

Discovering Free Energy Basins for Macromolecular Systems via Guided Multiscale Simulation

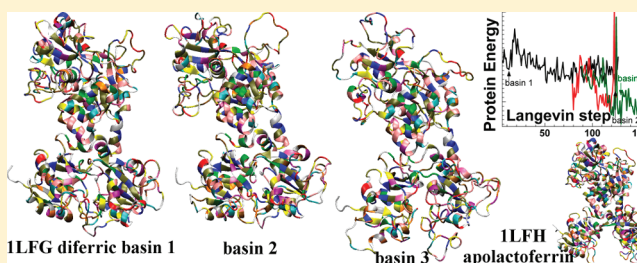
Yuriy V. Sereda, Abhishek B. Singharoy, Martin F. Jarrold, and Peter J. Ortoleva*

Center for Cell and Virus Theory, Department of Chemistry, Indiana University, 800 East Kirkwood Avenue, Bloomington, Indiana 47405, United States

S Supporting Information

ABSTRACT: An approach for the automated discovery of low free energy states of macromolecular systems is presented. The method does not involve delineating the entire free energy landscape but proceeds in a sequential free energy minimizing state discovery; i.e., it first discovers one low free energy state and then automatically seeks a distinct neighboring one. These states and the associated ensembles of atomistic configurations are characterized by coarse-grained variables capturing the large-scale structure of the system. A key facet of our approach is the identification of such coarse-grained variables. Evolution

of these variables is governed by Langevin dynamics driven by thermal-average forces and mediated by diffusivities, both of which are constructed by an ensemble of short molecular dynamics runs. In the present approach, the thermal-average forces are modified to account for the entropy changes following from our knowledge of the free energy basins already discovered. Such forces guide the system away from the known free energy minima, over free energy barriers, and to a new one. The theory is demonstrated for lactoferrin, known to have multiple energy-minimizing structures. The approach is validated using experimental structures and traditional molecular dynamics. The method can be generalized to enable the interpretation of nanocharacterization data (e.g., ion mobility–mass spectrometry, atomic force microscopy, chemical labeling, and nanopore measurements).



I. INTRODUCTION

Multiple macromolecular conformational states are observed in numerous experimental studies. However, it is often difficult to directly extract the 3D structures of a biomolecule from such data. Coarse-grained information on macromolecular assemblies is obtained in ion mobility–mass spectrometry experiments,¹ atomic force microscopy (AFM),² chemical labeling,³ and nanopore measurements.⁴ Such experimental data leave much ambiguity regarding the detailed structure. These data provide only a few parameters, while many configurational variables are required to capture the secondary structure of a large macromolecular system. Here, an information theory based method for the sequential discovery of conformational states of a macromolecular system as free energy minimizing structures is presented.

The dynamics of macromolecular systems involves the coupling of processes across multiple time and space scales. This multiscale character of macromolecular systems presents a challenge for identifying their numerous structural states. Here, this is addressed using a deductive multiscale approach⁵ as the basis of a structure discovery method.

The structural states usually of interest for macromolecular systems are those which minimize the free energy (FE). Each such state represents an ensemble of all-atom structures. The set of such structures, to which the nearby structures evolve spontaneously, is denoted a FE basin. However, due to thermal fluctuation, there are no all-atom structures that initiate

trajectories which subsequently always reside in the basin, although high-energy barriers may trap these trajectories within a basin for an exceedingly long time. Here, a framework for characterizing a FE basin and associated ensemble of all-atom configurations is presented. The ensemble for a given basin is required to compute associated average quantities that mediate the evolution of its all-atom configurations.

As one is typically interested in analyzing the kinetics of the transitions between FE minimizing structures, i.e., FE basins, a dynamical theory of the quantities characterizing the ensembles of all-atom configurations is needed. Here, the set of variables used to characterize such dynamical ensembles is denoted order parameters (OPs) Φ . Implicit in the above discussion is that there is a time scale separation between the dynamics of Φ and the individual all-atom states. Thus, while the system rapidly visits many all-atom configurations within the basin, the character of the ensemble, as tracked by Φ , is slowly varying. Therefore, a dynamical theory of OPs allows tracking the ensemble as the system evolves from one basin to another.

Interest in the biomolecular structure–function relationship has led to the development of theoretical structure discovery

Special Issue: Macromolecular Systems Understood through Multiscale and Enhanced Sampling Techniques

Received: December 30, 2011

Revised: March 14, 2012

methods inspired by quickly growing sequencing data. However, progress in obtaining experimental structures has been much slower. Theoretical and computational methods for discovering the structure of macromolecular systems have recently been reviewed⁶ and include the following: combinatorial methods^{7,8} for finding global minimum energy conformations; all-atom structure reconstruction methods⁹ which start with experimental $C\alpha$ trace and rely on combinatorial side-chain optimization and standard molecular dynamics (MD) energy minimization; approaches which employ specific energy functions,¹⁰ simulated annealing,¹¹ and mean-field optimization;¹² global optimization approaches for the structure prediction and FE calculations of solvated peptides;¹³ Monte Carlo method¹⁴ in conjunction with simulated annealing;^{15,16} genetic optimization algorithms;^{17,18} rigid-cluster elastic network interpolation technique¹⁹ and normal-mode analysis²⁰ for generating feasible transition pathways between known macromolecular conformations; and enhanced rare-event sampling techniques such as transition path sampling²¹ and metadynamics^{22,23} (including molecular dynamics flexible fitting²⁴ for resolving low-resolution cryo-EM structures).

These approaches provide insights into global protein structure optimization, transition paths, and FE landscapes. However, there is still room for improvement of theoretical structure determination strategies in light of one or more of the following. (1) They are usually limited to small systems and biologically short times. (2) Most implementations use simplified potential models. (3) Coarse-grained models require extensive recalibration for each new application, and governing dynamics equations must be postulated. Calibration is often limited to a small set of equilibrium states, even when used to describe nonequilibrium processes. Extensive data are used for calibration including system-specific information such as the area and volume accessible to the solvent, the sparse sets of NMR data, or subsystem structural information integrated via bioinformatics methods. (4) No criterion is provided for determining the completeness of the set of coarse-grained variables used to characterize the FE landscape. (5) All-atom interactions and states are not used or obtained. (6) Only unsolvated structures are addressed. (7) Only a single FE minimizing structure is provided. (8) Potential energy minimizing structures, and not FE minimizing ones, are provided. (9) Long computational times are needed to simulate transitions between energy basins. (10) Guide a system through a path that is not necessarily natural for the molecular physics. (11) The history of configurations generated in a Monte Carlo sequence is not always accounted for when mapping the energy landscape. (12) No guidelines are provided for optimization, although performance may be very sensitive to the specific implementation, e.g., as in the genetic optimization algorithm. (13) The simulation may be trapped in local, but not global, minima. (14) Knowledge of initial and/or final states may be required.

The objective of the present approach is to overcome most of these difficulties using the following: an all-atom underlying formulation and continuous (and not discrete) configuration space; coarse-grained structural variables; a multiscale methodology to derive Langevin equations for the OPs and algorithms for computing all factors in these equations from an interatomic force field; a FE basin discovery method using modification of FE driving thermal-average forces for OP evolution that integrates prior knowledge of known FE minimizing structures to guide the evolution to yet-unknown ones; an efficient, calibration-free multiscale simulation methodology on which to

build the methodical search algorithm and which is flexible enough to incorporate experimental data of a range of resolutions; the FE basin sequential elimination technique introduced here does not require prior knowledge of the reaction path, nor the final or initial structure.

A FE basin is defined as an ensemble of all-atom configurations consistent with a set of OPs that minimize the FE, i.e., for which thermal-average forces vanish. As the multiscale simulations progress via Langevin timesteps, it is often necessary to modify the definition of the OPs,²⁵ so that other variables (denoted descriptors here) are also used to characterize a FE basin. These descriptors are defined directly in terms of all-atom configurations and are typically directly measurable characteristics (e.g., moments of inertia or electrical dipole and quadrupole moments). Thermal-average OP forces modified by using the descriptors characterizing known structures are introduced to guide a multiscale simulation away from the known basins to a new one.

The multiscale formalism for simulating macromolecular assemblies is reviewed (Section II). The thermal-average forces arising in multiscale analysis are modified such that they drive macromolecular systems to new FE basins, enabling a sequential basin discovery algorithm; details on implementation are provided (Section III). Validation is presented (Section IV), and conclusions are drawn (Section V).

II. BRIEF REVIEW OF THE DEDUCTIVE MULTISCALE APPROACH

A natural framework for casting an FE basin discovery algorithm is deductive multiscale analysis.^{5,26,27} Let $\underline{r} \equiv \{\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N\}$ denote the all-atom configuration of the structure of interest. In the approach adopted here, OPs $\Phi \equiv \{\vec{\Phi}_1, \vec{\Phi}_2, \dots, \vec{\Phi}_{N_{OP}}\}$ describe the overall structure of the system. As earlier,^{26,28,29} the starting point of the analysis is the Φ - \underline{r} relationship²⁹

$$\vec{r}_i = \sum_{k=1}^{N_{OP}} U_{ki} \vec{\Phi}_k + \vec{\sigma}_i \quad (1)$$

The residuals $\vec{\sigma}_i$ are introduced to address the truncation of the k -sum in eq 1 resulting from taking a relatively small number of OPs, $N_{OP} \ll N$. With this, the k -sum generates the continuous deformation of the N -atom assembly via changes in Φ , while the $\vec{\sigma}_i$ account for more random individual atomic motions.³⁰ A reference structure \underline{r}° is used to construct the basis functions U_{ki} as stated earlier.²⁵ Using mass-weighted orthogonalized³¹ U_{ki} ,²⁶ one obtains

$$\vec{\Phi}_k = \frac{1}{\mu_k} \sum_{i=1}^N m_i U_{ki} \vec{r}_i, \quad \mu_k = \sum_{i=1}^N m_i U_{ki}^2 \quad (2)$$

m_i being the mass of atom i . This formulation has the only difference from the one in ref 26 wherein the μ_k was directly embedded in U_{ki} . This more explicit formulation suggests that the $\vec{\Phi}_k$ are generalized CM variables and the μ_k are associated masses. For example, if U_{ki} is independent of i , then the related OP is the center of mass. Other $\vec{\Phi}_k$ characterize finer details of the distribution of mass.³² Equation 2 does not provide the reciprocal relation, i.e., does not imply \underline{r} for given Φ , since the $\vec{\sigma}_i$ are yet-unspecified. This is expected since a given coarse-grained description (Φ here) corresponds to an ensemble of all-atom configurations, as addressed in more detail below.

The deductive multiscale approach²⁵ starts with the N -atom probability density ρ that depends on the $6N$ atomic coordinates and momenta (denoted Γ). While the N atoms constitute the structure of interest, atoms in the remainder of the system are labeled with $i > N$. The starting point of the multiscale analysis is the ansatz that ρ depends on Γ both directly and, via the OPs, indirectly. When the U_{ki} change slowly as the reference structure \underline{r}° varies, use of the OPs as defined in eq 2 introduces a smallness parameter ε in the Liouville equation obtained through the ansatz

$$\rho = \rho(\Gamma, \underline{\Phi}; t_0, \underline{t}; \varepsilon), \quad \underline{t} \equiv \{t_1, t_2, \dots\}, \quad t_n = \varepsilon^n t \quad (3)$$

Since ε is related to the ratio of the mass of a typical atom to that of a subset of the atoms, it is small and thereby enables a perturbation analysis.²⁶ The result is a Langevin equation for $\underline{\Phi}$ and the coevolving OP-constrained, quasi-equilibrium probability density $\hat{\rho}$ of all-atom structures^{25,26,30}

$$\begin{aligned} \hat{\rho} &= \exp(-\beta H)/Q, \\ Q(\underline{\Phi}) &= \int d\Gamma^* \Delta^+(\underline{\Phi} - \underline{\Phi}^*) \exp(-\beta H^*), \\ \beta &= 1/k_B T \end{aligned} \quad (4)$$

where $*$ denotes evaluation at Γ^* over which integration is taken, and H is the Hamiltonian. The Δ^+ factor is the product of N_{OP} narrow Gaussian-like functions introduced to impose an OP constraint on the ensemble of all-atom configurations.

With the above, the ensemble of all-atom configurations characterized by $\hat{\rho}$ evolves with $\underline{\Phi}$. In turn, $\underline{\Phi}$ evolves via the following Langevin dynamics

$$\frac{\partial \Phi_{k\alpha}}{\partial \tau} = \beta \sum_{k'\alpha'} D_{kak'\alpha'} \dot{f}_{k'\alpha'} + \xi_{k\alpha}, \quad \tau = \varepsilon^2 t \quad (5)$$

$\Phi_{k\alpha}$ is the α th Cartesian component of $\vec{\Phi}_k$, and \vec{f}_k is the thermal-average force given by the phase space average of the corresponding OP force^{25,26}

$$\vec{f}_k(\underline{\Phi}) = \int d\Gamma^* \Delta^+(\underline{\Phi} - \underline{\Phi}^*) \exp(-\beta H^*) \sum_{i=1}^N U_{ki} \vec{F}_i \quad (6)$$

where \vec{F}_i is the net force on atom i . The diffusivity factors $D_{kak'\alpha'}$ in eq 5 are related to correlation functions of OP time derivatives.^{5,26} A random noise term ξ_k determines the stochastic part of Langevin evolution and is constructed by requiring the integral of its autocorrelation function to be proportional to the diffusion coefficient D_{kaka} .²⁶

The above multiscale methodology was implemented as the DeductiveMultiscaleSimulator (DMS) software,^{5,26} originally as the MD/OPX software,^{29,33} and recently redesigned, optimized, and seamlessly integrated with NAMD³⁴ via a new Python interface. In the present implementation, the thermal-average forces \vec{f}_k were calculated using Monte Carlo integration. The ensemble of all-atom configurations \underline{r} needed was generated in two steps. First, eq 1 was used with statistically chosen $\underline{\sigma}$ to generate a preliminary ensemble of all-atom configurations consistent with instantaneous values of the OPs as they evolve according to the Langevin dynamics (eq 5). This initial ensemble was enriched via short isothermal MD runs over which $\underline{\Phi}$ does not change appreciably. The method

takes advantage of the special properties of the OPs introduced as in eq 1,^{25,26,30} originally cast as a space-warping framework.²⁸

III. FORMULATION AND IMPLEMENTATION OF THE SEQUENTIAL BASIN DISCOVERY METHOD

III.A. Discovery Concepts. The foundation of the sequential elimination method for FE basin discovery is as follows. Free energy is thermal energy minus temperature times entropy. Entropy depends on the available information on the constraints to which the system is subjected (e.g., temperature or specific values of $\underline{\Phi}$). In the sequential elimination approach, this information includes the fact that some FE basins are known and one seeks to discover new ones.

To implement this basin discovery method, a set of N_d descriptors $\underline{\eta} \equiv \{\eta_1, \dots, \eta_{N_d}\}$ is used to characterize a basin. While these descriptors characterize overall system structure as do the OPs, they are not used directly in the multiscale formulation since they may not serve as the basis of the \underline{r} - $\underline{\Phi}$ relation 1.

In what follows, we develop formulas for these descriptors and show how they can be used to automatically guide the multiscale dynamics (Section II) to a new basin given the descriptors for the known ones. For simplicity, we present the method for the case when one basin is known and a second one is sought. Generalization for multiple known basins is straightforward (see below).

The search algorithm we provide combines elements of (1) the multiscale analysis of macromolecular systems;^{25,26,30,35–48} (2) the notion of a stepwise procedure that precludes evolution into basins of attraction identified in earlier steps in the computation; (3) an OP method for simplifying the FE landscape to eliminate thermally irrelevant basins of attraction. In addition, (4) an algorithm for accounting for experimentally determined structural information can be incorporated in the search algorithm.³⁰

III.B. Implementation of the Basin Discovery Algorithm. The methodology of Section III.A for FE basin discovery was implemented by modifying the DMS software;^{5,25,26} this implementation is denoted here DeductiveMultiscaleSimulator-BasinDiscovery (DMS.BD). DMS uses NAMD with the CHARMM force field to perform selected calculations to construct forces and diffusions in the Langevin equations (Section II). Details on these MD calculations are provided in the Supporting Information. DMS was modified by changing the expressions for thermal-average forces (Section III.D) and similarly for the diffusion factors via averaging over restricted phase space. The workflow of DMS.BD is shown in Figure 1.

The entire FE landscape is not calculated since the discovery of adjacent basins does not require it.⁴⁹ Instead, the natural thermal-average forces \vec{f}_k (eq 6) are used to locate basins in the space of OPs or descriptors. The bottom of a basin is defined to be the point in OP space where all natural thermal-average forces vanish. If a simulation winds up in a local minimum (which, coincidentally, may be physically interesting), then in the next stage of sequential elimination the system will evolve to another basin, and so on. Like in any other method, the local minima are distinguished from a global minimum according to the depth and breadth of the basins of attraction discovered. At the end of the given step of sequential elimination, the system will be at the bottom of a well and will have departed from the saddle point.

DMS enables evolution of OPs along with an ensemble of all-atom configurations constrained by the instantaneous OP

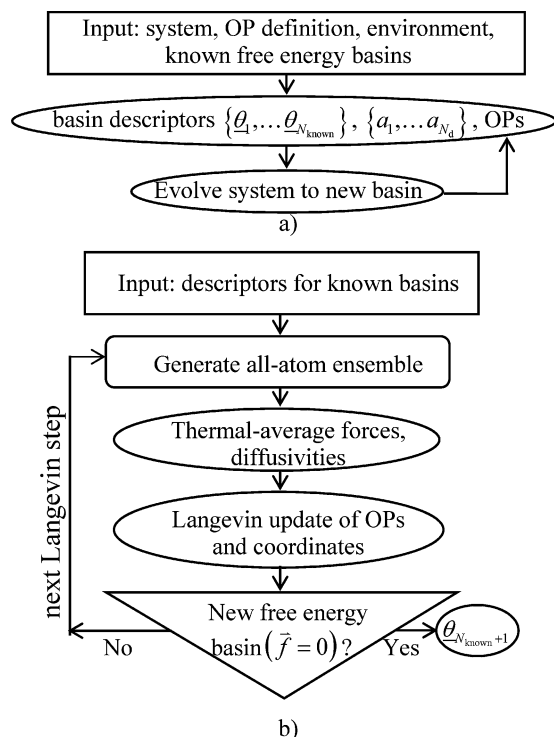


Figure 1. Workflow of the DMS.BD algorithm that enables traversal of FE barriers and discovery of new basins. (a) Input includes an initial all-atom structure in solvent, definition of basis functions and OPs, size of the Langevin time step, update frequency for the reference structure, and conditions in the host medium. Discovery of a new FE basin starts with establishing descriptors and values of OPs at the bottom of known basins and choosing width parameters for the Δ^- factors. (b) Flowchart for evolution to new basin via guided Langevin dynamics. For a detailed explanation of each step, see the first paragraph of the third subsection in the Supporting Information.

values. Such ensembles are constructed using a set of MD runs that capture a time scale much shorter than those of the OPs and are initialized to a given value of the OPs using higher-order OP-like variables as earlier²⁶ and in Section II. Thus, OPs remain essentially constant when the all-atom configurations are sampled using such MD runs. As a result, the state-counting factor Δ^+ appearing in the partition function (eq 11) is accounted for in thermal force and diffusivity calculations. At each Langevin time step, OPs constrain the quasi-equilibrium ensemble of atomic states which, in turn, enables the computation of the thermal-average forces (eq 19) and diffusivities^{5,26} that mediate Langevin OP dynamics (eq 5). With this, the modified thermal-average forces guide the system to new FE basins as in Section III.D.

In DMS.BD, the coevolving quasi-equilibrium ensemble is modified using the method of Section III.D. Information on known basins is accounted for in the state-counting factor Δ^- in the form of the product, with one factor (eq 10) for each basin. In the current implementation of DMS.BD, each of these factors involves several descriptors, as follows. The set $\{a_1, \dots, a_{N_d}\}$ of accompanying exponential width factors (Section III.D) was taken to be identical for all basins.

III.C. Descriptors. Examples of descriptors that can be used for basin discovery include total mass, charge, length of the dipole moment, and eigenvalues of the moment of inertia or electrical quadrupole moment tensors. Such descriptors have

the important property that they are independent of system orientation. In the current implementation, the three eigenvalues of the moment of inertia tensor of the structure relative to its center of mass are chosen. To discriminate between more complex structures and associated FE basins, more descriptors can be used.

The moment of inertia tensor $\vec{\vec{M}}$ is defined via

$$\vec{\vec{M}} = \sum_{i=1}^N m_i \begin{pmatrix} y_i^2 + z_i^2 & -x_i y_i & -x_i z_i \\ -x_i y_i & x_i^2 + z_i^2 & -y_i z_i \\ -x_i z_i & -y_i z_i & x_i^2 + y_i^2 \end{pmatrix} \quad (7)$$

Being the eigenvalues of $\vec{\vec{M}}$, molecular descriptors satisfy the cubic equation

$$\begin{aligned} \eta^3 - (M_{xx} + M_{yy} + M_{zz})\eta^2 + (M_{xx}(M_{yy} + M_{zz}) \\ + M_{yy}M_{zz} - M_{xy}^2 - M_{yz}^2 - M_{xz}^2)\eta \\ + M_{yy}(M_{xz}^2 - M_{xx}M_{zz}) + M_{zz}M_{xy}^2 \\ + M_{yz}(M_{xx}M_{yz} - 2M_{xy}M_{xz}) = 0 \end{aligned} \quad (8)$$

whose coefficients are determined by the elements of matrix 7.

To proceed in a sequential elimination calculation, all-atom configurations which yield descriptors close to those for the known basin are eliminated from the ensemble as follows. First one must specify the descriptors that characterize the known basin. However, a basin includes an ensemble of all-atom configurations. For isothermal systems, this ensemble is generated as earlier^{25,26} using short isothermal MD runs initialized with configurations consistent with instantaneous values of Φ (Section II). Out of this ensemble, a most probable structure with the lowest potential energy is chosen to calculate descriptor values characterizing the known basin.

III.D. Modification of Thermal-Average Forces to Include Known Basin Information. The starting point for the sequential basin discovery is entropy maximization to determine the quasi-equilibrium probability density $\hat{\rho}$ constrained by the known information. These constraints include the isothermal condition and fixed system volume, as well as the instantaneous values of the OPs at a given stage of the Langevin dynamics. In addition, states that resemble those in the known basin are excluded from the counting of states in the entropy for a sequential elimination computation. With this, the entropy \hat{S} takes the form

$$\hat{S} = -k_B \int d\Gamma^* \Delta^+(\Phi - \Phi^*) \Delta^-(\eta^* - \eta_{\text{known}}^*) \hat{\rho}^* \ln \hat{\rho}^* \quad (9)$$

where an additional factor Δ^- is introduced to discount the known FE basin via the descriptors at its bottom, $\eta_{\text{known}} = \{\theta_1, \dots, \theta_{N_d}\}$. Since the descriptors are coarse-grained variables, they can be expressed in terms of a complete set of OPs (see second subsection of Supporting Information). Therefore, they are used here to introduce information on the known basin to enable the discovery.

The factor Δ^- has the character of $1 - \Delta^+$ and therefore excludes configurations in the known basin, i.e., configurations which have descriptors close to η_{known} . In other words, to give

preference to the states that are different from those in the known FE basin, this counting factor is set to one for configurations distinct from the known one and is zero within the known basin. The particular form of Δ^- was chosen as

$$\Delta^-(\underline{\eta} - \underline{\theta}_{\text{known}}) = 1 - \exp\left(-\sum_{d=1}^{N_d} a_d |\eta_d(\underline{r}) - \theta_{d,\text{known}}|\right) \quad (10)$$

The parameter a_d is proportional to the inverse width of the Gaussian-like exponential function associated with descriptor d in eq 10. The a_d values are chosen to ensure escape from the known basin (see below).

Using eq 9 and entropy maximization, one arrives at the OP-constrained all-atom probability density $\hat{\rho}$ (eq 4). The associated partition function Q takes the form

$$Q(\underline{\Phi}; \underline{\theta}_{\text{known}}) = \int d\Gamma \Delta^+ (\underline{\Phi} - \underline{\Phi}^*) \Delta^- (\underline{\eta}^* - \underline{\theta}_{\text{known}}) \times \exp(-\beta H^*) \quad (11)$$

This yields the Helmholtz free energy F

$$F = -\frac{1}{\beta} \ln Q(\underline{\Phi}; \underline{\theta}_{\text{known}}) \quad (12)$$

By analogy with the developments of Section II, the modified thermal-average forces \vec{f}_k^m are obtained (see below). Then, the Langevin eq 5 can be used to evolve the system from the known basin to a new one. This calculation is carried out in the present implementation using methods as described earlier^{25,26,29} but with the present modified OP forces described in detail below. At the end of each Langevin time step, the updated OPs are obtained along with the associated ensemble of all-atom configurations. In turn, the latter are used to generate the next Langevin time step. This Langevin time-stepping is stopped when the natural (not modified; see Section II) thermal-average forces are negligible, indicating arrival at the bottom of the new FE basin.

Here we derive the expression for the thermal-average forces, modified by the state-counting factors Δ^- (eq 10) accounting for the earlier discovered FE basins. Provided the set of known FE basins with associated low-energy states, characterized by $N_d N_{\text{known}}$ descriptor values $\theta_{d,b}$, the Δ^- is calculated for each of the atomic configurations \underline{r} generated by MD sampling at a given Langevin time step. The associated contribution to the OP forces is composed of the derivatives of Δ^- with respect to the OPs $\vec{\Phi}_k$. The latter can be computed using the derivatives with respect to atomic coordinates, the chain rule, and the $\underline{\Phi}-\underline{r}$ relationship 1. When deriving new thermal-average forces \vec{f}_k^m , one brings the $\partial/\partial\vec{\Phi}_k$ derivative into the integral (eq 11) and uses the property of the Hamiltonian H that it does not depend on OPs explicitly

$$\beta Q \vec{f}_k^m = \int d\Gamma \exp(-\beta H) \sum_{i=1}^N \frac{\partial}{\partial \vec{r}_i} \{\Delta^+ \Delta^-\} \frac{\partial \vec{r}_i}{\partial \vec{\Phi}_k} \quad (13)$$

Assume that \underline{r} can be obtained from an augmented set of OPs (i.e., those including the residual parameters as in eq 1). Then the following approximation holds²⁹

$$\frac{\partial \vec{r}_i}{\partial \vec{\Phi}_k} = \sum_{\alpha=\{x,y,z\}} \frac{\partial r_{i\alpha}}{\partial \Phi_{k\alpha}} = \sum_{\alpha} U_{k\alpha^i} (r_{i\alpha}^0) \quad (14)$$

Using eq 14, one obtains

$$\beta Q \vec{f}_{k\alpha}^m = \int d\Gamma \exp(-\beta H) \sum_{i=1}^N U_{k\alpha^i} \frac{\partial}{\partial r_{i\alpha}} \{\Delta^+ \Delta^-\} \quad (15)$$

With this, the thermal-average force is that obtained earlier (eq 6), with the extra Δ^- weighting factor (eq 10), plus a new term \vec{f}^b arising from the following integral

$$\begin{aligned} \vec{f}^b &= \sum_{k=1}^{N_{\text{OP}}} \vec{f}_k^b, f_{k\alpha}^b \\ &= \frac{1}{\beta Q} \int d\Gamma \Delta^+ \exp(-\beta H) \sum_{i=1}^N U_{k\alpha^i} \sum_{d=1}^{N_d} \frac{\partial \Delta^-}{\partial \eta_d} \frac{\partial \eta_d}{\partial r_{i\alpha}} \end{aligned} \quad (16)$$

In view of eq 16, the derivatives of Δ^- with respect to the descriptors $\underline{\eta}$ should achieve their maximum values in the already discovered stable states to maximize biasing force \vec{f}^b .

The derivatives of the descriptors with respect to the atomic coordinates $\partial \eta_d / \partial r_{i\alpha}$ in the new thermal-average force term \vec{f}_k^b in eq 16 are to be taken numerically for each of the OP-restricted configurations within same Langevin time step. These derivatives were calculated using $3N$ independent offsets in the x , y , and z coordinates of each atom with subsequent recalculations of η_d . The speed-up in calculating these derivatives was achieved by using the analytical roots of polynomial 8, as opposed to using the eigenvalue calculation subroutines.

The thermal-average forces \vec{f} in our earlier approach (Appendix C in ref 26) are obtained from the $\partial \Delta^+ / \partial \vec{r}_i$ term in eq 15

$$\begin{aligned} \exp(-\beta H) \frac{\partial \Delta^+}{\partial r_{i\alpha}} &= \frac{\partial}{\partial r_{i\alpha}} \{\Delta^+ \exp(-\beta H)\} \\ &\quad - \Delta^+ \frac{\partial \exp(-\beta H)}{\partial r_{i\alpha}} \end{aligned} \quad (17)$$

Using the property of Δ^+ that it does not depend on spatial coordinates explicitly, but rather via OPs, and employing the divergence theorem, we present the first term in eq 17 in a form of full gradient and note that its contribution to the integral 15 is zero. The space derivative of H in the second term of eq 17 is a negated α -component of the corresponding atomic force, \vec{F}_i . Thus, one obtains

$$\begin{aligned} \vec{f}^- &= \sum_{k=1}^{N_{\text{OP}}} \vec{f}_k^-, \\ f_{k\alpha}^- &= -\frac{1}{Q} \int d\Gamma \exp(-\beta H) \Delta^+ \Delta^- \sum_{i=1}^N U_{k\alpha^i} F_{i\alpha} \end{aligned} \quad (18)$$

Here, the \vec{f}_k^- are thermal-average forces modified by the Δ^- factor in the phase space integral 18.

It was verified that by neglecting the biasing force (eq 16) and using only the term \vec{f}^- (eq 18), i.e., by simply multiplying the integrand in the expression for thermal-average forces (eq 6) by the anti-Gaussian-like probability function Δ^- of descriptors, one does not provide the desired driving force for the system to evolve out of the discovered FE basins. This is because the noise term dominates over the OP forces (eq 5). Attempts to increase the Langevin time step Δt and narrowing the widths a_d^{-1} of the anti-Gaussian Δ^- did not lead to the increase in OP forces.

The overall thermal-average force consists of two components: the FE driving forces $f_{k\alpha}$ modified by Δ^- factor and the biasing information theory-guiding ones ($f_{k\alpha}^b$),

$$\begin{aligned} f_{k\alpha}^m &= f_{k\alpha}^- + f_{k\alpha}^b \\ &= \frac{1}{Q} \sum_{i=1}^N U_{k\alpha^i} \int d\Gamma \exp(-\beta H) \Delta^+ \left(-\Delta^- F_{i\alpha} \right. \\ &\quad \left. + \frac{1}{\beta} \sum_{d=1}^{N_d} \frac{\partial \Delta^-}{\partial \eta_d} \frac{\partial \eta_d}{\partial r_{i\alpha}} \right) \end{aligned} \quad (19)$$

Discussed in Section IV, the mutually opposing nature of these forces underlies the discovery of basins and associated free-energy minimizing structures via DMS.BD.

III.E. Guided Evolution from the Known to Unknown Basins. The calculation from a known basin to a new one proceeds as follows. One starts the calculation within the known basin and then evolves the system via Langevin eq 5 with modified thermal-average forces of eq 19. A Langevin evolution course is tracked by the values of potential energy and \vec{f}_k (eq 6). After a high FE barrier is overcome and the system descends toward the bottom of a new FE basin, the basin discovery simulation is carried on until the thermal-average forces become negligible. A new basin structure is chosen from the time step at which \vec{f}_k are negligible, signifying that minimum free energy was achieved within the new basin.

Let the biasing thermal-average force \vec{f}_k^b (eq 16) be the $\vec{\Phi}_k$ gradient of the FE associated with the modified partition function (eq 11). Specifically, the \vec{f}_k^b are computed using the derivatives of the state-counting factor 10 with respect to the descriptors

$$\frac{\partial \Delta^-}{\partial \eta_d} = \Delta^- \sum_{b=1}^{N_{\text{known}}} \frac{\text{sgn}(\eta_d(r) - \theta_{d,b}) a_d}{\exp(\sum_{d=1}^{N_d} a_d |\eta_d(r) - \theta_{d,b}|) - 1} \quad (20)$$

It is also necessary to make an estimate of the N_d inverse width parameters a_d . In the present implementation, this is accomplished via adjustment to ensure escape from the known basin (see third subsection of Supporting Information).

The present approach not only allows the system to escape previously discovered minima but also ensures that it reaches the next state at which the modified free energy is minimal. When the system departs far enough from all known free energy basins such that the molecular descriptor-dependent Δ^- factor grows to almost one, it becomes insensitive to the escape

factors in the Monte Carlo integration formula (see eqs 19 and 20). As a result, the biasing force (eq 16) approaches zero as the barrier separating the new minimum from the old one is surmounted. The remaining part of the thermal-average force then is the natural force modified by the (almost constant) Δ^- factor (see eq 18). Thus, after escaping the known basins of attraction, the system is essentially only driven by the natural free energy force which, by construction, drives the system to a minimum in the true free energy landscape.

At the stage when the system is driven downhill by only the natural free energy force, the simulation algorithm becomes the same as implemented in our DeductiveMultiscaleSimulator. The latter proceeds one or more orders of magnitude faster than traditional MD and preserves accuracy. Therefore, the search for a new energy minimum is not random but is directed both out of known basins and to new ones. Unlike other approaches, our multiscale methodology provides the way to find multiple free energy minima in a sequential manner, as explained below. In contrast, a random search based on randomly chosen initial data could lead to many evolution scenarios ending in the same basin and, therefore, wasting simulation time.

III.F. Generalization for Sequential Discovery of Multiple Free Energy Basins. The case of a single known basin and the discovery of a new one was considered above. This algorithm can readily be generalized to the case of sequential discovery in a stepwise procedure. At each step, the system is guided away from the basins discovered in earlier steps to a new one. In a given step, the Δ^- factor (eq 10) can be generalized to be a product of similar factors $\Delta^-(\eta - \theta_b)$, one for each of the known basins labeled $b = 1, \dots, N_{\text{known}}$. For basin b , the N_d descriptors $\theta_{d,b}$ are accounted for, and the set $\{a_1, \dots, a_{N_d}\}$ of factors in the exponential function (eq 10) is chosen to ensure escape from each of the known basins.

The Δ^- factors artificially lower the FE of a system as it evolves out of the known basins. When Δ^- are incorporated in a sequential calculation, the system is driven away from all basins discovered in earlier phases of the calculation by the modified thermal-average forces \vec{f}_k^m (eq 19). Within a given phase of such a calculation, a number of Langevin steps are to be carried out to arrive at the set of OP values at the bottom of a newly discovered basin.

It is possible that an artificial minimum is created due to the alterations of the free energy functional. Such artifacts can be detected and eliminated via a subsequent traditional MD simulation. Alternatively, one can restart this process at an artificial minimum, and the system will be driven away to a new minimum. However, this was not found to be a problem for the case of lactoferrin studied.

IV. VALIDATION FOR HUMAN LACTOFERRIN

The FE basin discovery method was validated by finding two new FE basins on the FE landscape for human lactoferrin.⁵⁰ A brief summary of observations on this system is as follows. Two crystal X-ray structures are available for this protein: diferric lactoferrin (PDB code 1LFG) and apolactoferrin (PDB code 1LFH).⁵⁰

To validate the general DMS.BD algorithm of Section III.F, an arbitrary structure, and notably the compact closed-lobe diferric conformation 1LFG¹⁹ with the iron and carbonate ions removed (Figure 2a), was used to start a DMS simulation to find the bottom of a first FE basin. Then DMS.BD was used to simulate traversal of the FE topography and discover a new

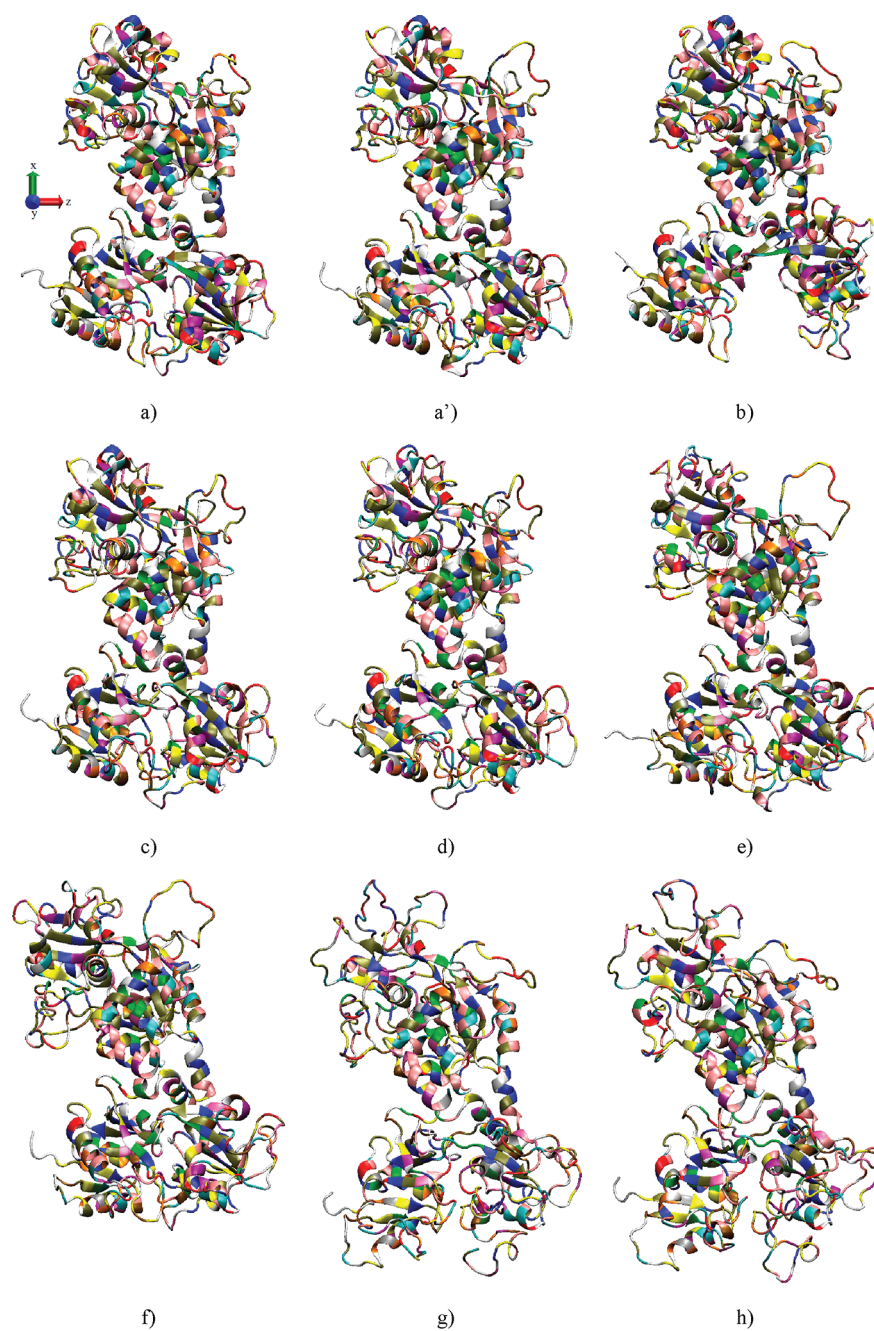


Figure 2. Following types of human lactoferrin structures are presented: crystallographic (a), (b); discovered basin bottoms (c), (e) and (f), (h); and transition points (d), (g). These are (a) closed 1LFG X-ray structure, (a') 1LFG MD used to start DMS simulation; (b) open 1LFH; (c) at the bottom of basin 1; (d) transition point along the basin 1→2 pathway; (e) arbitrarily chosen Langevin time step from basin 2 (called “descent 2” structure, see subsection five in Supporting Information); (f) bottom of basin 2; (g) basin 2→3 transition point (starting at the “descent 2” structure); (h) bottom of basin 3. All these structures are of lowest potential energy among those in the ensemble consistent with the instantaneous OP values. A transition point between basins is taken to be at the Langevin time step for which the potential energy goes through a maximum. Such points are close to transition regions where FE force changes sign (see paragraph 6 in Section IV).

basin for lactoferrin starting from the diferric basin. A set of descriptors characterizing the closed-lobe lactoferrin structures from the bottom of this first basin were used to guide simulation away from it. Lactoferrin was guided to a second basin with slightly opened structures. Next, a set of descriptors characterizing the second “pseudodiferric” basin were incorporated to guide protein to a third basin. The third basin contains open-lobe structures (Figure 2h) which are similar to the apolactoferrin conformation 1LFH but are less open than the

X-ray structure 1LFH (Figure 2b). Additional details on DMS and DMS.BD simulations are provided in the Supporting Information.

Implementation of DMS.BD is based on the interplay of Δ^- -modified (\vec{f}_k^-) and biasing (\vec{f}_k^b) components of the modified FE driving forces \vec{f}_k^m (eq 19). Inclusion of the Δ^- factor in state counting reduces those FE minimizing forces \vec{f}_k^- (eq 18) that would have otherwise kept the system within the known

basin (i.e., if \vec{f}_k and not \vec{f}_k^m was used). The biasing forces \vec{f}_k^b , by design, oppose the \vec{f}_k and, by the choice of inverse width parameters a_d in eq 10, drive the system away from known basin(s). Once out of a known basin, \vec{f}_k^b becomes smaller than \vec{f}_k^- if the structure changes appreciably relative to those characterizing the known basins. Thus, after a barrier is crossed, the \vec{f}_k drive the system toward the FE minimizing structure for which descriptors differ from those in the known basin(s). With this, the \vec{f}_k^m drive the system away from known basins and to new ones.

To rationalize the above effects, we compute the thermal-average forces \vec{f}_k (eq 6) at each Langevin step of a DMS.BD trajectory. For structures near the bottom of a basin, all \vec{f}_k are close to zero. However, stochastic forces $\vec{\xi}_k$ (eq 5) force the system to fluctuate about the bottom. If the system is far from the bottom, the \vec{f}_k are appreciable and drive the system to the bottom. With this, the \vec{f}_k provide information on the FE landscape topography along a Langevin evolution path. They indicate the location of FE barriers along the path (i.e., places where the \vec{f}_k vanish). For extensive sampling, integration over these forces yields an estimate of FE barrier height when the \vec{f}_k are integrated (see subsection four in Supporting Information).

For lactoferrin oriented as in Figure 2, lobes open in the xz -plane accompanying the transition from the diferric to the apolactoferrin basin. In particular, OPs Φ_{100X} and Φ_{001Z} track extension-compression along the x and z directions, capturing the structural transition. The closed and open states are also characterized by values of the descriptors (i.e., moment of inertia eigenvalues, Figure 3). Thermal-average forces along the DMS.BD trajectory from the diferric to the pseudodiferric basin are shown in Figure 4. Most forces fluctuate around zero along this trajectory, suggesting that local topography along the guided trajectory has the character of a valley. However, forces f_{100X} and f_{001Z} along the trajectory suggest that a barrier is crossed; i.e., they change from negative to positive as the barrier is traversed, stop growing, and ultimately go to zero as the system approaches the bottom of a new basin (Figure 4). This illustrates that our method explores local topography in the vicinity of high probability pathways; i.e., the FE is minimum along the directions orthogonal to the path.

In Figure 5, we plot potential energies of the most probable atomic configurations from constant OP ensembles at every Langevin step during transitions between specified basins. The potential energy profile also suggests a barrier crossing. The presence of such barriers suggests that the FE and potential energy landscapes are related, but not identical (Figure 4 versus Figure 5). This is expected because entropy effects at finite temperatures are not reflected in the potential energy profile and, therefore, can shift the location of potential energy features (minima or transition points) relative to the FE ones. The above transition path and topography are not readily accessible via traditional MD, as follows.

To confirm that different FE basins were discovered, an ensemble of all-atom configurations in the vicinity of the bottom of each basin was explored using traditional isothermal MD. For a given basin, the MD was initialized with an all-atom state of minimum potential energy, which was consistent with

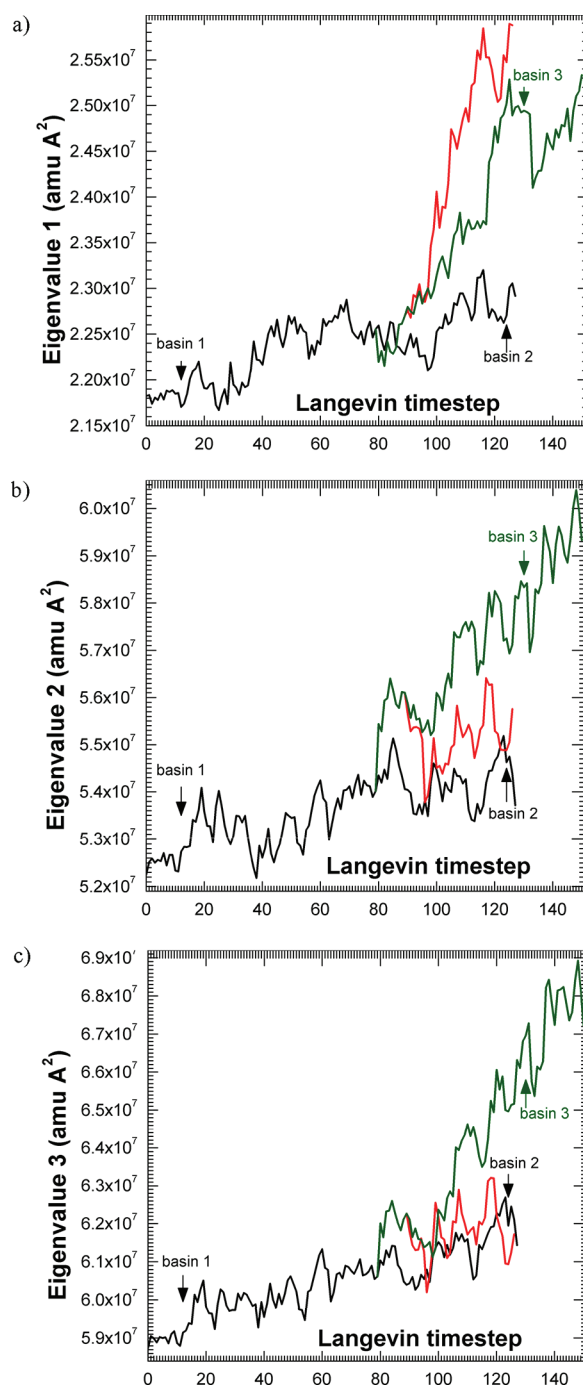


Figure 3. DMS.BD Langevin timecourses for descriptors showing distinct differences between the basins. (a) Eigenvalue 1, (b) eigenvalue 2, and (c) eigenvalue 3 of the moment of inertia tensor for human lactoferrin. (black) Basin 1 \rightarrow 2 transition; (red) control simulation launched from basin 2; (green) basin 2 \rightarrow 3 transition proving robustness of the basin discovery method (subsection five in Supporting Information). The eigenvalues remain fairly constant at the bottom of basins (Figure S2) and change during interbasin transitions. Eigenvalues 2 and 3 behave similarly during the basin 1 \rightarrow 2 transition; however, this similarity is lost in the next transition. If one starts in basin 1, only modest changes in descriptors are observed. In contrast, when precluding basins 1 and 2 in going to basin 3, descriptors change much more, implying greater extent of lobe opening.

the OPs at the bottom of the basin. Then, 10 ns NAMD runs were performed to show that an all-atom trajectory starts and ends in the same basin (Figures 6 and 4c). This MD sampling

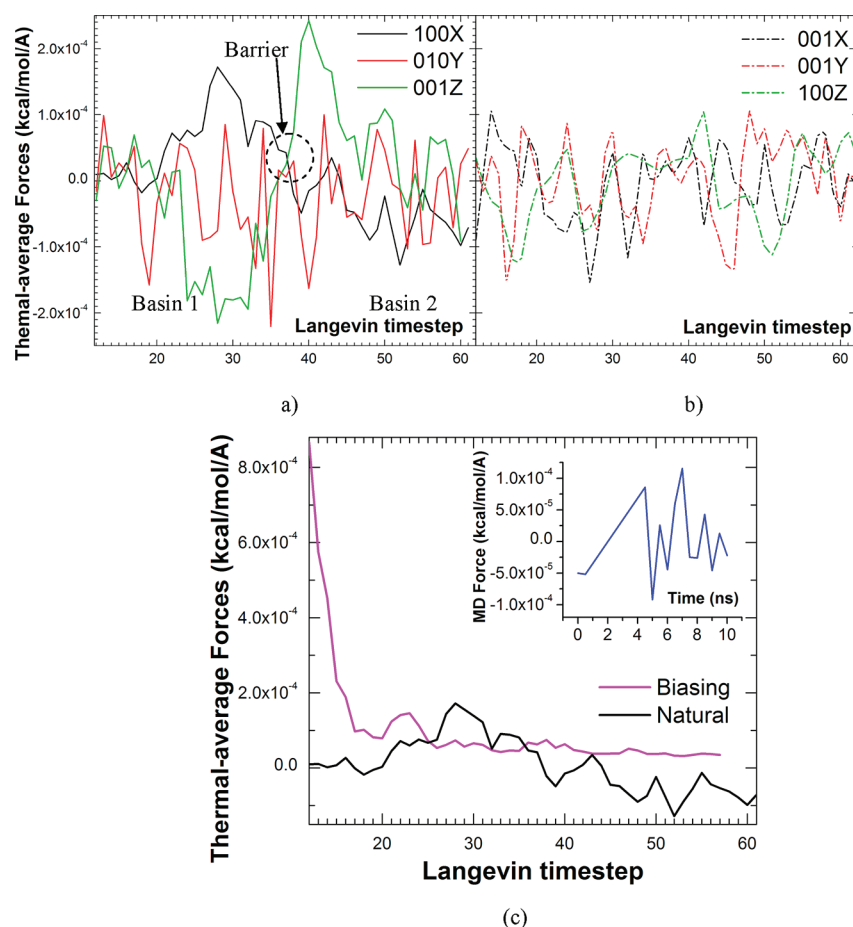


Figure 4. Thermal-average forces \vec{f}_k are shown along a transition path from basin 1 to 2: (a) forces of maximum magnitude showing clear interbasin transition pattern and (b) second-highest amplitude ones oscillating around zero. Values of k are shown in legends. The figure suggests that the FE minimizing tendency of lactoferrin in basin 1 makes it contract in the z -direction and expand in the x -direction (implied by the negative f_{001Z} and positive f_{100X}). During barrier crossing, the sign of these forces changes to positive along z and negative along the x direction. This leads to further expansion of lactoferrin along the z direction and hence lobe opening in basin 2. (c) Contrasting behavior of thermal-average f_{100X} (black) and biasing f_{100X}^b (magenta) forces. f_{100X}^b is maximum near the bottom of the basin and gradually decreases as the system escapes from the FE minimum as indicated by the increase in f_{100X} . Inset shows that f_{100X} computed from MD simulations are small and random when sampling structures in basin 1 and, therefore, do not drive transitions between basins.

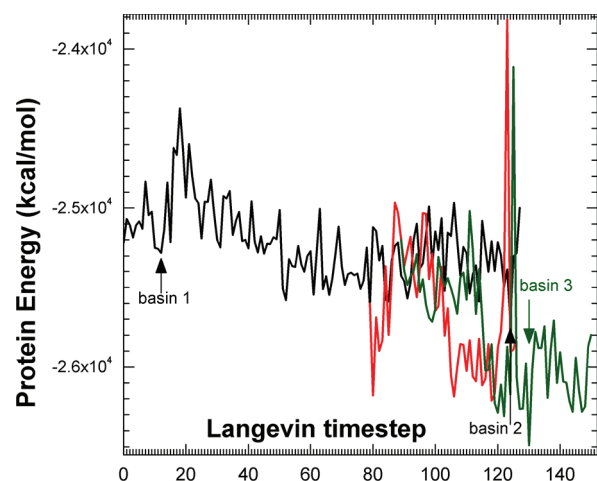


Figure 5. Energy timecourse of lowest potential energy structures of human lactoferrin generated from constant OP ensembles during the discovery of FE basins. Line styles and simulations are the same as in Figure 3.

validates our method, i.e., trajectories remain for long times in a given FE basin. That a trajectory remains in a basin is indicated

by the fact that the OPs do not change appreciably over their timecourse. In these samplings one does not obtain structures whose set of descriptors (and, therefore, OPs) falls in the domain sampled by MD in any other basin (Figure S2, Supporting Information). In addition, the DMS.BD is robust to the choice of initial all-atom structure. The analysis described in the fifth subsection of the Supporting Information suggests that DMS.BD can guide a system away from a known basin through the arbitrary choice of initial structure, which does not necessarily characterize the bottom of the basin. In this context, we probe the basin 2 to 3 transition using an arbitrary initial structure denoted “descent 2”.

We compare our results with those from experimental and previous theoretical observations. Transition of lactoferrin from the diferric to apolactoferrin states is accompanied by changes in the vicinity of residues THR90 and VAL250. These residues act as hinges that facilitate the lobe-opening transition.⁵¹ We observe substantial differences in the backbone dihedral angles of these residues between the closed state and the discovered slightly open one (Figure S1, Supporting Information). In particular, more differences in dihedral angles are observed for residues in the loop region than in the highly structured parts of

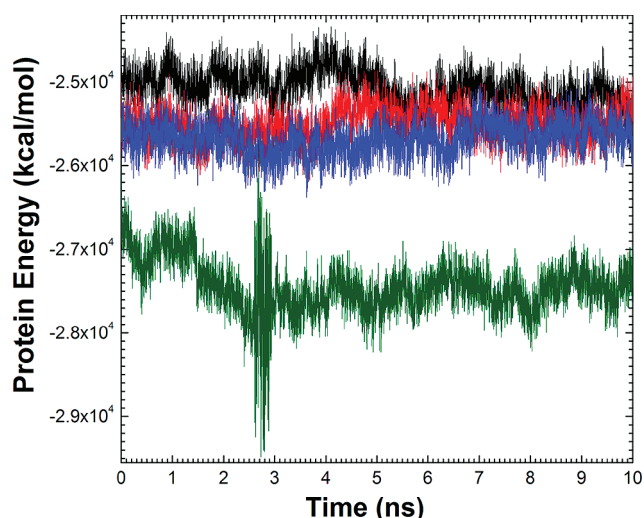


Figure 6. Potential energy timecourse of lactoferrin from MD sampling simulations starting at structures near the bottom of discovered basins. (black) Basin 1, potential energy minimum of basin 1 was achieved at time instance $t = 9.940$ ns; (red) “descent 2” structure from basin 2, with a potential energy minimum at $t = 3.686$ ns; (blue) lowest-energy basin 2 structure, with a potential energy minimum at $t = 3.621$ ns; (green) basin 3, with a potential energy minimum at $t = 2.733$ ns. Distinct energy bands indicate that each MD trajectory is confined to a given basin.

lactoferrin. This validates that most of the secondary structure is preserved during the transition, as has been suggested by previous theoretical results.¹⁹ A residue-by-residue rmsd comparison of the backbone C_{α} atoms between structures from basin 1, 2, and 3 with respect to that of the X-ray structure 1LFG of closed lactoferrin (Figure 7) is performed. To

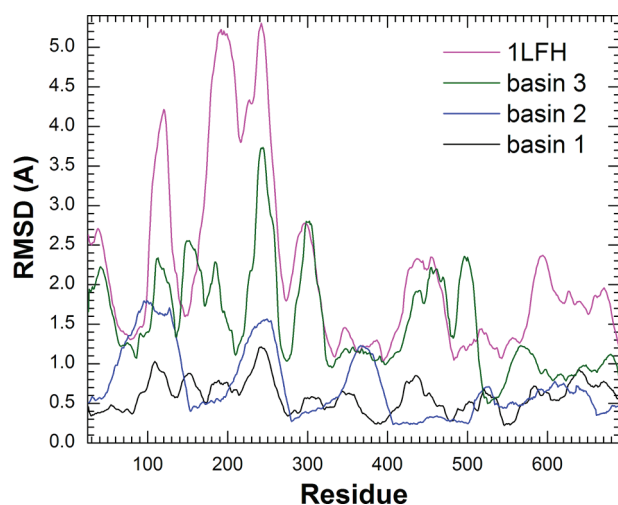


Figure 7. Residue-by-residue rmsd relative to closed-lobe diferric X-ray structure 1LFG averaged over a 25 residue window: (black) diferric basin 1, (blue) pseudodiferric basin 2, (green) open-lobe apo-like basin 3, and (magenta) the apolactoferrin structure 1LFH. This analysis implies that the lactoferrin gradually opens during DMS.BD simulation, exploring a range of states from diferric to apolactoferrin character.

understand differences with the apolactoferrin structure 1LFH, we also plot the rmsd between diferric and apolactoferrin structures. The rmsd gradually increases from basin 1 to 3 indicating lobe opening. The deviations are significant in the

vicinity of residues 90 and 250, indicating that the hinge motion is captured through DMS.BD simulations. Thus, the DMS.BD predicted FE minimizing structures approach the experimentally observed open state 1LFH.

V. CONCLUSIONS

A methodology for the sequential discovery of FE basins for macromolecular systems was presented. Structural information from known basins is used to escape/avoid them and thereby enable the discovery of yet-unknown basins. The approach was implemented via our DMS software and validated using two X-ray structures for human lactoferrin. Two new FE basins were discovered. The method has the potential for discovering pathways of transitions between basins, including estimates of FE barriers along the transition paths. Comparison of nano-characterization data with values calculated for the discovered all-atom states provides an approach for interpretation of such data. One example of nanocharacterization data to which this approach can be applied is collision cross sections from ion mobility–mass spectroscopy experiments for charged biomolecules.

The basin discovery algorithm is built on multiscale techniques. The latter provide orders of magnitude increase in the efficiency of simulation for large macromolecular assemblies.^{5,26} These efficiencies allow the methodology and the implementation of interest in biophysical studies such as on structural transitions in viruses.^{33,52}

For high temperature, the distribution of likely states within a FE basin is very broad, and therefore, the basin becomes less well-defined. In particular, FE barriers that would otherwise sequester all-atom trajectories to lie within the basin are lower, enabling more frequent escape. It was shown here that descriptors chosen at the state of minimum FE in the basin can be used to guide multiscale simulations from known to yet-unknown ones (Section IV).

The present method achieves system evolution and FE landscape exploration via a trifold approach. OPs provide the coarse-grained description via an expression that facilitates the construction of the ensemble of all-atom states consistent with the instantaneous OP values. However, as the system departs significantly from an initial reference all-atom structure, the OPs may not provide a viable description. Thus, in our implementation a new all-atom reference configuration and resulting newly defined OPs are established when needed. This implies that the present OPs do not serve as an appropriate coarse-grained description for mapping the broader FE landscape. In contrast, the system descriptors can serve as the coarse-grained state variables with which to define the landscape since their definition does not involve a reference configuration. However, the descriptors do not provide a convenient way to generate the ensembles of all-atom states needed to construct the thermal forces and diffusion factors mediating the evolution of the coarse-grained state. Thus, the present OPs facilitate ensemble generation and coarse-grained evolution; the descriptors provide a coarse-grained variable for a continuous mapping of the FE landscape despite the changing definition of the OPs. Thus, our method integrates the OPs, the descriptors, and ensembles of all-atom states to enable multiscale simulations across a FE landscape. This is the logic behind our trifold simulation and basin discovery approach.

■ ASSOCIATED CONTENT

■ Supporting Information

MD settings, relationship between descriptors and OPs, details on DMS.BD workflow and simulations performed, FE expression via thermal-average forces, table showing characteristics of the discovered structures for human lactoferrin, figures showing Ramachandran plots for some of these structures, and timecourse of descriptors sampled by MD. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ortoleva@indiana.edu. Phone: 1 812 856 6000. Fax: 1 812 855 8300.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This project was supported in part by the National Science Foundation (Collaborative Research in Chemistry program, award #0832651), National Institute of Health (NIBIB), METAcyt, Indiana University College of Arts and Sciences through the Center for Cell and Virus Theory, and the Indiana University office of the Vice President for Research (IU Collaborative Research Grant, "All-atom Theory of Virus Behavior: Applications to Vaccine Discovery"). The authors appreciate computing time provided by Indiana University on the Big Red supercomputer.

■ REFERENCES

- (1) Ruotolo, B. T.; Giles, K.; Campuzano, I.; Sandercock, A. M.; Bateman, R. H.; Robinson, C. V. *Science* **2005**, *310*, 1658–1661.
- (2) Binning, G.; Quate, C. F.; Gerber, C. *Phys. Rev. Lett.* **1986**, *56*, 930.
- (3) Beardsley, R. L.; Running, W. E.; Reilly, J. P. *J. Proteome Res.* **2006**, *5*, 2935–2946.
- (4) Keyser, U. F.; Koeleman, B. N.; van Dorp, S.; Krapf, D.; Smeets, R. M. M.; Lemay, S. G.; Dekker, N. H.; Dekker, C. *Nat. Phys.* **2006**, *2*, 473–477.
- (5) Joshi, H.; Singharoy, A. B.; Sereda, Y. V.; Chelvaraja, S. C.; Ortoleva, P. J. *Prog. Biophys. Mol. Biol.* **2011**, *107*, 200–217.
- (6) Rangwala, H.; Karypis, G. Introduction to Protein Structure Prediction. In *Introduction to Protein Structure Prediction*; John Wiley & Sons, Inc.: New York, 2010; pp 1–13.
- (7) Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. *Nature* **1992**, *356*, 539–542.
- (8) Georgiev, I.; Donald, B. R. *Bioinformatics* **2007**, *23*, i185–i194.
- (9) Heath, A. P.; Kavrakli, L. E.; Clementi, C. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 646–661.
- (10) Bower, M.; Cohen, F.; Dunbrack, R. J. *Mol. Biol.* **1997**, *267*, 1268–1282.
- (11) Lee, C.; Subbiah, S. *J. Mol. Biol.* **1991**, *217*, 373–388.
- (12) Lee, C. *J. Mol. Biol.* **1994**, *236*, 918–939.
- (13) Floudas, C. A. *Biotechnol. Bioeng.* **2007**, *97*, 207–213.
- (14) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.
- (15) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671–680.
- (16) Lee, C.; Levitt, M. *Nature* **1991**, *352*, 448–451.
- (17) Holland, J. H. *Adaptation in natural and artificial systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
- (18) Unger, R. The Genetic Algorithm Approach to Protein Structure Prediction. In *Applications of Evolutionary Computation in Chemistry*; Johnston, R. L., Ed.; Springer: Berlin/Heidelberg, 2004; Vol. 110, pp 2697–2699.

- (19) Kim, M. K.; Jernigan, R. L.; Chirikjian, G. S. *Biophys. J.* **2005**, *89*, 43–55.
- (20) Marques, O.; Sanejouand, Y.-H. *Proteins: Struct., Funct., Bioinf.* **1995**, *23*, 557–560.
- (21) Dellago, C.; Bolhuis, P. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*; Holm, C., Kremer, K., Eds.; Springer: Berlin/Heidelberg, 2009; Vol. 221, pp 167–233.
- (22) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
- (23) Barducci, A.; Bonomi, M.; Parrinello, M. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 826–843.
- (24) Trabuco, L. G.; Villa, E.; Schreiner, E.; Harrison, C. B.; Schulten, K. *Methods* **2009**, *49*, 174–180.
- (25) Chelvaraja, S.; Ortoleva, P. *J. Chem. Phys.* **2010**, *132*, 075102.
- (26) Singharoy, A.; Chelvaraja, S.; Ortoleva, P. *J. Chem. Phys.* **2011**, *134*, 044104.
- (27) Singharoy, A.; Joshi, H.; Chelvaraja, S.; Brown, D.; Ortoleva, P. J. Simulating Microbial Systems: Addressing Model uncertainty/incompleteness via Multiscaling and Entropy methods. In *Microbial Systems Biology: Methods and Protocols*; Navid, A., Ed.; Springer Science: New York, 2010; Vol. 881.
- (28) Jaqaman, K.; Ortoleva, P. J. *J. Comput. Chem.* **2002**, *23*, 484–491.
- (29) Miao, Y.; Ortoleva, P. J. *J. Comput. Chem.* **2009**, *30*, 423–437.
- (30) Pankavich, S.; Miao, Y.; Ortoleva, J.; Shreif, Z.; Ortoleva, P. J. *J. Chem. Phys.* **2008**, *128*, 234908–234920.
- (31) Schmidt, E. *Math. Ann.* **1907**, *63*, 433–476.
- (32) Singharoy, A.; Sereda, Y. V.; Ortoleva, P. J. *J. Chem. Theory Comput.* **2012**, DOI: <http://dx.doi.org/10.1021/ct200574x>.
- (33) Miao, Y.; Ortoleva, P. J. *Biopolymers* **2010**, *93*, 61–73.
- (34) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (35) Shreif, Z.; Ortoleva, P. J. *NSTI Nanotech. 2008, Nanotechnol. Conf. Trade Show, Tech. Proc.* **2008**, *3*, 741–744.
- (36) Shreif, Z.; Ortoleva, P. J. *Comput. Math. Methods Med.* **2009**, *10*, 49–70.
- (37) Ortoleva, P. J.; Adhangale, P.; Chelvaraja, S.; Fontus, M. W. A.; Shreif, Z. *IEEE Eng. Med. Biol.* **2009**, *28*, 70–79.
- (38) Pankavich, S.; Shreif, Z.; Ortoleva, P. J. *Phys. A* **2008**, *387*, 4053–4069.
- (39) Shreif, Z.; Ortoleva, P. J. *Phys. A* **2009**, *388*, 593–600.
- (40) Ortoleva, P. J. *J. Phys. Chem. B* **2005**, *109*, 21258–21266.
- (41) Miao, Y.; Ortoleva, P. J. *J. Chem. Phys.* **2006**, *125*, 44901–44908.
- (42) Bose, S.; Ortoleva, P. J. *J. Chem. Phys.* **1979**, *70*, 3041–3056.
- (43) Bose, S.; Ortoleva, P. J. *Phys. Lett. A* **1979**, *69*, 367–369.
- (44) Bose, S.; Medina-Noyola, M.; Ortoleva, P. J. *J. Chem. Phys.* **1981**, *75*, 1762–1771.
- (45) Shreif, Z.; Ortoleva, P. J. *Stat. Phys.* **2008**, *130*, 669–685.
- (46) Pankavich, S.; Shreif, Z.; Miao, Y.; Ortoleva, P. J. *J. Chem. Phys.* **2009**, *130*, 194115–194124.
- (47) Shreif, Z.; Pankavich, S.; Ortoleva, P. J. *Phys. Rev. E* **2009**, *80*, 031703.
- (48) Pankavich, S.; Ortoleva, P. J. *Math. Phys.* **2010**, *51*, 063303–063316.
- (49) Li, Z.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611–6615.
- (50) Norris, G. E.; Anderson, B. F.; Baker, E. N. *Acta Crystallogr., Sect. B* **1991**, *47*, 998–1004.
- (51) Gerstein, M.; Anderson, B. F.; Norris, G. E.; Baker, E. N.; Lesk, A. M.; Chothia, C. *J. Mol. Biol.* **1993**, *234*, 357–372.
- (52) Miao, Y.; Johnson, J. E.; Ortoleva, P. J. *J. Phys. Chem. B* **2010**, *114*, 11181–11195.