Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2012, Article ID 674832, 19 pages
doi:10.1155/2012/674832

Research Article

Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules

Brian Olson,[1] Irina Hashmi,[1] Kevin Molloy,[1] and Amarda Shehu[1,2]

[1] *Department of Computer Science, George Mason University, Fairfax, VA 22030, USA*
[2] *Department of Bioengineering, George Mason University, Fairfax, VA 22030, USA*

Correspondence should be addressed to Amarda Shehu, amarda@gmu.edu

Received 29 June 2012; Revised 23 September 2012; Accepted 19 October 2012

Academic Editor: Zhiyuan Luo

Since its introduction, the basin hopping (BH) framework has proven useful for hard nonlinear optimization problems with multiple variables and modalities. Applications span a wide range, from packing problems in geometry to characterization of molecular states in statistical physics. BH is seeing a reemergence in computational structural biology due to its ability to obtain a coarse-grained representation of the protein energy surface in terms of local minima. In this paper, we show that the BH framework is general and versatile, allowing to address problems related to the characterization of protein structure, assembly, and motion due to its fundamental ability to sample minima in a high-dimensional variable space. We show how specific implementations of the main components in BH yield algorithmic realizations that attain state-of-the-art results in the context of ab initio protein structure prediction and rigid protein-protein docking. We also show that BH can map intermediate minima related with motions connecting diverse stable functionally relevant states in a protein molecule, thus serving as a first step towards the characterization of transition trajectories connecting these states.

1. Introduction

Global optimization is an objective of many disciplines, both in academic and industrial settings [1, 2]. Characterization of complex systems often poses very hard global optimization problems with many variables [3, 4]. Algorithms that target such problems largely build on or combine four main approaches: deterministic, stochastic, heuristic, and smoothing [3, 5–7]. All these algorithms are challenged by systems where the variable space contains multiple distinct minima. While most algorithms can efficiently find a minimum, not all can feasibly locate the global minimum.

Some of the most successful applications of global optimization algorithms on characterizing physical and biological systems build on the stochastic Monte Carlo (MC) procedure and its Metropolis variant [8]. For instance, simulated annealing is one of the most widely used algorithms for finding the global minimum of a multivariable function for different complex systems [4, 9, 10]. Adaptations that build on deterministic and stochastic numerical procedures, such as molecular dynamics (MD) and MC, are abundant in computational biology for the structural characterization of biological macromolecules (cf. [11, 12]).

Basin hopping (BH) is a global optimization framework that is particularly suited for multivariable multimodal optimization problems [13], and it is our thesis in this paper that BH is an effective framework for the characterization of biological macromolecules. The basic BH framework is well studied and understood, but modifications to its core components are necessary for application to complex biological systems. In what follows, we first summarize the basic BH framework and some of its salient properties before proceeding to identify modifications necessary for application to biological macromolecules.

BH combines heuristic procedures with local searches to enhance its exploration of the given variable space, conducted as a series of perturbations followed by local optimization. As shown in pseudocode in Algorithm 1,

```
(1) i ← 0
(2) X_i ← random initial point in variable space
(3) Y_i ← LOCALSEARCH (X_i)
(4) while STOP not satisfied do
(5)     X_{i+1} ← PERTURB (Y_i)
(6)     Y_{i+1} ← LOCALSEARCH (X_{i+1})
(7)     if f(Y_{i+1}) < f(Y_i) then
(8)         i ← i + 1
```

Algorithm 1: Basic BH framework in pseudocode.

the framework can be described in terms of a local search procedure LOCALSEARCH that maps a point X_i in variable space to its nearest minimum Y_i, a perturbation move PERTURB that modifies a current minimum Y_i to obtain a new point X_{i+1} in variable space, and a stopping criterion STOP that terminates these repeated applications of a structural perturbation followed by a local optimization. The repeated applications result in a trajectory of local minima Y_i. As shown in Algorithm 1, only the lowest minimum needs to be retained in memory when seeking the global minimum of some function f. It is important to note that Algorithm 1 shows a specific realization of the BH framework, known as monotonic BH (MBH), where the current minimum is not accepted if it does not lower the lowest value obtained for the function f so far. In this case, another perturbation is attempted in order to obtain a new starting point for the local optimization that follows.

While this basic framework is easy to describe and employ for global optimization, effective implementations exploit specific domain expertise about the system at hand [14–19]. Heuristics are designed based on specific system knowledge to implement an effective perturbation component. Domain-specific expertise is also employed for an effective implementation of the local search component. The stopping criterion is often implemented in terms of a maximum number of function evaluations or in terms of no improvements over a window of the last sampled minima. It is important to note that the stochasticity in BH is mainly due to the implementation of the perturbation component, which seeks to take the exploration out of the current local minimum. The local optimization component, on the other hand, can employ deterministic numerical techniques to locate the local minimum with arbitrary accuracy [20].

The core advantage of the BH framework over a multistart method that essentially samples local minima at random is that BH moves between adjacent local minima in the variable space. This strategy is more effective when exploring high-dimensional variable spaces associated with complex physical systems, where the addition of new dimensions can result in an exponential increase in the number of minima in the space [21]. The adjacency is a result of a deep connection between the perturbation and local optimization. Despite the application setting, a good general rule is for the perturbation to preserve some structural characteristics of the local minimum Y_i it is disrupting to obtain a new starting point X_{i+1} for the next application of the local optimization.

If the magnitude of the perturbation jump in the variable space, measured through some distance function $d(Y_i, X_{i+1})$, is small, then X_{i+1} may remain in the basin of attraction of Y_i, and the local optimization will bring X_{i+1} back to Y_i. On the other end of the spectrum, the perturbation can completely disrupt Y_i and obtain an X_{i+1} that could have essentially been obtained at random. While the local optimization will yield a new local minimum $Y_{i+1} \neq Y_i$, the BH will degenerate to a multistart method in this case. Different studies have shown that the perturbation needs to preserve some of the structure of the current minimum for BH to be more effective than the multistart method [20, 22]. It is the careful implementation of the perturbation component that allows BH to organize the local minima it samples according to an adjacency relationship [21].

The BH framework is sometimes referred to as a funnel-descent method, because its core behavior of iterating over adjacent local minima has turned out to be an effective optimization strategy for functions with a funnel landscape [21]. The generality of the framework and its ease of adaption for different systems has resulted in diverse applications, which span from geometry problems, such as packing circles in circular containers [20], to statistical physics problems of characterizing low-energy states of small atomic clusters [3].

The BH framework originated in the computational biology community dating back to the pioneering work of Wales [3], where the objective was to characterize the minima of the Lennard-Jones energy function in small atomic clusters. The term basin hopping was coined in this work, though to an extent, the stated motivation for the BH framework was from related optimization algorithms in the evolutionary search community. In fact, BH can be viewed as a special case of Iterated Local Search, which is popular for solving discrete combinatorial optimization problems [23]. An algorithmic realization of the BH framework was available prior to the work of Wales, most notably in Scheraga's MC with minimization algorithm [9, 24].

The BH framework is particularly suited to deal with molecular spaces, where the function sought for optimization is a complex nonconvex potential energy function summing over the interactions among atoms in a 3-dimensional molecular structure. The global minimum of the function corresponds to the structural state of the molecule that is most stable under equilibrium conditions and so relevant for biological activity. Structural characterization of the biologically active (native) state of biological macromolecules is

an important problem in computational structural biology. A grand-standing challenge nowadays is to characterize such states for protein molecules, which are central in many chemical pathways in the cell and are the focus of this paper.

Proteins are complex systems with hundreds to thousands of atoms. These atoms are organized in amino-acid building blocks which connect serially to form a polypeptide chain (the N-terminus of one amino acid connects to the C-terminus of the other to form a peptide). Figure 1(a) shows a short polypeptide chain. Depending on the representation employed, a spatial arrangement of the atoms that constitute a polypeptide chain, also referred to as a conformation, may require the specification of a prohibitive number of variables. A popular representation in computational structural biology employs only the angles shown in Figure 1(a). These angles can be used to define the variable space, as their modification gives rise to different conformations.

The variable, or conformational, space of a polypeptide chain is associated with a funnel-like energy surface [25, 26]. The size and ruggedness of this surface, illustrated in Figure 1(b), are the primary reasons why obtaining structural information on native state of a protein polypeptide chain based on the chain's amino-acid sequence alone is an outstanding challenge in computational structural biology [27]. Meeting this challenge, often known as ab initio protein structure prediction, is needed, however, to close the gap between the wealth of protein sequence data and the scarce information on their native structures. Obtaining structural information ab initio promises to elucidate the structure-function relationship and advance structure-driven studies and applications on protein molecules [28–30].

The funnel-like but rugged energy surface of protein molecules seems suitable for the BH framework. In general, it is challenging to locate the global minimum in this surface and so elucidate the native structure of a protein. One of the main reasons relates to imperfect modeling. The energy functions currently available to probe the protein energy surface are semiempirical and contain inherent errors [31]. Due to the specific process undertaken in computational chemistry to design such functions, the actual global minimum of a designed protein energy function may deviate significantly from the true global minimum (the native structure obtained by experiment in the wet laboratory). Studies report deviations in the 2–4 Å range [32]. Due to these deviations, computational approaches that aim to obtain a broad view of the energy surface are more appropriate, particularly if they are to be followed by detailed heavy-duty optimization techniques on select conformations.

A common strategy among protocols for ab-initio protein structure prediction is the sampling of a large number of low-energy conformations. These are end points of many independent MD or MC trajectories optimizing some chosen energy function [28, 33–39]. Alternatively, the trajectories can be integrated in a tree to better control the exploration and use online analysis to bias the tree away from high-energy oversampled regions [40, 41]. The conformations are then grouped by structural similarity to reveal local minima from which it is worth continuing the exploration at higher representational detail. The goal then becomes obtaining convergence to a region of the space that can be predicted to represent the native state.

In the context of ab-initio protein structure prediction, BH can be employed to explicitly sample local minima in the protein energy surface. At a superficial level, this would require the retainment of an ensemble of local minima and not just the current one. In addition, while the pseudocode in Algorithm 1 shows a simple realization of the BH framework, MBH, applications of BH on molecular spaces often make probabilistic decisions on whether to accept a current minimum. Procedurally, the framework still consists of repeated applications of a structural perturbation followed by an energy minimization. However, a Metropolis criterion [42] biases the sampling of local minima towards lower energy ones over time. Essentially, the decision to accept Y_{i+1} is made with probability $\exp(-[E(Y_{i+1})-E(Y_i)]/[K_B T])$, where E refers to the energy function, K_B is the Boltzmann constant, and T is temperature. Temperature does not need a physical meaning, as its main role is to scale the height of an energy barrier.

The appeal of the BH framework is that it transforms the protein energy surface into a collection of interpenetrating staircases, as illustrated in Figure 1(c). A succinct discrete representation is obtained for this surface in terms of local minima. It is important to note that BH does not modify the energy surface in any way. Instead, it projects each point (conformation) to its closest local minimum to effectively reveal a map of the energy surface in terms of local minima. The details of the energy surface between local minima are lost, but this degree of resolution is still very useful for a structural characterization of protein molecules.

Given its ease of implementation, BH is starting to gain popularity as an optimization framework for biological systems. Current applications of BH for structural characterization of biological molecules essentially differ in the specific implementations for the perturbation and local optimization components. Local optimization, for instance, is implemented as gradient descent or Metropolis MC at low temperature, whereas the perturbation component, on the other hand, directly modifies atomic coordinates in existing work. These implementation choices have allowed BH algorithms to capture local minima of small atomic clusters and even map energy surfaces of polyalanines and other small proteins [14, 32, 43, 44]. However, applications to structure prediction [45] have been limited to small proteins, mainly because representation of conformations through atomic coordinates results in a prohibitive variable space. In particular, the BH algorithm in [45] succeeds in locating conformations closer to the experimentally determined native structure than MD with simulated annealing, but its efficiency drops on sequences longer than 75 amino acids.

In this paper, we show that BH is a useful framework for structural characterization beyond structure prediction. We recognize that BH is general and can be employed to map the equilibrium conformational space of a biological system. For instance, we show that with suitable modifications to the perturbation and local optimization components, BH can be applied to protein-protein docking to reveal native

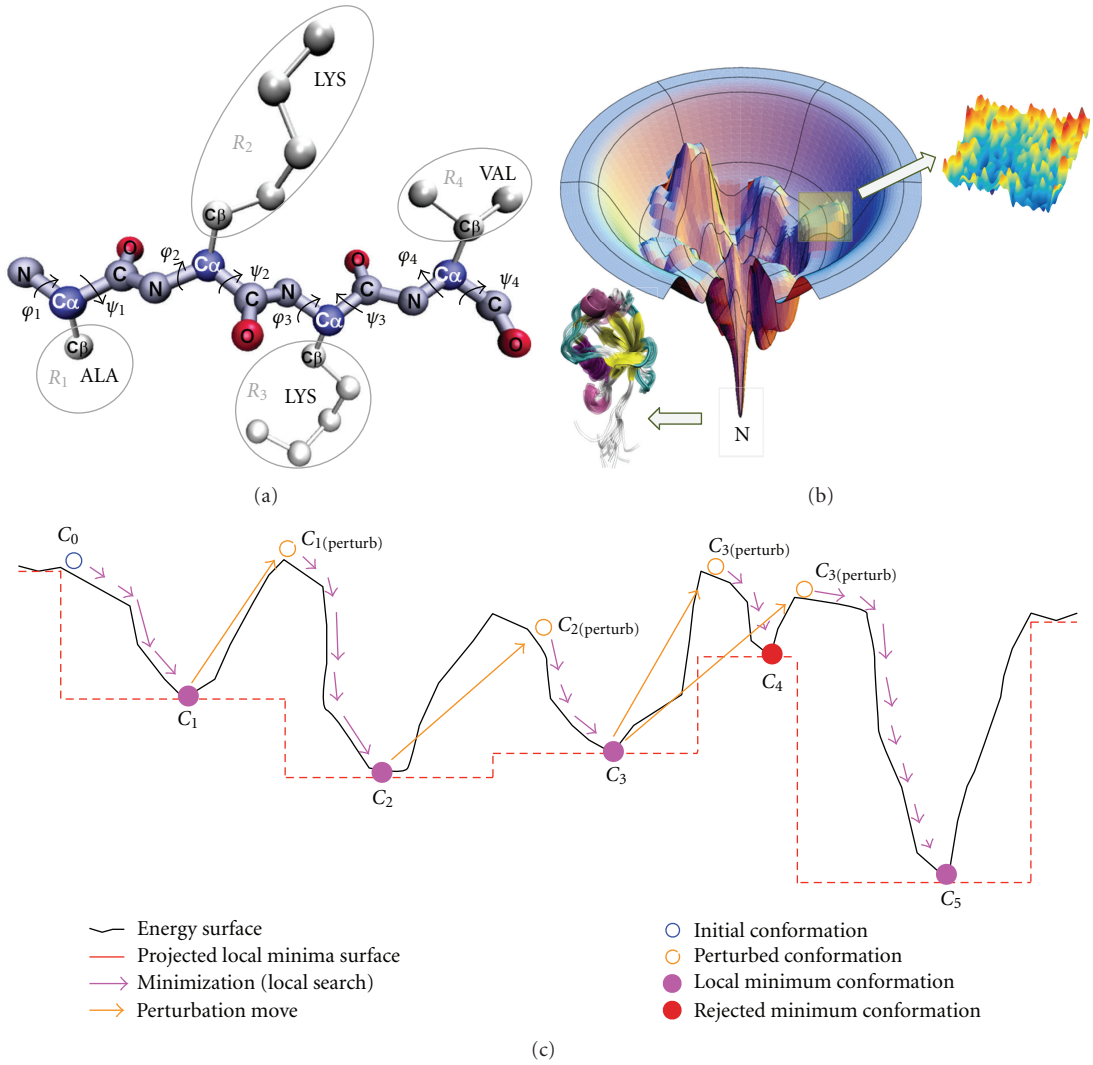

FIGURE 1: (a) A short polypeptide chain of 4 amino acids, alanine, lysine, lysine, and valine, is shown. The backbone atoms shared by all amino acids are N, C_α, C, and O. Side-chain atoms unique to 20 types of amino acids are in gray. The backbone (ϕ, ψ) dihedral angles are annotated over the chain. (b) A model energy surface is illustrated, adapted from [25]. The surface is funnel-like but rugged. The native state at the bottom is denoted by N. Conformations associated with it (obtained from experiment) are illustrated for a particular protein molecule. (c) The BH framework essentially converts the function into a stepwise one. The perturbation and local optimization components are illustrated here with differently colored arrows. A minimum is shown here which fails the Metropolis criterion and is thus not accepted, prompting a new perturbation move.

lowest-energy configurations of protein molecules resulting from the assembly of various polypeptide chains. Specifically, in the context of ab-initio protein structure prediction, we show that implementations of the main components in BH that employ domain-specific knowledge result in increased efficiency and allow application to longer protein chains.

We also show that the ability of BH to provide a map of the energy surface in terms of minima is useful not only when the goal is to locate the global minimum (whether that minimum corresponds to the native structure of one protein polypeptide chain or of a complex resulting from assembly of multiple chains), but also when the objective is to characterize proteins with more than one functionally relevant state. Such proteins are abundant in biology as effective biological machines that can tune their biological function through molecular motions [46–49]. A map of the minima surrounding the functionally relevant states is useful for understanding how the protein hops between minima in transition trajectories connecting these states [33, 49, 50].

The presentation in this paper of BH as a general, versatile framework builds over our recent work on ab-initio structure prediction and rigid protein-protein docking [22, 51]. In particular, in the context of structure prediction, we show that employment of the molecular fragment replacement technique allows BH to efficiently capture the native structure. In the context of protein-protein docking, we incorporate geometric hashing to efficiently obtain structural perturbations of a dimeric configuration.

Additional information from evolutionary sequence analysis allows restricting the variable space. Finally, we provide here a proof-of-concept demonstration that BH can be applied to understand the connectivity between functionally relevant states in a protein in terms of the minima surrounding these states. Obtaining a view of minima in the equilibrium conformational space of a protein molecule is the first step into elucidating motions and transition trajectories that take a protein between the states it uses for biological function.

2. Methods

The basic BH framework was showcased in Algorithm 1 in Section 1. Our algorithmic treatment in this section focuses on modifications to the basic components of BH which allow its application to the three different problems on which we focus in this paper. As described in Section 1, two modifications that allow application to these problems concern accepting a newly obtained local minimum according to the Metropolis criterion (unlike the basic MBH algorithm) and adding that minimum to a growing ensemble of BH-obtained local minima (unlike recording only the last one as in the basic MBH framework). The description of BH below is organized according to the three different applications showcased in this paper in Sections 2.1, 2.2, and 2.3, respectively. The treatment of BH in each application is limited to description of four main components: (1) representation of the system being modeled, which allows defining the variable space; (2) description of the energy function being optimized by BH; (3) implementation of the structural perturbation move; and (4) implementation of the local search procedure for the local optimization.

2.1. BH for Sampling Decoy Conformations for Ab Initio Protein. As described above, the BH framework can be employed to obtain a broad view of the energy surface in terms of low-energy local minima. This can be done efficiently at a coarse-grained level of detail, employing an energy function that sacrifices detail and some accuracy to save computational time. The sampled conformations corresponding to the local minima are low-energy decoy conformations, which can then be fed to any structure prediction protocol for further analysis and refinement of select conformations with dedicated computational resources. The refinement will allow adding further detail, discriminating between decoy conformations, and making a prediction on which refined conformation can be considered to represent the native structure.

2.1.1. Employed Representation. As illustrated in Figure 1(a), a polypeptide chain of n amino acids contains $2n$ backbone (ϕ, ψ) dihedral angles. Our representation of a protein conformation employs only these angles, which constitute the variable space. Side chains are sacrificed, as any structure prediction protocol can pack them as part of the ensuing refinement of decoy conformations [52]. The representation here is essentially the idealized geometry model, which fixes bond lengths and angles to idealized (native) values. Forward kinematics allows computing Cartesian coordinates of the backbone atoms (on which the energy function described below operates) from the ϕ, ψ angles in the representation [53].

2.1.2. Energy Function. The energy function is a modification of the associative memory Hamiltonian with water (AMW) [54]. This function has been used previously by us and others in the context of ab-initio structure prediction [40, 41, 55–57]. AMW sums nonlocal terms (local interactions are kept at ideal values in the idealized geometry model): $E_{\text{AMW}} = E_{\text{Lennard-Jones}} + E_{\text{H-Bond}} + E_{\text{contact}} + E_{\text{burial}} + E_{\text{water}} + E_{\text{Rg}}$. The $E_{\text{Lennard-Jones}}$ term is implemented after the 12–6 Lennard-Jones potential in AMBER9 [58] but allows a soft penetration of van der Waals spheres. The $E_{\text{H-Bond}}$ term allows modeling hydrogen bonds and is implemented as in [59]. The other terms, E_{contact}, E_{burial}, and E_{water}, allow formation of nonlocal contacts, a hydrophobic core, and water-mediated interactions, and are implemented as in [39].

The listed energy terms of E_{AMW} sum over pairwise interactions. For instance, the 12–6 functional form of the Lennard-Jones term is $-4\epsilon_{ij}[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}]$, where ϵ_{ij} is a constant characteristic of the types of atoms at positions i and j, σ_{ij} is the average diameter of the atoms, and r_{ij} is their distance. This functional form illustrates the quadratic running time of a typical energy function modeling pairwise interactions. More importantly, the terms summed together in an energy function are competing; minima of one term are obtained by suboptima of the other. This competition is known as frustration and refers to the fact that slight changes in atomic positions may lower the value of one term but increase that of another term. The result of summing competing terms in an energy function is a complex multimodal function, whose optimization is nontrivial. More details on the functional form of the other terms of the AMW energy function can be found in [33, 54].

2.1.3. Implementation of Structural Perturbation. The realization of the BH framework we describe here hops between two conformations representing two consecutive minima C_i and C_{i+1} through an intermediate $C_{\text{perturb},i}$ conformation. The perturbation modifies C_i to obtain a higher-energy conformation $C_{\text{perturb},i}$ to escape the current minimum. Essentially, 6 backbone dihedral angles of a fragment of the polypeptide chain associated with three consecutive amino acids in the current conformation C_i are modified simultaneously. This process is referred to as the molecular fragment replacement technique, because it allows replacing the current configuration (in terms of angles) of a selected fragment with another fragment configuration [60].

Molecular Fragment Replacement. The fragment replacement technique has allowed ab-initio structure prediction methods to make great advancements [28, 34–37]. Its key advantage is that it allows obtaining physically realistic modifications if the fragment configurations are sampled from a library of actual native structures obtained in the wet laboratory. The basic idea is that a subset of nonredundant

protein structures are obtained from the Protein Data Bank [61], and configurations of all fragments that can be defined for k consecutive amino acids are excised from these structures and stored in a library. We direct the reader to [28, 41] for a detailed description of how the library is constructed. In this work, we employ a fragment of length 3 rather than a longer fragment, so that the magnitude of the jump resulting from the fragment replacement in variable space is limited.

The perturbation component is implemented as follows. Given a conformation C_i, a fragment of length 3 is selected at random over the polypeptide chain ($n - 2$ fragments can be defined with overlap over a chain of n amino acids). Once the fragment is selected, a configuration for that fragment is then sampled at random over those available for the fragment from the fragment configuration library. The replacement of the angles of the fragment in C_i with those of the configuration obtained from library results in $C_{\text{perturb},i}$.

Since low-energy conformations tend to be compact and leave little room for movement without raising energy (a concept known as frustration in protein biophysics), this implementation of the perturbation component is sufficient to obtain a high-energy conformation through which to escape the current local minimum. Additionally, $C_{\text{perturb},i}$ will share nearly all of its local structural features with C_i, but the new conformation will have a higher energy and a different overall global structure. We note that the first conformation to initiate BH is obtained after $n - 2$ fragment configuration replacements over an extended conformation.

2.1.4. Implementation of Local Optimization. Our implementation of the local optimization conducts a series of modifications starting from $C_{\text{perturb},i}$ to reach a new minimum C_{i+1}. While numerical techniques can be used here, they tend to be inefficient [45]. We employ instead a greedy search, which essentially attempts a maximum of m consecutive fragment replacements (as described above for the perturbation component) until k consecutive attempts fail to lower energy. The resulting C_{i+1} conformation is added to the trajectory according to the Metropolis criterion based on the energetic difference with C_i.

Our implementation of the local optimization is probabilistic due to the fragment replacement technique. Moreover, the true bottom of a current basin may not be found. A working definition of a local minimum is employed instead in terms of the parameter k. Finding true local minima in the energy surface can be computationally intensive while unnecessary. For instance, analysis of the AMW surface in related work in [22] shows that the native structure is near but not at a minimum. In addition, the results in Section 3 make the case that a working definition of a local minimum is sufficient to discover near-native conformations.

2.2. BH for Sampling Decoy Configurations for Rigid Protein-Protein Docking. In this application, the native structures of two protein polypeptide chains (referred to as monomers) are known atomic coordinates obtained for each of the chains from experiment or structure prediction protocols. The objective is to find the native quaternary structure that brings the two monomers together. The assumption here is that the monomers do not change structure upon docking but bind rigidly with each other. Under this assumption, the objective is to find the spatial arrangement that brings one monomeric structure over the other and results in a dimeric configuration of lowest energy.

2.2.1. Employed Representation. In rigid docking, the only variables of interest are those that allow representing a spatial arrangement of one monomeric structure over another. A natural way to do so is through rigid-body transformations, which can be represented as vectors of 6 variables (3 for translation and 3 for rotation in 3-dimensional space). Hence, the variable space here is the 6-dimensional SE(3) space consisting of rigid-body motions or transformations.

The variable space we consider here is not the entire SE(3) but is constrained to rigid-body motions that align geometrically complementary and evolutionary conserved regions of the molecular spaces associated with each of the monomers. This builds upon earlier work by us on rigid docking which makes use of geometric hashing [51, 62]. Geometric hashing is a popular technique that essentially discretizes the space of rigid-body transformations by defining these transformations as alignments of geometrically complementary regions on monomeric molecular surfaces [63–66]. In recent work [51, 62] we show that the number of regions relevant for alignment can be further reduced by focusing on regions with high evolutionary conservation. Such regions are often found to be on contact interfaces [67].

While details of the process through which rigid-body transformations are defined are available in previous work [51, 62], we provide here a brief summary. The Connolly representation is first obtained for each monomeric surface [68]. The representation stores geometrical information for points on the surface, including whether the point represents a convex, saddle, or concave region. The representation is made less dense by only storing key locations on the molecular surface, known as critical points [69]. Triangles can be defined over these points. Associating evolutionary information with a critical point (through an analysis of related biological sequences [67]) allows focusing on triangles with high sequence conservation. We refer to these as active triangles. Once two geometrically complementary (e.g., concave with convex) active triangles T_A and T_B are obtained (from the molecular surfaces of monomers A and B, resp.), a rigid body transformation is easily defined as the one that aligns the local coordinate frame associated with T_B over that associated with T_A.

2.2.2. Energy Function. Each rigid-body motion can be represented as a transformation, a vector of 3 translation and 3 rotation components (details below) that when applied to the moving monomer (one monomer is designated as moving and the other as reference or base) move that monomer in space and bring it over the reference monomer. Atomic coordinates are then obtained for the resulting dimeric configuration, which can now be evaluated in terms

of the interaction energy. The energy function we employ combines three nonlocal terms useful for contact interfaces: $E = E_{\text{VdW}} + E_{\text{electrostatic}} + E_{\text{hydrogen-bonding}}$. The first term implements the standard 12–6 Lennard-Jones potential as in the CHARMM force field [70]. The electrostatic term implements Coulomb's law, also as in the CHARMM force field [70]. The hydrogen-bonding term is calculated as in [71] through the 12–10 hydrogen potential: $E_{\text{hydrogen-bonding}} = 5 \times [(r_0/d_{ij})^{12} - 6 \times (r_0/d_{ij})^{10}]$, where d_{ij} is the distance between acceptor and donor atoms i and j, and $r_0 = 2.9$ Å is the optimal distance for hydrogen bonding. Energy is computed only for the contact interface, which is defined over pairs of atoms in one monomer in contact with the atoms in the other monomer. Two atoms are in contact if their Euclidean distance is not higher than 4.0 Å.

2.2.3. Implementation of Structural Perturbation. The exposition above describes that a rigid-body motion is obtained by aligning an active triangle T_B on the surface of monomer B with a geometrically complementary active triangle T_A on the surface of the base monomer A. Let the current minimum C_i be the configuration corresponding to the transformation aligning T_B with T_A. In other words, the contact interface in C_i is that obtained by aligning T_B with T_A. As described in Section 1, an effective structural perturbation needs to preserve the adjacency relationship. For this reason, an effective perturbation in this context needs to modify the contact interface in C_i but limit the magnitude of the perturbation. The implementation we pursue here seeks a new pair of triangles, T'_A and T'_B, to perturb C_i and obtain $C_{\text{perturb},i}$. In order to limit the magnitude of the perturbation and preserve some of the contact interface of C_i in $C_{\text{perturb},i}$, T'_A needs to be close to T_A, and T'_B needs to be close to T_B.

This is implemented as follows. The molecular surface of each monomer is precomputed and represented in terms of a finite list of active triangles. The center of mass of each triangle is computed, and reverse indexing is used in order to sample a triangle T'_A and a triangle T'_B whose center of mass is within d Å of the center of mass of triangles T_A and T_B, respectively. The process repeats until a pair T'_A and T'_B are found which are geometrically complementary. A new rigid-body transformation aligning T'_B with T'_A is then defined, resulting in the perturbed configuration $C_{\text{perturb},i}$. Sampling in a d-radius neighborhood allows controlling and limiting the extent to which $C_{\text{perturb},i}$ perturbs the structural features of C_i (in this context, the contact interface).

2.2.4. Implementation of Local Optimization. As in the realization of the BH framework for protein structure prediction, the local optimization here also attempts at most m structural modifications starting with $C_{\text{perturb},i}$. The optimization terminates early if k consecutive modifications fail to lower energy. A naive implementation of the local optimization could employ the same structural modifications as the perturbation component; that is, new pairs of geometrically complementary active triangles are sought, but using a smaller d value. Our recent work on docking shows, however, that it becomes difficult to find geometrically complementary active triangles with smaller values of d [72]. A more effective alternative is to sample new rigid-body transformations directly rather than through new pairs of geometrically complementary active triangles and to do so in a continuous small neighborhood of an initial transformation.

Let the vector $\langle t, u, \theta \rangle$ be a rigid-body transformation, where t refers to the translation component, and $\langle u, \theta \rangle$ is an axis-angle representation of the orientation component (implemented through quaternions). In each move in the local optimization, a new random transformation is sampled in a small neighborhood of the transformation representing the configuration resulting from the previous modification. A new translation component t' is sampled in a δ_t neighborhood of t. A new axis u' is sampled by rotating around u by a sampled angle value δ_ϕ; a new angle θ' for the rotation component is obtained by sampling in a δ_θ neighborhood around θ. The result is that each move is a small modification of the contact interface to project a configuration onto its nearest local minimum. We note that, as before in the context of structure prediction, a working definition is employed here for the local minimum.

The result is a trajectory of low-energy dimeric configurations that are useful as decoys for the purpose of docking protocols. As in protein structure prediction, protein-protein docking protocols rely on first obtaining low-energy decoy configurations. Structural and energetic analysis then allows selecting a subset for further refinement in order to make a prediction on the native quaternary structure.

2.3. BH for Mapping Minima Connecting Diverse Stable States of a Protein Molecule. Many proteins employ motions to access different structures that allow them to tune their biological function [73, 74]. An important problem is to understand how a protein transitions between different functionally relevant states [50, 75, 76]. The problem of obtaining transition trajectories is directly related to that of obtaining the connectivity of the space around stable states. Computing transition trajectories is challenging [77], as such trajectories can connect structural states far away in the variable space. By taking into account a system's dynamics, the typical MD framework is in principle desirable to provide information on the time scales associated with conformational changes in a transition trajectory. However, its practical application is limited. Long simulation times may be needed to observe a transition trajectory to go over energy barriers.

In this proof-of-concept application, we propose that the BH framework can be a valuable tool as a first step towards elucidating transition trajectories. BH can be employed to map the minima connecting two given structural states and thus elucidate energetically credible conformational paths. Treating conformations in the path as important milestones, MD-based techniques can then be employed to locally deform a conformational path into an actual transition trajectory that incorporates dynamics [78].

In this application, the representation, energy function, and the implementations of the perturbation and local search

components of BH are as in Section 2.1. Here we pursue a proof-of-principle demonstration as follows. Let us suppose we are given two stable structural states of a protein, A and B. One of them can be regarded as the initial conformation to initiate a BH trajectory of local minima, and the other as the goal. A given number, let us say h, of BH trajectories can be initiated from the initial structure. The trajectories are allowed to grow for a fixed number of energy evaluations. In the unbiased scenario, the trajectories do not employ information about the location of the goal conformation in variable space. The results in Section 3 show that with sufficient sampling, if the initial and goal conformations are low-energy (i.e., stable), even unbiased BH trajectories are successful at approaching the goal conformation.

In a second scenario, the trajectories can be biased. Let us define an ϵ-radius ball around the goal conformation. As long as a BH trajectory stays outside this volume of the variable space (i.e., no minima are ϵ or closer to the goal), the BH exploration proceeds unbiased. When the trajectory enters the designated goal region of the variable space, say through its current minimum C_i, the exploration is biased towards obtaining minima that stay within the goal region. Given C_i, multiple perturbations followed by local optimization are attempted until a C_{i+1} is found which remains in the goal region. While the number of attempts is limited to a maximum of l consecutive failures before the BH exploration returns to its unbiased setting, in practice it is possible to remain in a goal region for a sufficiently large ϵ. The exploration terminates when the goal conformation is approached within some determined tolerance.

The value of ϵ is related to that of l. Moreover, a meaningful value for ϵ depends on the distance metric used and its effectiveness on a particular system. For instance, on small proteins, lRMSD can be used to determine the radius of the goal region. On other systems, instead, other measurements allow circumventing some of the issues with lRMSD. For instance, the TM-score [79] and GDT_TS [80] allow better capturing structural similarities than lRMSD when motions are localized to specific regions. The two are also less sensitive to noise. Familiarity with the system to be modeled allows better determining which measurement should be used and what values will be effective for ϵ and l. As the goal in this paper is to show a proof-of-concept demonstration that BH can be useful to obtain information on minima connecting diverse stable states in the equilibrium conformational space of a protein, we do not devote time to fine tuning parameters. The results in Section 3 show that values exist for these parameters that allow BH to come in closer proximity to the goal structural state in the biased over the unbiased implementation. Further tuning of the parameters is expected to improve the results and provide interesting directions for researchers to explore the viability of BH in this application context.

3. Results

Experimental Setup. The stopping criterion in each experimental setting to evaluate the performance of BH is set to a fixed number of energy evaluations. This number is 10^7 energy evaluations for the application of BH on structure prediction and 10^6 on our last application of connecting between different stable states. Results for the protein-protein docking do not change after 10,000 conformations, so this number is employed as a stopping criterion. Additionally, m and k are set to 100 and 20, respectively. In the application on docking, different values are tried for δt, $\delta\phi$, $\delta\theta$, and the ones employed for the experiments presented below are 1.5 Å, 10°, and 30°, respectively. On the last application of BH, h is set to 10 trajectories, ϵ is set to a TM-score of 0.4, $l = 100$, and the exploration terminates earlier than 10^6 energy evaluations when the current minimum is within a TM-score of 0.9 of the goal conformation.

We present three main sets of results according to the three different BH applications analyzed here. Where possible, results are compared to those reported by other state-of-the-art structure prediction or protein-protein docking methods (Tables 1 and 2 in the results presented in Sections 3.1 and 3.2, resp.). In addition, analysis of the BH-obtained minima is conducted, and distributions of the distances between consecutive minima are shown. This allows evaluating whether the implementations for the perturbation and local optimization in each application setting preserve the adjacency relationship between consecutively obtained minima. The comparison with state-of-the-art methods and the adjacency analysis employ the least Root-Mean-Squared Deviation (lRMSD) semimetric. Analysis of results obtained on the third application of BH on connecting stable states of a protein molecule employs additional measurements, such as GDT_TS and TM-score (Tables 3 and 4 in the results presented in Section 3.3). The minima sampled by BH in the context of this third application are also visualized on a low-dimensional projection of the variable space (the projection coordinates are detailed below) that reveals where the BH sampling focuses.

The main measurement used in the analysis below is lRMSD. Briefly, lRMSD measures the weighted Euclidean distance between corresponding atoms after optimal superposition of the two conformations under comparison (or configurations, if consisting of more than one polypeptide chain). The optimal superposition refers to the rigid-body motion or transformation in SE(3) minimizing this weighted Euclidean distance [81]. lRMSD captures structural dissimilarity, but it is not a Euclidean metric, as it does not obey the triangle inequality. Low values indicate high similarity, and high values indicate high dissimilarity, but interpretation of intermediate values is difficult. Interpretation has been the subject of many studies [82]. For instance, lRMSD has been found to depend on system size. A 5 Å lRMSD between a computed conformation and the native structure of a short protein chain of no more than 30 amino acids is considered a large deviation, but the same dissimilarity is less significant for a medium-size protein of 100 amino acids or more. Working interpretations abound. In general, for medium-size proteins, if the lowest lRMSD obtained over computed conformations to the known native structure is more than 6-7 Å, the native structure is not considered to have been captured in silico.

Table 1: Comparison of the lowest lRMSDs obtained by BH to those obtained by other methods on the protein dimers studied here. MBH refers to monotonic BH. lRMSDs reported by BH, MBH, and the work in [56] in columns 5–7 are over backbone atoms, whereas those reported by the work in [36, 84] in columns 8-9 are over alpha carbons of the backbone chain.

Number	PDB ID	Length	Fold	BH (Å)	MBH (Å)	[56] (Å)	[36] (Å)	[84] (Å)
1	1dtdB	61	α/β	6.9	6.6	7.5	6.5	5.7
2	1isuA	62	α/β	6.3	6.5	6.5	6.5	6.9
3	1c8cA	64	α/β	6.5	5.7	7.2	3.7	5.0
4	1sap	66	α/β	6.5	6.0	7.36	4.6	6.6
5	1hz6A	67	α/β	5.7	6.0	6.6	3.8	3.4
6	1wapA	68	β	7.4	8.1	7.3	8.0	7.7
7	1fwp	69	α/β	6.3	6.7	7.1	8.1	7.3
8	1ail	70	α	3.2	4.2	4.0	5.4	6.0
9	1aoy	78	α/β	5.7	6.1	5.8	5.7	5.7
10	1cc5	83	α	5.8	5.6	5.8	6.5	6.2
11	2ezk	93	α	4.3	5.8	6.0	5.5	6.6
12	1hhp	99	β	10.4	10.5	11.0	NA	NA
13	2hg6	106	α/β	8.8	9.3	9.7	NA	NA
14	3gwl	106	α	4.9	4.9	6.3	NA	NA
15	2h5nD	123	α	7.5	7.8	8.6	NA	NA

Table 2: Comparison of the lowest lRMSDs obtained by BH to those obtained by other methods. Systems that are CAPRI targets are denoted by an asterisk.

Number	PDB ID (chains)	Size	BH (Å)	[66] (Å)	[91] (Å)
1	1c1y (A,B)	1376, 658	1.8	1.2	N/A
2	1ds6 (A,B)	1413, 1426	3.4	1.2	N/A
3	1tx4 (A,B)	1579, 1378	2.4	1.4	N/A
4	1www (W,Y)	862, 782	2.6	11.4	N/A
5	1flt (V,Y)	770, 758	2.7	1.5	N/A
6	1vcb (A,B)	755, 692	3.4	0.8	N/A
7	1vcb (B,C)	692, 1154	2.7	13.1	N/A
8	1ohz* (A,B)	1027, 416	2.7	1.7	0.6
9	1t6g* (A,C)	2628, 1394	3.6	1.7	3.8
10	1zhi* (A,B)	1597, 1036	4.6	25.3	3.4
11	2hqs* (A,C)	3127, 856	2.6	29.1	2.5
12	1qav (A,B)	663, 840	2.6	1.4	N/A
13	1g4y (B,R)	682, 1156	4.1	0.8	N/A
14	1cse (E,I)	1920, 522	2.4	0.7	N/A
15	1g4u (R,S)	1398, 2790	3.2	1.0	N/A

However, high values of lRMSD cannot be automatically interpreted to indicate significant structural dissimilarity. Since lRMSD weighs each atom equally, it cannot capture global topology changes and overly penalizes cases where the differences are localized to a specific region of the molecule due to, say, a large-scale motion. For instance, the lRMSD of two conformations can be high even if structural deviations are limited to a loop that has a different orientation in the two conformations under comparison [83]. In such cases, other measurements, such as GDT_TS and TM-score, can be more appropriate. While different in implementation details, these two scores essentially locate a maximum subset of atoms between two conformations under comparison which are close in space after optimal superposition and minimizes an overall lRMSD-based error. While GDT_TS is reported in %, TM-score is unitless. Both capture similarity, so higher values are better. While lRMSD and GDT_TS depend on system size, TM-scores are found to be more reliable [83], which is why we employ them here in the analysis in Section 3.3 on the third BH application on connecting diverse stable states.

3.1. Analysis of BH-Obtained Decoy Conformations of a Protein Polypeptide Chain. Our realization of the BH framework for the purpose of ab-initio structure prediction is applied to a comprehensive list of 15 target protein systems. These systems, listed in Table 1, range from 61–123 amino acids in length and cover the α, β, and α/β folds. Many of them are selected due to the availability of data reported on them by structure prediction protocols. On these systems, computing 10^7 energy function evaluations takes 1–4 days of CPU time on a 2.4 Ghz Core i7 processor, depending on chain length.

3.1.1. Comparison with State-of-the-Art Methods. Table 1 shows comparisons to state-of-the-art methods in ab-initio structure prediction in terms of lRMSD. Over all minima obtained from each amino acid sequence by BH, the conformation with the lowest lRMSD to the known native structure of that sequence (experimentally obtained structure with PDB ID shown in this table) is recorded, and that value is reported in column 5 in Table 1. To take into account stochasticity, we report in Table 1 the average lowest lRMSD obtained over 3 independent runs. This value is compared to lowest lRMSDS reported by methods that are popular in ab-initio structure prediction [36, 84]. We also compare to data obtained with our previous work on ab-initio structure prediction with a robotics-inspired tree-based exploration

Table 3: Initial conformations of calmodulin are in the rows, whereas goal conformations are denoted in the columns. Three measurements, lowest lRMSD, highest TM-score, and highest GDT_TS scores to any of the goal conformations, are reported over the 10 trajectories started from each initial conformation.

PDB	1cfd			1cll			2f3y		
ID	lRMSD (Å)	TM-score	GDT_TS (%)	lRMSD (Å)	TM-score	GDT_TS (%)	lRMSD (Å)	TM-score	GDT_TS (%)
1cfd	6.70	0.57	56	6.48	0.60	50	8.20	0.44	38
1cll	5.84	0.50	43	2.38	0.84	78	6.24	0.53	54
2f3y	6.7	0.47	40	2.50	0.82	74	2.87	0.77	73

Table 4: Initial conformations of adenylate kinase are in the rows, whereas goal conformations are denoted in the columns. Three measurements, lowest lRMSD, highest TM-score, and highest GDT_TS scores to any of the goal conformations, are reported over the 10 trajectories started from each initial conformation.

PDB	1dvr		2aky		2ak3		4ake	
ID	TM-score	GDT_TS (%)	TM-score	GDT_TS (%)	TM-score	GDT_TS (%)	TM-score	GDT_TS (%)
1dvr	0.99	99	0.83	74	0.41	25	0.32	20
2aky	0.84	76	1.00	100	0.44	28	0.31	18
2ak3	0.41	25	0.44	28	0.99	99	0.39	25
4ake	0.31	18	0.29	17	0.41	26	1.00	100

method [40, 41]. Monotonic BH (MBH) is also included in the comparisons (column 6).

The results in Table 1 make the case that BH performs just as well as state-of-the-art methods in structure prediction in terms of its ability to obtain low-lRMSD conformations in an ab-initio setting. The role of the energy function may partially explain some differences among the methods, as they employ different energy functions (MBH and [41] also employ AMW). It is interesting to note that, while our realization of BH (which uses a Metropolis criterion to add the next minimum to its trajectory) obtains lower lowest lRMSDs on more proteins than MBH, the performance of MBH is comparable to the other methods in many cases. MBH can be regarded as a special case of the BH framework with the Metropolis criterion, where temperature $T = 0$. The T value we use here for our realization of BH allows a 2.6 kcal/mol energy increase between two consecutive local minima with probability 0.1.

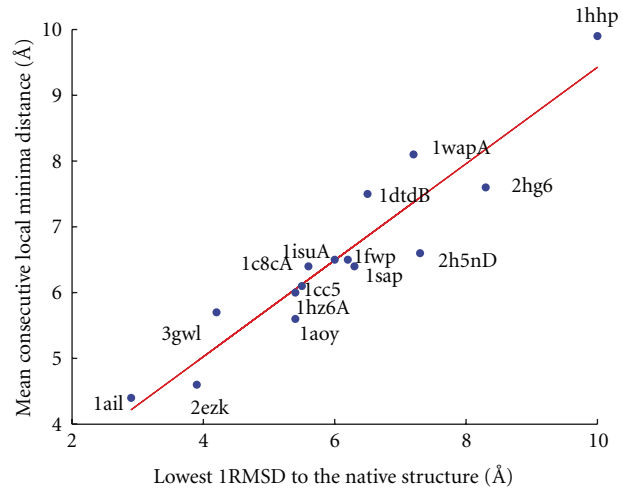

Figure 2: The mean consecutive local minima distance is drawn against the lowest lRMSD obtained for each protein.

3.1.2. Evaluation of Adjacency Relationship. Adjacency between local minima obtained consecutively by BH is often stated as important for global optimization. Here we show concretely, in the context of ab-initio structure prediction, how this adjacency correlates with the lowest lRMSD reported by BH to the known native structure of each of the protein systems studied. The lRMSD between two consecutive local minima is computed, and the average is recorded for a given protein system. This value is plotted against the lowest lRMSD from the native structure obtained by BH on each protein system in Figure 2. A strong correlation of 94% is observed in Figure 2 between the average consecutive local minima distance and the lowest lRMSD to the native structure. This result suggests that adjacency of consecutively sampled local minima is related to the ability of BH to locate the native structure. Figure 2 shows that, in cases where the average consecutive local minima distance is large, BH does not come close to the native structure.

Further detail is provided in Figure 3 on two protein systems. These systems are selected to represent two diametrical cases that correspond to the bottom left and top right portions in Figure 2. The lowest lRMSD structures obtained for each of these two systems by BH are superimposed over their respective native structures in Figures 3(a)-3(b). The entire distribution of consecutive local minima distances is shown for these two proteins in Figures 3(c)-3(d). Figures 3(c)-3(d) further show that, in cases where the majority of consecutive minima are not adjacent in variable space, the overall performance of BH in terms of lowest lRMSD to the native structure suffers. One reason for the poor adjacency is that the fragment replacement may perturb too much

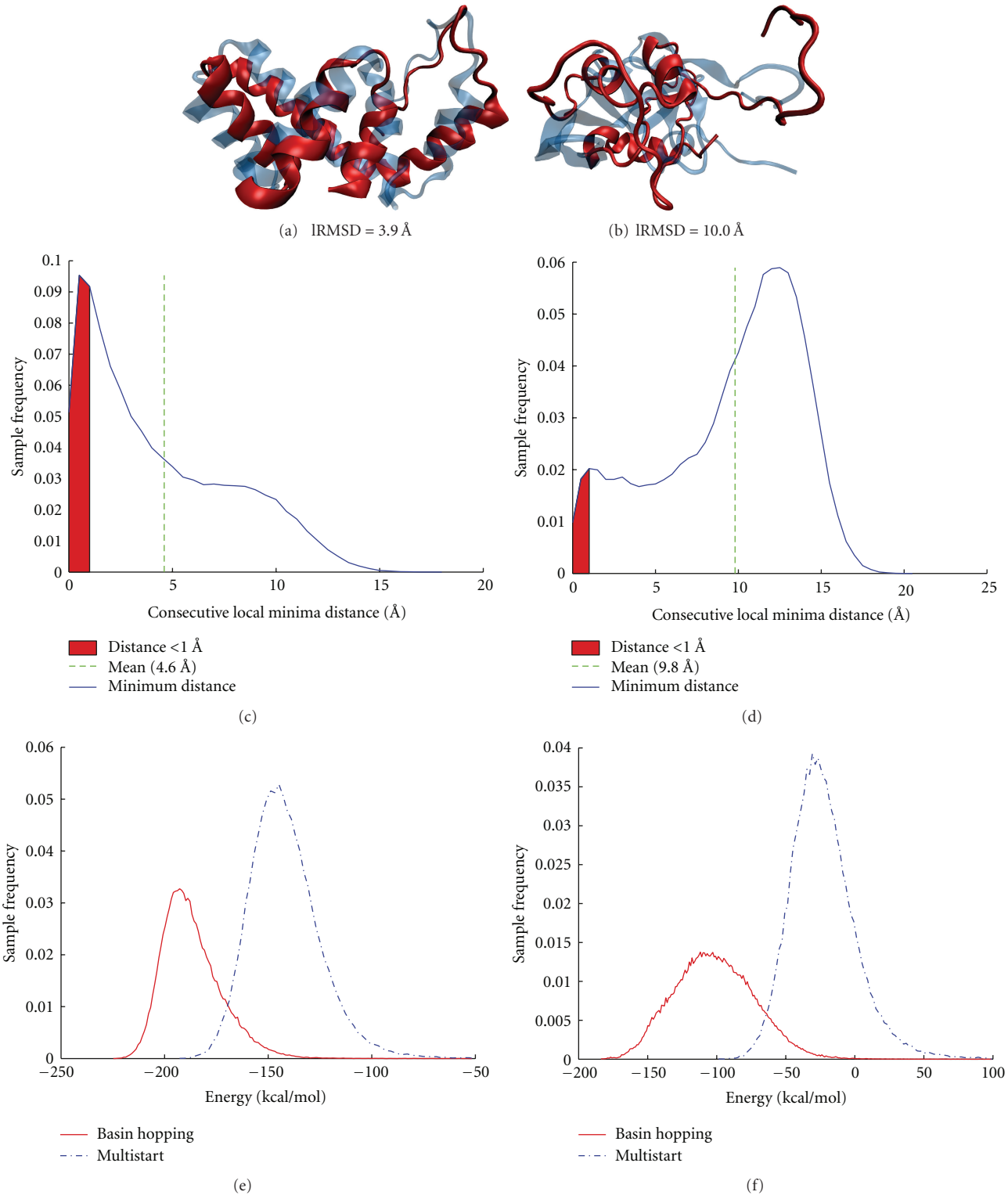

Figure 3: (a and b) The lowest-lRMSD conformation (in opaque red) is superimposed over the known native structure (in transparent blue) for the protein with native PDB ID 2ezk in (a) and 1hhp in (b). (c and d) The distribution of consecutive local minima distances in terms of lRMSD is shown in (c and d) for the two proteins, respectively. (e and f) The distribution of energies obtained by BH is superimposed over that obtained by the multistart method on each of the two proteins.

of a conformation. In a recent analysis [85], we show that this can be controlled by biasing the sampling of fragment replacements towards those that will result in small structural changes in terms of lRMSD between C_i and $C_{\text{perturb},i}$.

Comparison of BH with Multistart Sampling. Adjacency of consecutive local minima in BH is often stated as a key distinguishing characteristic over a multistart method, where initial points for local optimization are essentially sampled at random over the variable space. Here we show the effect of the adjacency relationship in a concrete setting in terms of the energetic quality of the sampled minima. On the same two protein systems where the above analysis highlights consecutive local minima distances, we show in Figures 3(e)-3(f) the distribution of energies. Figures 3(e)-3(f) superimposes the distribution obtained by BH over that obtained by the multistart method. The results show that BH obtains lower-energy minima than the multistart method. In the context of ab-initio structure prediction, the quality of decoy conformations obtained by BH is superior over that obtained by a multistart method.

Sampling Redundancy in BH. It is interesting to determine how often our realization of the BH framework here comes close to the native structure. We show this visually through a projection of the variable space in a few dimensions. The projection coordinates we choose are based on the ultrafast shape recognition (USR) features [86], which we have employed in previous work to guide a tree-based exploration of the variable space with measurements taken over a low-dimensional projection [40, 56]. These coordinates give a coarse representation of the molecular shape. They are first momenta of distance distributions of atoms in a molecule from selected points on the molecule. The selected (reference) points are the centroid (ctd), point closest to centroid (cst), point farthest from centroid (cfd), point farthest from cfd, and so on. More reference points can be defined this way, but the ones we employ for the visual representation here use only ctd and cfd. It is worth noting that, while coarse, the USR-based projection is fast to compute for each conformation, unlike PCA- or ISOMAP-based decompositions [87–89], which are time consuming and hard to use in an online setting and contain several other shortcomings noted for conformational space [90].

Figure 4 shows the projection of BH-sampled minima over the two USR-based coordinates measured using the ctd and cfd reference points. The projection is discretized so that cells can be defined in a 2d grid for the purpose of measuring how often BH samples similar minima in terms of their coarse 2d USR-based representation. The 2d grid in Figure 4 is color coded with a blue-to-red color scheme that corresponds to cells with low-to-high number of minima projected to them. The cell that contains the projection of the native structure is marked with an $\times$.

The projection of the sampled minima in this USR-based 2-dimensional space allows visualizing highly sampled regions by BH. The representation is coarse (e.g., cell widths used here for visualization can be made smaller), as conformations that map to the same cell may be several Å apart, but the projection is useful to draw two conclusions. First, compared to the vast variable space (sea of blue in Figure 4), BH sampling seems to focus in regions near the native structure. These regions represent the equilibrium conformational space. Second, sampling can be redundant; some regions are more populated than others. Future research can address redundancy in order to enhance the capability of BH to sample the equilibrium conformational space of a protein molecular in terms of local minima.

3.2. Analysis of BH-Obtained Decoy Configurations of Protein-Protein Dimers. Our realization of the BH framework for the purpose of protein-protein docking is applied to a comprehensive list of 15 different dimers. These vary in size, represent diverse functional classes, and have been tested by other protein-protein docking methods, and some are even CAPRI targets. Testing is carried out on a 2.66 GHz Opteron processor with 8 GB of memory. Depending on system size, obtaining 10,000 conformations takes 6–12 CPU hours.

3.2.1. Comparison with State-of-the-Art Methods. Table 2 shows the lowest lRMSD from the known native structure (with PDB ID shown in column 1) obtained by BH in column 3. Lowest lRMSDs reported on these systems by other methods are shown in columns 4-5. System size in terms of number of atoms in each of the chains is shown in column 2. Table 2 shows that BH achieves low lRMSDs to the native structure on each system. Moreover, these are comparable to the lRMSDs reported by other related methods. In particular, the method presented in [66] employs geometric hashing, whereas that in [91] uses long optimizations with a carefully designed energy function that employs information on evolutionary conservation to sample low-energy conformations. In addition to a comparable performance with these methods, BH samples many configurations within 5 Å lRMSD of the native structure (data not shown). These configurations, if selected and further refined in the course of a multistage docking protocol, will allow obtaining the native structure in great detail.

3.2.2. Evaluation of Adjacency Relationship. We investigate here the adjacency between consecutively sampled local minima. Figure 5(a) plots the mean consecutive local minima distance in terms of lRMSD for each protein against the lowest lRMSD obtained to the native structure. A positive correlation of 73% is observed. The mean consecutive local minima distance is less than 15 Å for about half of the systems. While this may seem like a large number compared to the related results on ab-initio structure prediction, the range is larger due to the size of the dimeric systems (lRMSD depends on size). The strong correlation suggests that adjacency of consecutively sampled minima directly relates with the ability of BH to locate a global minimum. Lower lowest lRMSDs (<5 Å) are obtained here compared to the ab-initio structure prediction setting. This is not surprising, as the variable space here is 6-dimensional, whereas the space

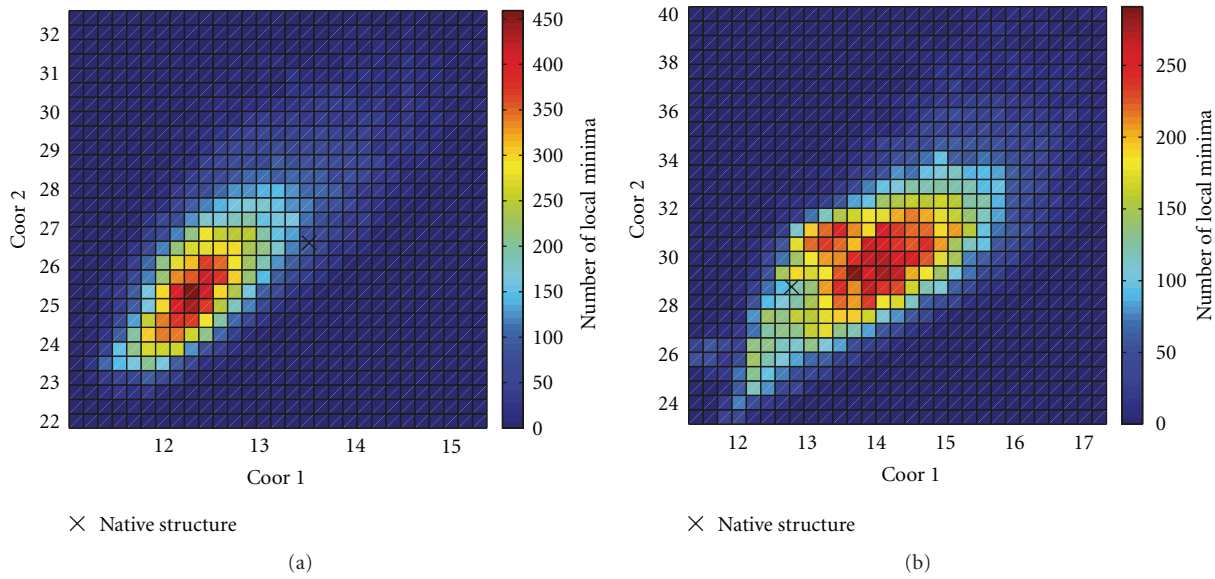

Figure 4: The 2d grids show projections of BH-sampled minima using two USR-based coordinates (with ctd and cfd as reference points). The projection for the protein with native PDB ID 2ezk is shown in (a), and that for the protein with native PDB ID 1hhp is shown in (b). The grids are color coded with a blue-to-red color scheme to show cells with low-to-high number of minima projected to them. The cell that contains the projection of the native structure is marked with an ×. The range of values of each of the coordinates is estimated as in [40]. Maximum values are based on an extended chain, and minimum values are based on the Flory compact self-excluding model of a chain of n amino acids. To improve visualization, the ranges are limited here, and the grids are clipped to allow focusing to regions with some minimal population.

in the ab-initio structure prediction application contains hundreds of dimensions.

Further detail is provided in Figures 5(d)-5(e), which shows the distribution of lRMSDs between consecutively sampled local minima on two protein systems. These systems are selected to represent two diametrical cases that correspond to the bottom left and top right portions in Figure 5(a). The actual lowest-lRMSD structures obtained on these systems are shown in Figures 5(b)-5(c), superimposed over the corresponding native structures of these proteins. The distributions in Figures 5(d)-5(e) show that more pairs of consecutive minima with low lRMSDs are obtained for the protein where BH also obtains a lower lowest lRMSD to the native structure.

3.3. Analysis of BH Trajectories in Connecting Diverse Stable States of a Protein. The unbiased setting of BH is tested here in detail on two proteins, calmodulin and adenylate kinase. Some encouraging results are shown for the biased setting as well, but a detailed investigation of the biased implementation and parameter tuning is beyond the purpose of this work.

3.3.1. Mapping Minima between Stable States in Calmodulin. Calmodulin is a 144 amino-acid long EF-hand protein that binds calcium and regulates more than 100 proteins, including kinases, phosphodiesterases, calcium pumps, and motility proteins [46–48]. The protein resembles a dumbbell, with the terminal domains linked by a flexible α-helix and the termini in a transorientation from each other on either side of the central linker. The partial unfolding of the central linker around position 77 gives calmodulin flexibility.

Calmodulin has been captured in three different functionally relevant structural states in the wet laboratory [92–94]. These states are documented in the PDB as X-ray structures under PDB IDs 1cfd (apo state), 1cll (calcium-binding state), and 2f3y (collapsed peptide-binding state). The central helix is fully formed in the calcium-binding state, unfolds in the middle in the apo state, and bends in the collapsed state. Transitions between the apo and collapsed states have been observed both in experiment and simulation [49, 50].

In order to test the unbiased setting of BH in this application, the following experiment is conducted. Each of the structures is obtained from the PDB and employed as an initial conformation. $h = 10$ Bh trajectories are launched independently from any of them. The proximity to any of the other two structures is reported in Table 3. Entry at row i and column j reports the best proximity over the 10 trajectories initiated from structure i to goal structure j.

Proximity to the goal is measured in three ways. Table 3 shows lowest lRMSD, highest TM-score [79], and highest GDT_TS [80]. The results in Table 3 show that BH is able to capture the structure with PDB ID 1cfd (apo) when initiated from 1cll (semicollapsed state) and vice versa (TM-scores above 0.5 are found to capture significant structural similarity [83]). BH also captures the structure with PDB ID 1cll (semicollapsed state) when initiated from 2f3y (collapsed state) and vice versa. A very low lRMSD and very high TM-score and GDT_TS score are obtained from 2f3y to 1cll.

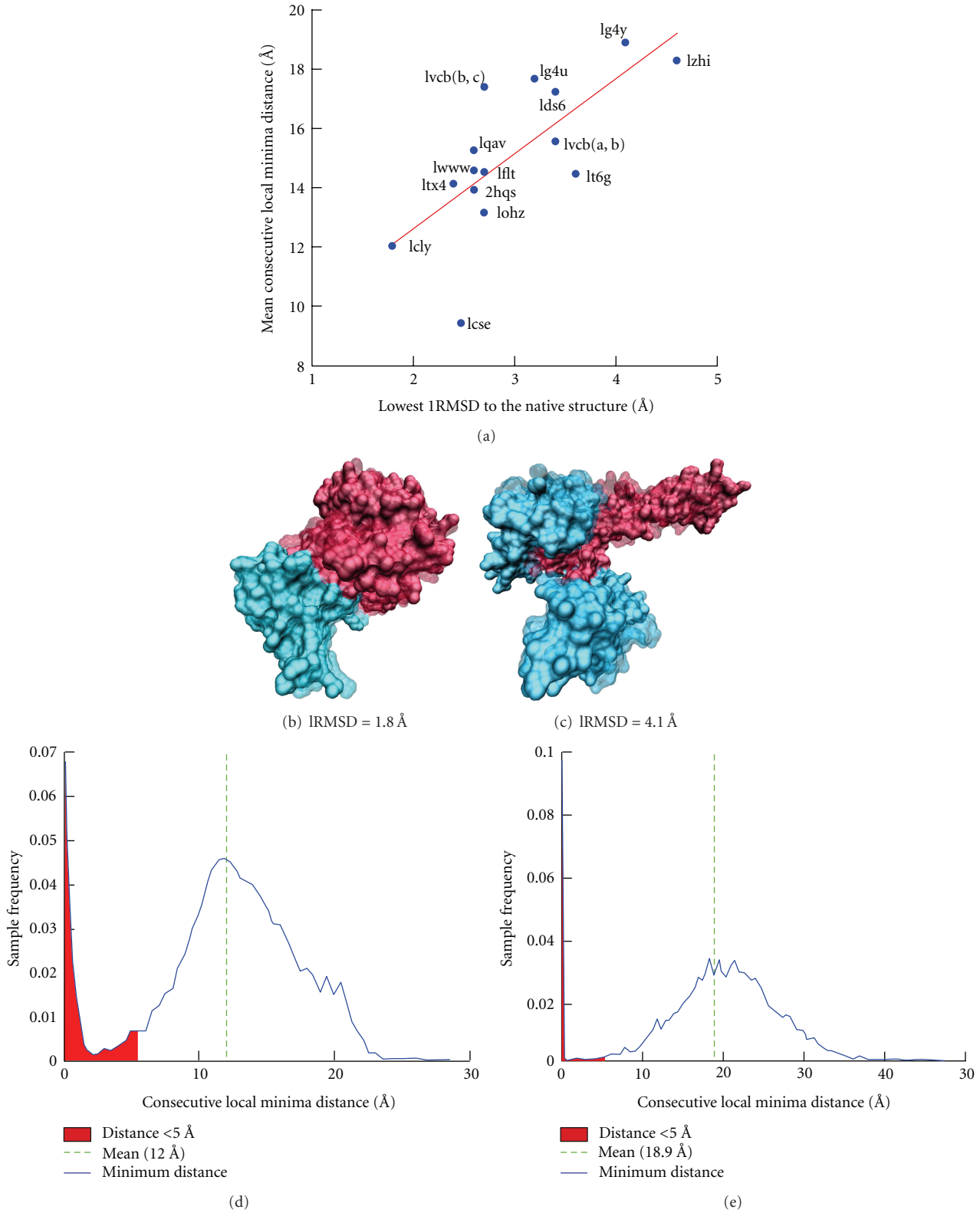

Figure 5: (a) The mean consecutive local minima distance is drawn against the lowest lRMSD obtained for each protein. (b and c) The lowest-lRMSD dimeric configuration (in opaque, with chains in different colors) is superimposed over the known native structure (in transparent) for the protein with native PDB ID 1c1y in (b) and 1g4y in (c). (d and e) The distribution of consecutive local minima distances in terms of lRMSD is shown in (d and e) for each of the proteins, respectively.

This is an encouraging result, since 1cll captures a partially closed state, whereas 2f3y captures calmodulin in its close state. Structurally, 1cfd, the apo state, is further away from the semicollapsed and collapsed states. Indeed, all three measurements in Table 3 indicate that BH has not captured state 1cfd from 2f3y and vice versa.

Table 3 also shows values along the main diagonal, which record the closest that a BH trajectory comes to its initial conformation. While this is achieved often through the very first minimum, the structure with PDB ID 1cfd is the only exception. This indicates that this structure is not at a local minimum, and BH quickly steers away from this initial conformation. This may also explain the difficulty of capturing that state when initiated from any of the other two. In addition, analysis of the biased setting for calmodulin reveals that when an ϵ value of 0.4 in terms of TM-score is used, BH is able to come closer to the goal structures. Improvements are around 0.5 Å (data not shown).

3.3.2. Mapping Minima between Stable States in Adenylate Kinase. Adenylate kinase is a 214 amino-acid long phosphotransferase enzyme that maintains energy balance in cells by catalyzing the reversible reaction $Mg^{2+} \cdot ATP + AMP \rightleftharpoons Mg^{2+} \cdot ADP + ADP$ [95]. The protein consists of a CORE domain and two substrate-binding (AMP- and ATP-) domains. The binding domains move and bind substrates independently, resulting in different functional states.

Adenylate kinase has been found in four different structural states in the wet laboratory [96–99]: the apo state, where both substrate-binding domains are open (available in the PDB under PDB ID 4ake), the collapsed state, where both domains are closed (available under PDB ID 2aky), and two intermediate states, where one of the domains is open and the other closed (PDB IDs 1dvr and 2ak3). Transitions between the apo and collapsed states have been observed both in experiment and simulation [76, 100, 101].

Unbiased BH trajectories are initiated from each of the four structures, and the best proximity to any of the other three is reported in Table 4 in terms of TM-score and GDT_TS. lRMSD is not employed, as the chains deposited under the PDB IDs listed above are of different lengths (due to differences in the setup of the structure resolution protocol in the wet laboratory). The results in Table 4 show that adenylate kinase is indeed a challenging system. The BH trajectories manage to come close only to the structure with PDB ID 2aky when initiated from the structure with PDB ID 1dvr and vice versa. This is an encouraging result, nonetheless, because the structures with PDB ID 1dvr and 2aky are structurally closer to each other.

Outstanding Challenges. Calmodulin and adenylate kinase are considered challenging systems for computational investigation due to their large size [33]. The results above support the fact that size limits sampling capability in BH. The upper bound of 10^6 energy evaluations limits BH to sample around 1,500 minima for calmodulin and 1,000 minima for adenylate kinase. Considering the small number of minima sampled, the results above are encouraging. They suggest that, with improvements to enhance the sampling capability, BH is a promising tool for mapping the equilibrium conformational space of a protein and elucidating the connectivity between different stable states.

4. Conclusion

We have shown that BH is a general, versatile framework that allows structural characterization of important biological macromolecules, such as proteins. We have selected three different applications of importance in computational structural biology on which to show the power and promise of the BH framework. Domain-specific expertise is used to implement effective perturbation and local optimization components. Important generally recognized characteristics of the BH framework, such as adjacency of local minima and its relation to the quality of the reported global minimum, are demonstrated in the applications selected in this paper.

Taken together, the results show that BH is an effective framework for structural characterization of protein systems. It is more effective than the multistart method, and the adjacency of consecutively sampled local minima is directly related with the ability of BH to come close to the global minimum. The presented results make the case that BH can be an effective tool for generating good-quality decoys for ab-initio structure prediction and protein-protein docking and a promising framework for mapping the connectivity of functionally relevant states in flexible proteins.

We note that the implementations we offer here for the key components in BH are a first step, and further tuning can result in better performance. Further analysis into different implementations is necessary to obtain a better understanding of the BH framework and its capability to enhance sampling of molecular spaces. We are currently pursuing such a comparative analysis. For instance, in recent work [85] we show that the implementation of the perturbation component employed here is sufficient to escape a current minimum and that the greedy search employed for the local optimization is just as effective but more efficient than Metropolis MC searches at low temperature [22, 102].

The application on proteins with diverse stable states serves as a proof of concept that BH can be employed to map the intermediate minima that connect stable states. The results presented here show that the framework is promising and merits further investigation in this context. The trajectory of minima obtained by BH in connecting two stable states can be considered a coarse conformational path. This path can be transformed into an actual trajectory that takes the protein through specific molecular motions from one state to another. The process is not dissimilar from how paths in robotics motion sampling are converted to actual execution trajectories with dynamics constraints [103]. The coarse transition paths can be refined through, for instance, short steered MD simulations connecting adjacent minima. Other path deformation techniques are available, and this is a direction we will explore in future research.

We believe the exposition of BH in this paper will bring more attention to this framework as a powerful global

optimization tool for biological systems. Its versatility, as we show here in the context of three different yet related applications on proteins, merits further investigation. In particular, different implementations for the main components in BH can be investigated to balance between accuracy and efficiency. Moreover, related ideas from the evolutionary computing community on population-based strategies can be employed to promote diversity of minima, as proposed in recent work on geometrical problems [20]. Related ideas from our robotics-inspired search of molecular conformational spaces [40] can be exploited to organize the BH-sampled minima, steer the exploration away from overpopulated regions in the variable space, and so enhance the sampling capability in BH.

Acknowledgments

This work is supported in part by NSF CCF no. 1016995 and NSF IIS CAREER Award no. 1144106.

References

[1] C. A. Floudas and P. M. Pardalos, *Encyclopedia of Optimization*, Kluwer Academic Publishers, Norwell, Mass, USA, 2001.

[2] "Nonconvex optimization and its applications," in *Global Optimization: Scientific and Engineering Case Studies*, J. Pinter, Ed., vol. 85 of *Mathematics and Statistics*, Springer Science and Business Media, New York, NY, USA, 2006.

[3] D. J. Wales and J. P. K. Doye, "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms," *Journal of Physical Chemistry A*, vol. 101, no. 28, pp. 5111–5116, 1997.

[4] A. Tiano, F. Pizzochero, and P. Venini, "A global optimization approach to nonlinear system identification," in *Conference on Control and Automation*, pp. 752–761, 1999.

[5] A. R. Leach, "A survey of methods for searching the conformational space of small and medium-sized molecules," in *Reviews in Computational Chemistry*, vol. 2, pp. 1–55, VCH Publishing, New York, NY, USA, 1991.

[6] H. A. Scheraga, "Predicting three-dimensional structures of oligopeptides," in *Reviews in Computational Chemistry*, K. B. Lipkowtiz and D. B. Boyd, Eds., vol. 3, pp. 73–142, VCH Publishing, New York, NY, USA, 1992.

[7] R. V. Pappu, R. K. Hart, and J. W. Ponder, "Analysis and application of potential energy smoothing and search methods for global optimization," *Journal of Physical Chemistry B*, vol. 102, no. 48, pp. 9725–9742, 1998.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[9] A. Nayeem, J. Vila, and H. A. Scheraga, "A comparative study of the simulated-annealing and monte carlowith-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin," *Journal of Computational Chemistry*, vol. 12, no. 5, pp. 594–605, 1991.

[10] R. C. Brower, G. Vasmatzis, M. Silverman, and C. Delisi, "Exhaustive conformational search and simulated annealing for models of lattice peptides," *Biopolymers*, vol. 33, no. 3, pp. 329–334, 1993.

[11] W. F. van Gunsteren, D. Bakowies, R. Baron et al., "Biomolecular modeling: goals, problems, perspectives," *Angewandte Chemie*, vol. 45, no. 25, pp. 4064–4092, 2006.

[12] A. Shehu, "Conformational search for the protein native state," in *Protein Structure Prediction: Method and Algorithms*, Rangwala and G. Karypis, Eds., Wiley Book Series on Bioinformatics, chapter 21, Fairfax, VA, USA, 2010.

[13] R. H. Leary, "Global optimization on funneling landscapes," *Journal of Global Optimization*, vol. 18, no. 4, pp. 367–383, 2000.

[14] M. Iwamatsu and Y. Okabe, "Basin hopping with occasional jumping," *Chemical Physics Letters*, vol. 399, no. 4–6, pp. 396–400, 2004.

[15] M. A. Miller and D. J. Wales, "Novel structural motifs in clusters of dipolar spheres: knots, links, and coils," *Journal of Physical Chemistry B*, vol. 109, no. 49, pp. 23109–23112, 2005.

[16] J. M. Carr and D. J. Wales, "Global optimization and folding pathways of selected alpha-helical proteins," *The Journal of Chemical Physics*, vol. 123, no. 23, p. 234901, 2005.

[17] T. James, D. J. Wales, and J. Hernández-Rojas, "Global minima for water clusters $(H_2O)n$, $n \leq 21$, described by a five-site empirical potential," *Chemical Physics Letters*, vol. 415, no. 4–6, pp. 302–307, 2005.

[18] R. Gehrke and K. Reuter, "Assessing the efficiency of first-principles basin-hopping sampling," *Physical Review B*, vol. 79, no. 8, Article ID 085412, 10 pages, 2009.

[19] D. J. Wales, *Energy Landscapes and Structure Prediction Using Basin-Hopping*, Wiley-VCH Verlag GmbH and Co. KGaA, 2010.

[20] A. Grosso, A. R. M. J. U. Jamali, M. Locatelli, and F. Schoen, "Solving the problem of packing equal and unequal circles in a circular container," *Journal of Global Optimization*, vol. 47, no. 1, pp. 63–81, 2010.

[21] M. Locatelli, "On the multilevel structure of global optimization problems," *Computational Optimization and Applications*, vol. 30, no. 1, pp. 5–22, 2005.

[22] B. Olson and A. Shehu, "Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface," *Proteome Science*, vol. 10, no. supplement 1, p. S5, 2012.

[23] O. M. H. R. Lourenco and T. Stutzle, "Iterated local search," in *Handbook of Metaheuristics*, F. Glover and G. Kochenberger, Eds., vol. 57, no. 513 of *Operations Research & Management Science*, pp. 321–353, Kluwer Academic Publishers, 2002.

[24] Z. Li and H. A. Scheraga, "Monte Carlo-minimization approach to the multiple-minima problem in protein folding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 19, pp. 6611–6615, 1987.

[25] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nature Structural Biology*, vol. 4, no. 1, pp. 10–19, 1997.

[26] J. N. Onuchic and P. G. Wolynes, "Theory of protein folding," *Current Opinion in Structural Biology*, vol. 14, no. 1, pp. 70–75, 2004.

[27] J. Moult, K. Fidelis, A. Kryshtafovych, and A. Tramontano, "Critical assessment of methods of protein structure prediction (CASP) round IX," *Proteins*, vol. 79, supplement 10, pp. 1–5, 2009.

[28] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[29] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nature Methods*, vol. 4, no. 6, pp. 466–467, 2007.

[30] T. Kortemme and D. Baker, "Computational design of protein-protein interactions," *Current Opinion in Chemical Biology*, vol. 8, no. 1, pp. 91–97, 2004.

[31] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins*, vol. 65, no. 3, pp. 712–725, 2006.

[32] A. Verma, A. Schug, K. H. Lee, and W. Wenzel, "Basin hopping simulations for all-atom protein folding," *The Journal of Chemical Physics*, vol. 124, no. 4, p. 044515, 2006.

[33] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins*, vol. 76, no. 4, pp. 837–851, 2009.

[34] R. Bonneau, C. E. M. Strauss, C. A. Rohl et al., "De novo prediction of three-dimensional structures for major protein families," *Journal of Molecular Biology*, vol. 322, no. 1, pp. 65–78, 2002.

[35] T. J. Brunette and O. Brock, "Guiding conformation space search with an all-atom energy potential," *Proteins*, vol. 73, no. 4, pp. 958–972, 2008.

[36] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, "Mimicking the folding pathway to improve homology-free protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 10, pp. 3734–3739, 2009.

[37] J. Debartolo, G. Hocky, M. Wilde, J. Xu, K. F. Freed, and T. R. Sosnick, "Protein structure prediction enhanced with evolutionary diversity: SPEED," *Protein Science*, vol. 19, no. 3, pp. 520–534, 2010.

[38] A. Shehu, L. E. Kavraki, and C. Clementi, "Unfolding the fold of cyclic cysteine-rich peptides," *Protein Science*, vol. 17, no. 3, pp. 482–493, 2008.

[39] M. C. Prentiss, C. Hardin, M. P. Eastwood, C. Zong, and P. G. Wolynes, "Protein structure prediction: the next generation," *Journal of Chemical Theory and Computation*, vol. 2, no. 3, pp. 705–716, 2006.

[40] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an Ab-initio tree-based exploration," *International Journal of Robotics Research*, vol. 29, no. 8, pp. 1106–1127, 2010.

[41] B. Olson, K. Molloy, and A. Shehu, "In search of the protein native state with a probabilistic sampling approach," *Journal of Bioinformatics and Computational Biology*, vol. 9, no. 3, pp. 383–398, 2011.

[42] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[43] R. Abagyan and M. Totrov, "Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins," *Journal of Molecular Biology*, vol. 235, no. 3, pp. 983–1002, 1994.

[44] P. N. Mortenson, D. A. Evans, and D. J. Wales, "Energy landscapes of model polyalanines," *Journal of Chemical Physics*, vol. 117, no. 3, pp. 1363–1376, 2002.

[45] M. C. Prentiss, D. J. Wales, and P. G. Wolynes, "Protein structure prediction using basin-hopping," *The Journal of Chemical Physics*, vol. 128, no. 22, Article ID 225106, 9 pages, 2008.

[46] A. S. Manalan and C. B. Klee, "Calmodulin," *Advances in Cyclic Nucleotide and Protein Phosphorylation Research*, vol. 18, pp. 227–278, 1984.

[47] A. R. Means, "Molecular mechanisms of action of calmodulin," *Recent Progress in Hormone Research*, vol. 44, pp. 223–262, 1988.

[48] K. T. O'Neil and W. F. DeGrado, "How calmodulin binds its targets: sequence independent recognition of amphiphilic α-helices," *Trends in Biochemical Sciences*, vol. 15, no. 2, pp. 59–64, 1990.

[49] B. E. Finn, J. Evenas, T. Drakenberg, J. P. Waltho, E. Thulin, and S. Forsen, "Calcium-induced structural changes and domain autonomy in calmodulin," *Nature Structural Biology*, vol. 2, no. 9, pp. 777–783, 1995.

[50] B. W. Zhang, D. Jasnow, and D. M. Zuckermann, "Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 46, pp. 18043–18048, 2007.

[51] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu, "Guiding protein docking with geometric and evolutionary information," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 3, Article ID 1242008, 16 pages, 2012.

[52] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr., "A graph-theory algorithm for rapid protein side-chain prediction," *Protein Science*, vol. 12, no. 9, pp. 2001–2014, 2003.

[53] M. Zhang and L. E. Kavraki, "A new method for fast and accurate derivation of molecular conformations," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 1, pp. 64–70, 2002.

[54] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 10, pp. 3352–3357, 2004.

[55] A. Shehu, "An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations," in *Robotics: Science and Systems*, pp. 241–248, Seattle, Wash, USA, 2009.

[56] B. S. Olson, K. Molloy, S. F. Hendi, and A. Shehu, "Guiding search in the protein conformational space with structural profiles," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 3, Article ID 1242005, 2012.

[57] J. A. Hegler, J. Laetzer, A. Shehu, C. Clementi, and P. G. Wolynes, "Restriction vs. guidance: fragment assembly and associative memory hamiltonians for protein structure prediction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 36, pp. 15302–15307, 2009.

[58] D. A. Case, T. A. Darden, T. E. I. Cheatham et al., *Amber 9*, University of California, San Francisco, Calif, USA, 2006.

[59] H. Gong, P. J. Fleming, and G. D. Rose, "Building native protein conformations from highly approximate backbone torsion angles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 45, pp. 16227–16232, 2005.

[60] K. F. Han and D. Baker, "Global properties of the mapping between local amino acid sequence and local structure in proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 12, pp. 5814–5818, 1996.

[61] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nature Structural Biology*, vol. 10, no. 12, p. 980, 2003.

[62] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu, "Protein docking with information on evolutionary conserved

interfaces," in *Bioinformatics and Biomedicine Workshops (BIBMW '11)*, pp. 358–365, November 2011.

[63] G. Terashi, M. Takeda-Shitaka, K. Kanou, M. Iwadate, D. Takaya, and H. Umeyama, "The SKE-DOCK server and human teams based on a combined method of shape complementarity and free energy estimation," *Proteins*, vol. 69, no. 4, pp. 866–872, 2007.

[64] D. Schneidman-Duhovny, Y. Inbar, R. Nussinov, and H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking," *Nucleic Acids Research*, vol. 33, no. 2, pp. W363–W367, 2005.

[65] Y. Inbar, H. Benyamini, R. Nussinov, and H. J. Wolfson, "Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies," *Physical Biology*, vol. 2, no. 4, pp. S156–S165, 2005.

[66] Y. Inbar, H. Benyamini, R. Nussinov, H. J. Wolfson, and B. Honig, "Prediction of multimolecular assemblies by multiple docking," *Journal of Molecular Biology*, vol. 349, no. 2, pp. 435–447, 2005.

[67] S. Engelen, L. A. Trojan, S. Sacquin-Mora, R. Lavery, and A. Carbone, "Joint evolutionary trees: a large-scale method to predict protein interfaces based on sequence sampling," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000267, 2009.

[68] M. L. Connolly, "Analytical molecular surface calculation," *Applied Crystallography*, vol. 16, no. 5, pp. 548–558, 1983.

[69] R. Norel, S. L. Lin, H. J. Wolfson, and R. Nussinov, "Examination of shape complementarity in docking of unbound proteins," *Proteins*, vol. 36, no. 3, pp. 307–317, 1999.

[70] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[71] T. Kortemme and D. Baker, "A simple physical model for binding energy hot spots in protein-protein complexes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 22, pp. 14116–14121, 2002.

[72] I. Hashmi and A. Shehu, "A basin hopping algorithm for protein-protein docking," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM '12)*, J. Gao, W. Dubitzky, C. Wu et al., Eds., pp. 466–469, Philadelphia, Pa, USA, 2012.

[73] J. R. Schnell, H. J. Dyson, and P. E. Wright, "Structure, dynamics, and catalytic function of dihydrofolate reductase," *Annual Review of Biophysics and Biomolecular Structure*, vol. 33, pp. 119–140, 2004.

[74] E. Z. Eisenmesser, O. Millet, W. Labeikovsky et al., "Intrinsic dynamics of an enzyme underlies catalysis," *Nature*, vol. 438, no. 7064, pp. 117–121, 2005.

[75] K. I. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: structure-based molecular dynamics simulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 32, pp. 11844–11849, 2006.

[76] Q. Lu and J. Wang, "Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles," *Journal of the American Chemical Society*, vol. 130, no. 14, pp. 4772–4783, 2008.

[77] P. Majek, H. Weinstein, and R. Elber, *Pathways of Conformational Conformational Transitions in Proteins*, chapter 13, Taylor and Francis group, 2008.

[78] D. R. Weiss and M. Levitt, "Can morphing methods predict intermediate structures?" *Journal of Molecular Biology*, vol. 385, no. 2, pp. 665–674, 2009.

[79] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins*, vol. 57, no. 4, pp. 702–710, 2004.

[80] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, 2003.

[81] A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallographica A*, vol. 26, no. 6, pp. 656–657, 1972.

[82] V. N. Maiorov and G. M. Crippen, "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins," *Journal of Molecular Biology*, vol. 235, no. 2, pp. 625–634, 1994.

[83] J. Xu and Y. Zhang, "How significant is a protein structure similarity with TM-score = 0.5?" *Bioinformatics*, vol. 26, no. 7, pp. 889–895, 2010.

[84] J. Meiler and D. Baker, "Coupled prediction of protein secondary and tertiary structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12105–12110, 2003.

[85] B. Olson and A. Shehu, "Efficient basin hopping in the protein energy surface," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '12)*, J. Gao, W. Dubitzky, C. Wu et al., Eds., pp. 119–124, Philadelphia, Pa, USA, 2012.

[86] P. J. Ballester and W. G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes," *Journal of Computational Chemistry*, vol. 28, no. 10, pp. 1711–1723, 2007.

[87] M. L. Teodoro, G. N. Phillips, and L. E. Kavraki, "Understanding protein flexibility through dimensionality reduction," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 617–634, 2003.

[88] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, "Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 26, pp. 9885–9890, 2006.

[89] H. Stamati, C. Clementi, and L. E. Kavraki, "Application of nonlinear dimensionality reduction to characterize the confonrmational landscape of small peptides," *Proteins*, vol. 78, no. 2, pp. 223–235, 2010.

[90] M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, "Determination of reaction coordinates via locally scaled diffusion map," *Journal of Chemical Physics*, vol. 134, no. 12, Article ID 124116, 2011.

[91] E. Kanamori, Y. Murakami, Y. Tsuchiya, D. M. Standley, H. Nakamura, and K. Kinoshita, "Docking of protein molecular surfaces with evolutionary trace analysis," *Proteins*, vol. 69, no. 4, pp. 832–838, 2007.

[92] H. Kuboniwa, N. Tjandra, S. Grzesiek, H. Ren, C. B. Klee, and A. Bax, "Solution structure of calcium-free calmodulin," *Nature Structural Biology*, vol. 2, no. 9, pp. 768–776, 1995.

[93] R. Chattopadhyaya, W. E. Meador, A. R. Means, and F. A. Quiocho, "Calmodulin structure refined at 1.7 Å resolution," *Journal of Molecular Biology*, vol. 228, no. 4, pp. 1177–1192, 1992.

[94] J. L. Fallon, D. B. Halling, S. L. Hamilton, and F. A. Quiocho, "Structure of calmodulin bound to the hydrophobic IQ domain of the cardiac Cav1.2 calcium channel," *Structure*, vol. 13, no. 12, pp. 1881–1886, 2005.

[95] D. G. Rhoads and J. M. Lowenstein, "Initial velocity and equilibrium kinetics of myokinase," *Journal of Biological Chemistry*, vol. 243, no. 14, pp. 3963–3972, 1968.

[96] C. W. Müller, G. J. Schlauderer, J. Reinstein, and G. E. Schulz, "Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding," *Structure*, vol. 4, no. 2, pp. 147–156, 1996.

[97] U. Abele and G. E. Schulz, "High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap5A, showing the pathway of phosphoryl transfer," *Protein Science*, vol. 4, no. 7, pp. 1262–1271, 1995.

[98] G. J. Schlauderer, K. Proba, and G. E. Schulz, "Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP," *Journal of Molecular Biology*, vol. 256, no. 2, pp. 223–227, 1996.

[99] K. Diederichs and G. E. Schulz, "The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution," *Journal of Molecular Biology*, vol. 217, no. 3, pp. 541–549, 1991.

[100] J. Ådén and M. Wolf-Watz, "NMR identification of transient complexes critical to adenylate kinase catalysis," *Journal of the American Chemical Society*, vol. 129, no. 45, pp. 14003–14012, 2007.

[101] C. Snow, G. Qi, and S. Hayward, "Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motions," *Proteins*, vol. 67, no. 2, pp. 325–337, 2007.

[102] B. Olson and A. Shehu, "Populating local minima in the protein conformational space," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM '11)*, pp. 114–117, November 2011.

[103] H. Choset, K. M. Lynch, S. Hutchinson et al., *Principles of Robot Motion: Theory, Algorithms, and Implementations*, MIT Press, Cambridge, Mass, USA, 1st edition, 2005.