method may prove to be an important tool for helping to connect the details of molecular conformational change with measured biomolecule function.

References

- [1]. Fischer S, Chemical Physics Letters, 1992, 194(3), 252–261.
- [2]. Gruia AD; Bondar A-N; Smith JC and Fischer S, Structure, 2005, 13(4), 617–627. [PubMed: 15837200]
- [3]. Elber R and Karplus M, Chemical Physics Letters, 1987, 139(5), 375–380.
- [4]. Olender R and Elber R, Theochem - Journal of Molecular Structure, 1997, 398, 63–71.
- [5]. Czerminski R and Elber R, International Journal of Quantum Chemistry, 1990, (suppl.24), 167–186.
- [6]. Ulitsky A and Elber R, Journal of Chemical Physics, 1990, 92(2), 1510–1511.
- [7]. Czerminski R and Elber R, Proceedings of the National Academy of Sciences of the United States of America, 1989, 86(18), 6963–6967. [PubMed: 2780552]
- [8]. Czerminski R and Elber R, Journal of Chemical Physics, 1990, 92(9), 5580–5601.
- [9]. Choi C and Elber R, Journal of Chemical Physics, 1991, 94(1), 751–760.
- [10]. Berkowitz M; Morgan J and Mccammon J, Journal of Chemical Physics, 1983, 78(6), 3256–3261.
- [11]. Dellago C; Bolhuis PG; Csajka FS and Chandler D, The Journal of Chemical Physics, 1998, 108(5), 1964–1977.
- [12]. Pratt LR, The Journal of Chemical Physics, 1986, 85(9), 5045–5048.

- [13]. Chandler D and Pratt LR, The Journal of Chemical Physics, 1976, 65(8), 2925–2940.
- [14]. Bolhuis PG and Chandler D, The Journal of Chemical Physics, 2000, 113(18), 8154–8160.
- [15]. Huo S and Straub JE, The Journal of Chemical Physics, 1997, 107(13), 5000–5006.
- [16]. Ren W; Eijnden EV; Maragakis P and Weinan E, The Journal of Chemical Physics, 2005, 123(13), 134109+. [PubMed: 16223277]
- [17]. Jónsson H; Mills G and Jacobsen KW, Classical and Quantum Dynamics in Condensed Phase Simulations, 1998, pages 385–404.
- [18]. Crehuet R and Field MJ, The Journal of Chemical Physics, 2003, 118(21), 9563–9571.
- [19]. Huang H; Ozkirimli E and Post CB, Journal of Chemical Theory and Computation, 2009, 5(5), 1304–1314.
- [20]. Paci E; Vendruscolo M; Dobson CM and Karplus M, Journal of Molecular Biology, 2002, 324(1), 151–163. [PubMed: 12421565]
- [21]. Marchi M and Ballone P, The Journal of Chemical Physics, 1999, 110(8), 3697–3702.
- [22]. Laio A and Parrinello M, Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(20), 12562–12566. [PubMed: 12271136]
- [23]. Huber G, Biophysical Journal, 1996, 70(1), 97–110. [PubMed: 8770190]
- [24]. Zhang BW; Jasnow D and Zuckerman DM, Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(46), 18043–18048.
- [25]. Wagner W, Journal of Computational Physics, 1987, 71(1), 21–33.
- [26]. Woolf T, Chemical Physics Letters, 1998, 289(5–6), 433–441.
- [27]. Zuckerman DM and Woolf TB, The Journal of Chemical Physics, 1999, 111(21), 9475–9484.
- [28]. Jang H and Woolf TBB, Journal of computational chemistry, 2006, 27(11), 1136–1141. [PubMed: 16721720]
- [29]. Zuckerman DM and Woolf TB, 2002.
- [30]. Zuckerman DM and Woolf TB, The Journal of Chemical Physics, 2002, 116(6), 2586–2591.
- [31]. Zuckerman DM and Woolf TB, Physical Review E, 2000, 63(1), 016702+.
- [32]. Brooks B; Bruccoleri R; Olafson D; States D; Swaminathan S and Karplus M, Journal of Computational Chemistry, 1983, 4, 187–217.
- [33]. Brooks BRR; Brooks CLL; Mackerell ADD; Nilsson L; Petrella RJJ; Roux B; Won Y; Archontis G; Bartels C; Boresch S; Caflisch A; Caves L; Cui Q; Dinner ARR; Feig M; Fischer S; Gao J; Hodoscek M; Im W; Kuczera K; Lazaridis T; Ma J; Ovchinnikov V; Paci E; Pastor RWW; Post CBB; Pu JZZ; Schaefer M; Tidor B; Venable RMM; Woodcock HLL; Wu X; Yang W; York DMM and Karplus M, Journal of computational chemistry, 2009.
- [34]. Schaefer M and Karplus M, The Journal of Physical Chemistry, 1996, 100(5), 1578–1599.
- [35]. Schaefer M; Bartels C and Karplus M, Journal of Molecular Biology, 1998, 284(3), 835–848. [PubMed: 9826519]
- [36]. 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 and Karplus M, Journal of Physical Chemistry B, 1998, 102(18), 3586–3616.
- [37]. Mackerell AD; Feig M and Brooks CL, Journal of Computational Chemistry, 2004, 25(11), 1400–1415. [PubMed: 15185334]
- [38]. Eastman P; Niels, and Doniach S, The Journal of Chemical Physics, 2001, 114(8), 3823–3841.
- [39]. Onsager L and Machlup S, Physical Review, 1953, 91(6), 1505–1512.
- [40]. Magnasco MO, Physical Review Letters, 1993, 71(10), 1477–1481. [PubMed: 10054418]
- [41]. Bartussek R; Reimann P and Hänggi P, Physical Review Letters, 1996, 76(7), 1166–1169. [PubMed: 10061650]
- [42]. Czernik T; Kula J; Luczka J and Hänggi P, Physical Review E, 1997, 55(4), 4057–4066.
- [43]. Marchesoni F, Physical Review Letters, 1996, 77(12), 2364–2367. [PubMed: 10061935]
- [44]. Van den Broeck C and Hänggi P, Physical Review A, 1984, 30(5), 2730–2736.
- [45]. Derényi I and Vicsek T, Physical Review Letters, 1995, 75(3), 374–377. [PubMed: 10060005]

- [46]. Ma J, Structure, 2005, 13(3), 373–380. [PubMed: 15766538]
- [47]. Miloshevsky GV and Jordan PC, Structure, 2007, 15(12), 1654–1662. [PubMed: 18073114]
- [48]. Cerjan CJ and Miller WH, The Journal of Chemical Physics, 1981, 75(6), 2800–2806.
- [49]. Baker J, Journal of Computational Chemistry, 1986, 7(4), 385–395.
- [50]. Culot P; Dive G; Nguyen VH and Ghuysen JM, Theoretical Chemistry Accounts: theory, computation, and modeling (Theoretica Chimica Acta), 1992, 82(3), 189–205.
- [51]. Bofill JM and Anglada J. M., Theoretical Chemistry Accounts: theory, computation, and modeling (Theoretica Chimica Acta), 2001, 105(6), 463–472.
- [52]. Kim MK; Jernigan RL and Chirikjian GS, Biophysical Journal, 2002, 83(3), 1620–1630. [PubMed: 12202386]
- [53]. Zheng W and Doniach S, Proceedings of the National Academy of Sciences of the United States of America, 2003, 100(23), 13253–13258. [PubMed: 14585932]
- [54]. Zhenggo W and Brooks BR, Biophysical Journal, 2005, 88(5), 3109–3117. [PubMed: 15722427]
- [55]. Tama Florence, Florent Xavier Gadea OM and Sanejouand Y-H, Proteins: Structure, Function, and Genetics, 2000, 41(1), 1–7.
- [56]. Gouda H; Torigoe H; Saito A; Sato M; Arata Y and Shimada I, Biochemistry, 1992, 31(40), 9665–9672. [PubMed: 1390743]
- [57]. Derrick JP and Wigley DB, Journal of Molecular Biology, 1994, 243(5), 906–918. [PubMed: 7966308]
- [58]. Schlitter J; Engels M; Kruger P; Jacoby E and Wollmer A, Molecular Simulation, 1993, 10, 291–309.
- [59]. Dijkstra EW, Numerische Mathematik, 1959, 1(1), 269–271.
- [60]. Otterbein L; Kordowska J; Wittehoffmann C; Wang C and Dominguez R, Structure, 2002, 10(4), 557–567. [PubMed: 11937060]
- [61]. Swift RV and Mccammon AJ, Journal of the American Chemical Society, 0000, 0(0).
- [62]. Bahar I; Chennubhotla C and Tobi D, Current Opinion in Structural Biology, 2007, 17(6), 633–640. [PubMed: 18024008]
- [63]. Arora K and Brooks CL, Proceedings of the National Academy of Sciences of the United States of America, 2007, pages 0706443104+.
- [64]. Borrok MJ; Kiessling LL and Forest KT, Protein science, 2007, 16(6), 1032–1041. [PubMed: 17473016]
- [65]. Quioco FA; Spurlino JC and Rodseth LE, Structure, 1997, 5(8), 997–1015. [PubMed: 9309217]
- [66]. Sharff AJ; Rodseth LE; Spurlino JC and Quioco FA, Biochemistry, 1992, 31(44), 10657–10663. [PubMed: 1420181]
- [67]. Jameson GB; Anderson BF; Norris GE; Thomas DH and Baker EN, Acta Crystallographica Section D, 1998, 54(6 Part 2), 1319–1335.
- [68]. Baker EN; Baker HM and Kidd RD, Biochemistry and Cell Biology, 2002, 80, 27–34. [PubMed: 11908640]
- [69]. Echols N; Milburn D and Gerstein M, Nucleic Acids Res, 2003, 31(1), 478–482. [PubMed: 12520056]
- [70]. Kim MK; Chirikjian GS and Jernigan RL, Journal of Molecular Graphics and Modelling, 2002, 21(2), 151–160. [PubMed: 12398345]
- [71]. van der Vaart A and Karplus M, The Journal of Chemical Physics, 2005, 122(11), 114903+. [PubMed: 15836253]
- [72]. Kumar S; Rosenberg JM; Bouzida D; Swendsen RH and Kollman PA, Journal of Computational Chemistry, 1992, 13(8), 1011–1021.
- [73]. Beckstein O; Denning EJ; Perilla JR and Woolf TB, Journal of Molecular Biology, 2009, In Press, DOI: 10.1016/j.jmb.2009.09.009i.

Figure 1:

Initial and final conformational states for the systems: a) Protein A, b) Protein G, c) Calcium Sensor (S100A6), d) Glucose-galactose binding protein, e) Maltodextrin and, f) Lactoferrin.

Figure 2:
Protein-G folding transition going from unfolded to folded. Helix is in purple and, β -strand is in yellow. Hydrophobic residues are highlighted.

Figure 3:
RMS and Energy changes along transitions generated by DIMS and targeted molecular dynamics for the folding transition of Protein A, a three helix bundle.

Figure 4:

Histogram analysis of the population of path sampling in the two-dimensional projected coordinate space of the folding transition of protein A (three helix bundle). The two axes are number of hydrogen bonds in the structure (representing the formation of regular secondary structure) and the percent of native contacts.

Figure 5:

Two different strategies to increase diversity, in both cases the same final structure is used for protein G. a) Generating sets of ten trajectories using six different unfolded states as the initial structure. b) Generating 60 trajectories going from a single unfolded state.

Figure 6: Pairwise comparison between DIMS trajectories. Figure a) shows, via a color scale, the RMS differences between all conformations along two different DIMS trajectories, in white the line that represents the shortest path in RMSD space. Note that the two sets of conformations have a large range of RMS differences relative to each other. Figure b) shows the RMS values for the shortest path for several trajectories, it can be seen that the structures between two any given trajectories have a high RMS distance that ranges from 3 Å up to 5 Å for the intermediate states.

Figure 7:
Relative distribution of OM scores is shown. The set reflects on the diversity of sampled transitions.

Figure 8:

Normal mode self avoidance selection for Calcium Sensor S100A6 over 1000 trajectories. In red is shown the progress variable. Several modes are selected along the transition allowing the algorithm to sample contributions from different modes.

Figure 9:

a) The angle probability distribution in time, for the hinge opening movement of the glucose-galactose binding protein. b) Force distribution along the hinge angle.

Figure 10:

a) Angle probability distribution in time, for the hinge bending of maltodextrin. b) Force distribution along the hinge angle.

Figure 11:

Total root mean square distance traveled along the trajectory per residue for three different maltodextrin trajectories (a), (b) and (c)). In black the distance for the initial part, while in red the final regions of the trajectory. Residues in the catalytic site are labeled with blue bullets, residues in the N-terminal domain with green triangles. d) The same color scheme is used to illustrate the location of the residues in the structure. The C-terminal domain is highlighted in orange.

Figure 12: Probability of acceptance (equation (9)) as a function of RMS distance to the target, along the transitions for maltodextrin.

(a)

(b)

Figure 13:

- a) The angle probability distribution in time, for the hinge opening movement of lactoferrin.
b) The force distribution along the hinge angle.

Figure 14:

Comparison between MolMovDB, TMD and DIMS. a) Angle displacement as a function of the time step. b) Pulling velocity.

Table I:

DIMS parameters for SoftRatcheting

System	Size (N_{res})	N_{traj}	$\rho_0(\text{\AA})$	Cutoff	DIMS atoms	Align, atoms
Protein-A	60	100	1×10^{-5}	1.5 $\text{\AA}$	Backbone	All C_{α}
Protein-G	61	120	3×10^{-5}	1.0 $\text{\AA}$	Backbone	All C_{α}
GGBP	305	150	1×10^{-5}	0.9 $\text{\AA}$	Backbone	Res 111–252, 293–305
Maltodextrin	370	50	5×10^{-5}	0.8 $\text{\AA}$	Backbone	Res 1 – 111, 259 – 316
Lactoferrin	691	32	6×10^{-5}	0.8 $\text{\AA}$	Backbone	Res 342 – 680

Table II:

DIMS parameters for Normal Modes biasing. The same DIMS and alignment atoms as for SR were used.

System	N_{traj}	β	N_{bias}	N_{mod}	N_{best}	N_{comb}	N_{avoid}	Cutoff
Protein-A	200	0.025	30	276	15	3	250	1.5 Å
S100A6	200	0.05	30	534	15	3	250	1.4 Å
Lactoferrin	20	0.05	40	534	15	3	250	1.2 Å