Title Page

Modeling Protein Conformational Ensembles: From Missing Loops to Equilibrium Fluctuations

Research Article

Authors

Amarda Shehu[1], Cecilia Clementi[2,3], Lydia E. Kavraki[1,3,4]

[1]*Department of Computer Science, Rice University, Houston, Texas, 77005*

[2]*Department of Chemistry, Rice University, Houston, Texas, 77005*

[3]*Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, 77030*

[4]*Department of Bioengineering, Rice University, Houston, Texas, 77005*

Corresponding Authors

Cecilia Clementi
Email: cecilia@rice.edu
Phone: +1-713-348-3485
Fax: +1-713-348-5155

Lydia E. Kavraki
Email: kavraki@rice.edu
Phone: +1-713-348-5737
Fax: +1-713-348-5930

Short Title: Modeling Protein Conformational Ensembles

Keywords: Protein flexibility; equilibrium mobility; loop modeling; inverse kinematics; robotics; Boltzmann statistics.

Modeling Protein Conformational Ensembles: From Missing Loops to Equilibrium Fluctuations

Amarda Shehu[1], Cecilia Clementi[2,3] *, Lydia E. Kavraki[1,3,4] †

[1]*Department of Computer Science, Rice University, Houston, Texas, 77005*

[2]*Department of Chemistry, Rice University, Houston, Texas, 77005*

[3]*Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas, 77030*

[4]*Department of Bioengineering, Rice University, Houston, Texas, 77005*

Abstract

Characterizing protein flexibility is an important goal for understanding the physical-chemical principles governing biological function. This paper presents a Fragment Ensemble Method to capture the mobility of a protein fragment such as a missing loop and its extension into a Protein Ensemble Method to characterize the mobility of an entire protein at equilibrium. The underlying approach in both methods is to combine a geometric exploration of conformational space with a statistical mechanics formulation to generate an ensemble of physical conformations on which thermodynamic quantities can be measured as ensemble averages. The Fragment Ensemble Method is validated by applying it to characterize loop mobility in both instances of strongly stable and disordered loop fragments. In each instance, fluctuations measured over generated ensembles are consistent with data from experiment and simulation. The Protein Ensemble Method captures the mobility of an entire protein by generating and combining ensembles of conformations for consecutive overlapping fragments defined over the protein sequence. This method is validated by applying it to characterize flexibility in ubiquitin and protein G. Thermodynamic quantities measured over the ensembles generated for both proteins are fully consistent with available experimental data. On these proteins, the method recovers non-trivial data such as order parameters, residual dipolar couplings, and scalar couplings. Results presented in this work suggest that the proposed methods can provide insight into the interplay between protein flexibility and function.

*Electronic mail: cecilia@rice.edu, Phone: +1-713-348-3485, Fax: +1-713-348-5155

†Electronic mail: kavraki@rice.edu, Phone: +1-713-348-5737, Fax: +1-713-348-5930

1 Introduction

Experimental and simulation studies have established that proteins are not rigid molecular objects[1,2] but instead exhibit internal motions that are often essential for their function[3–5]. As a flexible molecule, a protein may populate a large ensemble of different structures. In particular, loop fragments are oftentimes highly mobile even in generally stable proteins. Such mobile loops are not easily characterized by X-ray crystallography as they may introduce significant disorder in a protein crystal. In fact, partially resolved protein structures are reported in these cases, with the loop fragment missing.

Finding a physically relevant conformation for a missing loop fragment in a given protein structure is an important problem (known as "loop modeling"[‡]) in automated crystallographic protein structure determination, homology modeling[6,7], and ab initio structure prediction[8,9]. The problem involves generating a peptide conformation whose N- and C- terminal residues attach to the fixed anchor residues of the two protein segments at either end of the loop. However, proposing a single peptide conformation fails to address the mobility of the missing loop. In light of the high variation of loop structures in proteins, one or few conformations may not adequately represent the diversity in the ensemble of conformations assumed by a mobile missing loop.

Loops are not the only flexible fragments in a protein. An entire protein can undergo conformational changes that may be essential to its biological function[2,10]. Compounding evidence from experiment, simulation, and theory indicates that the characterization of protein functions, such as enzymatic reactions, ligand binding, and protein/protein interactions, requires considering a protein native state as a dynamical ensemble of conformations rather than one single structure[2–5,10].

In this work we address both the problem of modeling mobile loops and the characterization of flexibility of an entire protein at equilibrium conditions. Motivated by recent computational techniques for studying protein flexibility[11–13], the presented work aims to provide a complete characterization of equilibrium fluctuations in proteins.

We propose the Fragment Ensemble Method (FEM) to address equilibrium mobility in the loop modeling problem. Given an incomplete protein structure and the amino acid sequence of the

[‡]Alternative names include loop/fragment completion, gap completion, loop closure, or fragment fitting

missing loop, the proposed method generates an ensemble of low-energy loop conformations that complete the given protein structure. The method combines a statistical mechanics formulation with an efficient exploration of conformational space[14–17], exploiting analogies between proteins and robots[18,19] to model a loop fragment as an open kinematic chain. FEM is based on a multi-scale approach. Many backbone-resolution loop conformations are first generated to satisfy the geometric and energetic constraints imposed by the given protein structure. These conformations are then structurally and energetically refined to obtain an ensemble of low-energy atomistic-resolution loop conformations. We validate the proposed method by using it to characterize loop structure and mobility in both instances of strongly stable and completely disordered loops. In each instance, fluctuations measured over a generated ensemble fully agree with experimental and simulation data.

FEM is not limited to applications on missing loops but can generate an ensemble of physical conformations for any fragment in a protein. We exploit this capability and extend this method into the Protein Ensemble Method (PEM) to characterize mobility over an entire protein. PEM generates ensembles of conformations for consecutive overlapping fragments defined over a protein sequence and combines results from ensembles of neighboring fragments. We validate this method by applying it to obtain a complete characterization of the structural flexibility of two proteins (ubiquitin and protein G) under equilibrium conditions. We show that for both proteins thermodynamic quantities measured over the generated ensembles are fully consistent with available experimental data.

Presented applications of the proposed methods indicate the potential of this work in obtaining valuable information on the interplay between protein flexibility and function. By characterizing fluctuations around a protein structure at an atomistic level of detail, the presented work can help design targeted wet-lab experiments and simulations to further improve our understanding of the physical-chemical principles governing biological function.

This article is organized as follows. We first provide more context for the loop modeling problem through a brief review of related work in section 2. The proposed methods are described in section 3. In section 4 we analyze ensembles generated by FEM for loops in chymotrypsin inhibitor 2 (CI2), the variable surface antigen (VlsE), and α-lactalbumin (α-Lac) (loops of length 12, 20, and 26 residue,

respectively). We show that there is good agreement between thermodynamic quantities measured over each ensemble and corresponding data from experiment and simulation. In section 4 we also validate the proposed PEM by analyzing protein ensembles generated for ubiquitin and protein G. We show that thermodynamic quantities measured over each ensemble correlate remarkably well with Nuclear Magnetic Resonance (NMR) data such as order parameters, residual dipolar couplings, and 3-bond scalar couplings. We conclude in section 5 with a summary and discussion of future work.

2 Current Methods for Modeling Missing Loops

FEM explores the equilibrium mobility of a missing protein fragment by dealing with the core problem of fitting a generated fragment conformation with a given protein structure. Driven by applications in X-ray crystallography, homology modeling, and ab initio structure prediction, existing work[14,20–34] focuses on fitting a generated loop conformation to model an unknown loop.

Database methods[21,26,29,30] search for candidate loops that satisfy constraints on length and geometry in homologous proteins available in structural databases such as the Protein Data Bank (PDB)[35]. Recently, the limited loop diversity in the PDB is addressed through a divide and conquer approach[29] or by constructing missing loops from short protein fragments sampled from structural libraries[34]. Database methods can model loops of up to 15 residues long[30].

Ab initio methods either sample from a discrete set of conformational parameters or adapt efficient robotics-inspired sampling algorithms to model loops of arbitrary length. Loop conformations can first be sampled from a discretized solution space through an exploration that is biased toward more populated regions of the (ϕ, ψ) map[27] and then refined through molecular dynamics simulations[20], Monte Carlo searches with simulated annealing[28], genetic algorithms[23], dynamic programming[22], bond scaling with relaxation[24], or multi-copy searches[25]. Robotics-inspired ab initio methods employ a probabilistic sampling framework[36]. Loop conformations are first sampled ignoring the constraints and later enforcing them through gradient descent[37], or the satisfaction of constraints is integrated in the sampling process[38]. In the latter case, a loop conformation that satisfies the constraints on its termini is found by solving an inverse kinematics (IK)[39] problem.

Methods that solve an IK problem to model missing loops exploit the fact that steering a terminal

residue of the loop so that it assumes the pose of the corresponding fixed anchor is very similar to controlling motions of a robot arm so that the robot hand/gripper assumes a specified target position and orientation. By modeling the polypeptide chain of a missing loop as an open kinematic chain[18], the problem of attaching the terminal residues of a loop to their corresponding fixed anchors can be posed as an IK problem: Solve for the degrees of freedom (DOFs) of the kinematic chain so that a terminal anchor of the loop assumes its target pose.

Robotics-inspired techniques[38,40] that employ exact IK solvers to enumerate all solutions[18,41–44] can do so on sub-chains of no more than 6 DOFs. More recently, this limitation has been pushed to 9 DOFs[45]. Currently, only optimization-based IK solvers[46,47] can deal with an arbitrary number of DOFs. Two such methods, random tweak[46] and cyclic coordinate descent[47], iteratively solve a system of equations until the constraints on the loop termini are satisfied. Due to a linear time complexity in the number of DOFs, numerical stability, and the ability to allow external constraints on the DOFs with predictable behavior, cyclic coordinate descent has become the method of choice in modeling missing loops of arbitrary length[31–33].

3 Materials and Methods

We first provide in section 3.1 a brief overview of the main ingredients of FEM and PEM. These methods are detailed in sections 3.2 and 3.3, respectively. An analysis of their robustness is presented in section 3.4. Finally, useful implementation details are provided in section 3.5.

3.1 Overview of Proposed Methods

Since the proposed FEM is generally applicable to any protein fragment and not just a loop, we describe it hereafter in terms of generating an ensemble of physical conformations for a protein fragment. Given an incomplete protein structure and the amino acid sequence of the missing fragment, FEM generates an ensemble of physical fragment conformations that fit with the given protein structure through essentially a three-step multi-scale approach:

(i) *Backbone Geometric Exploration:* The conformational space available to the backbone of a missing fragment is explored to generate fragment conformations that fit with a given protein structure without introducing steric clashes (details are found in section 3.2.1). The obtained conformations are passed on to step (ii).

(ii) *Side-chain Exploration for a Fixed Backbone:* The configurational space available to the side chains of a fragment is explored to add all-atom detail to each fitted fragment conformation without introducing collisions (details are found in section 3.2.2). The obtained conformations are passed on to step (iii) for energetic refinement.

(iii) *All-atom Energy Refinement:* The conformations obtained are subjected to an extensive energy minimization that seeks stabilizing interactions between atoms of the fitted fragment and the rest of the protein (details are found in section 3.2.3). Each fragment conformation is retained in the ensemble if the corresponding completed protein conformation has energy lower than a given cutoff value.

Steps (i) through (iii) of FEM allow us to efficiently generate a large ensemble of fragment conformations whose corresponding completed protein conformations are physically relevant. A statistical mechanics formulation is employed to weight each generated conformation according to its Boltzmann probability. Such statistical weighting, detailed in section 3.2.4, leads to the definition of a statistical ensemble which allows us to measure thermodynamic quantities as ensemble averages for direct validation with data from experiment and simulation studies.

Additionally, we extend the proposed FEM into PEM, which allows us to study the flexibility of an entire protein. PEM, detailed in section 3.3, consists of three steps:

(i) A window is slided over a protein sequence to define consecutive overlapping fragments.

(ii) FEM is applied to obtain an ensemble of low-energy conformations for each fragment.

(iii) Ensembles of consecutive overlapping fragments are combined to define a statistical ensemble of physically relevant conformations for the entire protein. Protein fluctuations measured over this ensemble are tested against available experimental data.

We now describe both methods in detail.

3.2 A Method for Addressing the Equilibrium Mobility of a Missing Fragment

FEM generates and fits fragment conformations to obtain an ensemble that represents the equilibrium conformational diversity of a missing fragment. A generated fragment conformation is fitted with a given protein structure by solving an inverse kinematics[39] problem. The formulation of this problem requires that we first define a missing fragment.

Let the residues of a protein from the N- to C- terminus be numbered 1 to n. We say that a fragment $[n_1, n_2]$ of the protein is missing if atomic coordinates are available only for residues from 1 to n_1 and residues from n_2 to n. We define the missing fragment $[n_1, n_2]$ as the polypeptide chain consisting of residues from n_1 to n_2, including n_1 and n_2. Finding a conformation for the missing fragment involves generating coordinates for all atoms of its polypeptide chain. Doing so in a way that fits the fragment with the given protein structure requires that the coordinates of residues n_1 and n_2 in the fragment conformation be as those of residues n_1 and n_2 in the given protein structure. In this sense, residues n_1 and n_2 are "duplicated": those in the given protein structure are fixed and so referred to as stationary or fixed anchors; those in the fragment move as one tries to find new coordinates for the fragment's atoms and are referred to as mobile anchors.

Hence, modeling an unknown fragment involves finding coordinates for its residues so that its mobile anchors attach to the stationary anchors in the given protein structure. Attaching a mobile anchor to its stationary counterpart means translating the mobile anchor so that one of its backbone atoms assumes its target position in the stationary anchor and orienting the anchor so that all its N, C_α, and C backbone atoms properly align with their counterparts in the stationary anchor residue. A mobile anchor is said to have reached its target pose when it assumes its target position and orientation in space.

The problem of modeling an unknown fragment typically consists of two steps: (i) obtain initial coordinates for the atoms of the polypeptide chain of the fragment; and (ii) modify the fragment conformation so the mobile anchors finally assume their target poses in the stationary anchors.

The first step can be addressed in different ways. In this work, a biologically relevant polypeptide chain for an unknown fragment such as a missing loop is initially obtained from a sequence-homologous protein structure selected from the PDB[35]. Any missing atom information[§] is completed through the PSFGEN[48] package. A large set of different conformations of the polypeptide chain of the fragment are then obtained by modifying the chain's dihedral angles, as described in section 3.2.1. Conformations obtained in this way do not generally fit with the given protein structure, as illustrated in Figure 1(a) for a loop, since the mobile anchors n_1 and n_2 may not be

[§]Structures reported in the PDB commonly miss hydrogen or side-chain atoms.

attached to their stationary counterparts. Indeed, fragment conformations depend in a non-trivial way on the amino acid sequence of the fragment and the environment provided by the rest of the protein.

The second step is the core of the problem. One mobile anchor of the fragment, such as n_1 in Figure 1(b), is easily attached to its stationary counterpart through rigid body transformations, a translation and two rotations to align the backbone atoms of the mobile anchor to their stationary counterparts in the fixed anchor. As illustrated in Figure 1(c), the resulting fragment conformation needs to be modified so as to attach the remaining mobile anchor n_2 to its stationary anchor. This problem is often referred to as "closing the fragment" or "closing the loop" in the context of loop modeling. It is solved in the *Backbone Geometric Exploration* step, which takes as inputs the given protein structure and the polypeptide chain of the missing fragment already attached to one fixed anchor and outputs fragment conformations that fit with the given protein structure.

3.2.1 Step (i): Backbone Geometric Exploration

We start by stripping away all but backbone atoms off the polypeptide chain obtained for the missing fragment. Working with a coarse resolution allows us to make direct use of analogies between proteins and robots[18,19] that are often exploited to adapt powerful robotic space exploration methods to the study of protein systems[14–17]. In keeping with these analogies, we model the backbone chain of a fragment as an open kinematic chain, where a protein's atoms are equivalent to robotic links and rotatable bonds connecting atoms to joints connecting links. We employ the idealized geometry, where the bond lengths and bond angles are kept fixed in their equilibrium values. The only DOFs employed at this stage are the ϕ, ψ backbone dihedral angles starting at residue $n_1 + 1$ and ending at residue $n_2 - 1$.

Many different initial conformations for the backbone of the fragment are generated by sampling values for these DOFs uniformly at random in $[-\pi, \pi]$. Considering only the backbone reduces the dimensionality of the sampled conformational space and allows for an efficient exploration. Each sampled initial conformation is closed through the cyclic coordinate descent (CCD) algorithm[47] already employed in loop modeling[31–33]. Our implementation of CCD follows closely that in ref.[31]

The CCD algorithm closes each generated fragment conformation by solving the following IK

problem: Given the positions of the backbone atoms of the stationary anchor n_2, assign values to the DOFs of the kinematic chain modeling the fragment so that the backbone atoms of the mobile anchor n_2 assume their target positions in the stationary anchor. CCD recasts this problem as a minimization problem. Given one particular DOF (*i.e.,* backbone dihedral angle) of the kinematic chain, the algorithm analytically finds the value yielding the minimum distance between residue n_2 of the fragment and its target pose in the given protein structure. CCD proceeds in cycles. At each cycle it iterates over all DOFs according to a prespecified order, updating each DOF one at a time, until the resulting pose of the mobile anchor is within a cutoff distance ϵ from the target pose. Details can be found in ref.[31]

Each conformation closed with CCD depends on the initial fragment conformation sampled. The dependence of CCD on an initial conformation is a useful feature that we exploit to generate many different fragment conformations that complete a given protein structure without introducing steric clashes (see the *Supplementary Material* for pseudocode-level details of the *Backbone Geometric Exploration*). The completed structure is deemed collision-free if its energy is below a maximum energy value $E_{\max}$ ¶.

We also investigate the potential dependence of CCD on the order the DOFs are updated in each CCD cycle. We explore two CCD implementations corresponding to two different orders: one where the DOFs are sequentially ordered from the N- to the C- terminus (as employed in loop modeling[31–33]) and another where the DOFs are randomly permuted in each CCD cycle. A comparison of two sets of closed fragment conformations, each generated with a particular CCD implementation, allows us to conclude that the order in which CCD modifies the DOFs does not significantly affect the properties of the final ensemble of conformations generated (see section 4 for the analysis and the *Supplementary Material* for details).

3.2.2 Step (ii): Side-chain Exploration for a Fixed Backbone

The *Backbone Geometric Exploration* step generates many different backbone-resolution conformations for a missing fragment. Since it only modifies the backbone of a fragment, the side chains of the polypeptide chain of the fragment are not in their optimal configurations in each generated

¶Parameters are introduced to keep the description of the methods general. Values to these parameters are empirically determined and listed in section 3.5.

backbone conformation. Therefore, values for the dihedral angles of these side chains are sampled uniformly at random in $[-\pi, \pi]$ to explore multiple side-chain configurations. For each backbone conformation, the side-chain dihedral space is explored until an all-atom fragment conformation is found whose corresponding completed protein conformation C is collision-free.

3.2.3 Step (iii): Energetic Refinement of a Modeled Fragment

To render interactions between atoms of a fragment conformation and the rest of the protein favorable, each completed protein conformation C is subjected to extensive energy minimization. Energy is measured by physical force fields such as CHARMM[49] or AMBER[50]. We design the energetic refinement of C to attribute unfavorable interactions mainly to a fragment's atoms, since the conformation corresponding to the given protein structure is considered feasible.

To achieve this goal, we interleave two strategies that mainly explore fluctuations of a closed fragment to minimize the energy of C while maintaining the given protein structure. The first, closure-constrained backbone refinement, inspired by ref.[32,33], modifies the backbone dihedrals of a fragment during minimization. The second, closure-constrained conjugate gradient descent, relaxes the idealized geometry model and allows all atoms' coordinates to change as dictated by the force field for crucial interactions of the fragment with the rest of the protein. While exploring small fluctuations of the given protein structure, this strategy attributes most of the mobility to the fragment's atoms.

Since both minimization strategies are local searches that may converge to local minima, they serve as relaxation steps for each other. If after N steps of the closure-constrained conjugate gradient descent, the improvement in energy is less than a cutoff value η, this indicates failure to escape from a local minimum of the energy landscape. Therefore, the minimization switches to the closure-constrained backbone refinement which can further minimize energy. The two strategies interleave with each other for a maximum of $N_{\max}$ minimization steps, testing after every N steps whether to terminate the minimization (if the improvement in energy is less than a convergence value μ).

Closure-constrained Backbone Refinement: Since the *Backbone Geometric Exploration* uses $m = 2(n_2 - n_1 - 1)$ dihedral DOFs to satisfy three positional and three orientational constraints of the mobile anchor n_2, the subspace defined by the remaining $m-6$ redundant DOFs, the self-motion

manifold[51], is explored as in ref.[32,33] to minimize the energy of a completed protein conformation while keeping the fragment anchored. The self-motion manifold is explored through a steepest descent which at each step updates the backbone dihedral angles of the fragment so as to minimize the energy of the completed conformation C (details provided in the *Supplementary Material*).

Closure-constrained Conjugate Gradient Descent: A conjugate gradient descent is performed on the energy landscape defined by the pseudo-energy function $E = E_{\text{forcefield}} + \sum_{atom\, i \notin \text{fragment}} K_{d_i} \cdot |\vec{x_i}(C) - \vec{x_i}(C_{\text{rest}})|^2$, where $\vec{x_i}$ indicates the 3D position of atom i during minimization and C_{rest} refers to the conformation corresponding to the rest of the protein structure. Minimizing the second term as well as the energy (measured in the first term) ensures that more mobility is asked of the fragment's atoms for stable interactions with the given protein structure. The extent to which an atom i outside the fragment moves away from its position in C_{rest} depends on the strength of interactions between atoms of the fragment and C_{rest} and is modeled through the damping constant K_{d_i}. This constant is empirically determined for each protein in this study.

3.2.4 Obtaining an Ensemble of Physical Fragment Conformations

Steps (i) through (iii) of FEM yield many all-atom closed fragment conformations of low energy. Closed fragment conformations whose corresponding completed protein conformations are of energy no higher than a cutoff value of 20 kcal/mol from a reference energy[‖] are deemed physically relevant and are added to an ensemble $\Omega_{[n_1,n_2]}$ of physical fragment conformations. We point out that fragment conformations can be generated independently from one another and so their computation is easily distributed. The issue of ensemble convergence, i.e., how many fragment conformations need to be generated to obtain a reliable equilibrium ensemble, is discussed in detail in the *Supplementary Material*.

Probability of a Local Fluctuation: A statistical mechanics formulation is employed to weight each conformation $C \in \Omega_{[n_1,n_2]}$ with energy E_C according to its Boltzmann probability $P(C) = P_{\text{ref}} e^{-\frac{E(C)-E_{\text{ref}}}{RT_0}}$, where P_{ref} and E_{ref} are the probability and the energy of C_{ref}, T_0 is the room temperature (300 K), and R is the gas constant. The reference probability P_{ref} can be arbitrarily

[‖] When a reference energy is not available, the minimum-energy completed conformation is used instead.

set equal to 1 as the calculation of average quantities is independent of the actual value of $P_{\rm ref}$. A cutoff value of 20 kcal/mol for $E(C) - E_{\rm ref}$ allows to discard generated conformations that do not contribute to thermodynamic averages measured over the ensemble of completed conformations (conformations where this cutoff is higher than 20 kcal/mol have an extremely low Boltzmann probability ($\lesssim 10^{-15}$) at room temperature T_0). The Boltzmann average $\langle X_i \rangle_{[n_1,n_2]}$ of a measurable quantity X_i at a given position i (such as, for instance, the value of the root-mean-square deviation - RMSD - for a given residue) is computed over all conformations $\{C\}$ of the ensemble $\Omega_{[n_1,n_2]}$ associated with fragment $[n_1, n_2]$ as:

$$\langle X_i \rangle_{[n_1,n_2]} = \frac{\sum_{C \in \Omega_{[n_1,n_2]}} e^{-\frac{E(C)-E_{\rm ref}}{RT_0}} X_i(C)}{Z}$$

where $Z = \sum_{C \in \Omega_{[n_1,n_2]}} e^{-\frac{E(C)-E_{\rm ref}}{RT_0}}$ is the partition function associated with the ensemble $\Omega_{[n_1,n_2]}$.

3.3 From Local to Global: A Method for Combining Local Fluctuations to Explore Protein Equilibrium Ensembles

Since FEM, described in section 3.2, generates an ensemble of physical conformations for any protein fragment, we employ it as a component in PEM to capture equilibrium fluctuations over an entire protein. We describe here the steps of this method in detail.

3.3.1 Step(i): Defining Consecutive Overlapping Fragments

Fragments are first defined by sliding a window of l residues along the polypeptide chain of the protein, with significant overlap of $\delta l \simeq l$ residues between two consecutive fragments. By using a significant overlap δl between two consecutive windows (*i.e.,* $\delta l \simeq l$) it is possible to characterize the flexibility of an entire protein self-consistently. Let us assume an initial window size of l_0 residues is selected. If significant discrepancies arise on the fluctuations at a given position as obtained from different overlapping windows**, then it means that the finite size of the window significantly distorts the fluctuations of the fragment of interest; therefore, the size of the sliding window must be increased by a finite number of residues dl to become $l = l_0 + dl$. Window size and overlap between neighboring windows are incremented by 5 residues until full consistency is reached in the fluctuations obtained from the analysis of overlapping windows enclosing each residue. The final

**The comparison of the fluctuations at a given position as obtained from different windows is performed after discarding the first few and last residues in each windows, as they are clearly constrained to be fixed in our algorithm.

window size and overlap are chosen so that no artificial constraints are introduced by the finite size of the window and so that fluctuations of neighboring overlapping fragments can be combined together to characterize the flexibility of the whole protein (see Figures 3(a2) and (b2) in section 4).

3.3.2 Steps (ii)-(iii): Obtaining and Combining Fragment Ensembles

An ensemble of relevant conformations for each of the defined fragments is generated as described in section 3.2.4. Obtained fragment ensembles are combined to provide information on the flexibility of the whole protein. Any measurable quantity X_i, for a given residue i, is obtained as a weighted average over the equilibrium ensembles of fragments overlapping in residue i. For example, if a window of size 30 and overlap of 25 is used, residue 19 is contained in fragments $[1, 30]$, $[5, 35]$, $[10, 40]$, and $[15, 45]$. Therefore, any averaged quantity for this residue can be obtained independently over these four fragment ensembles as $\langle X_{19}\rangle_{[1,30]}$, $\langle X_{19}\rangle_{[5,35]}$, $\langle X_{19}\rangle_{[10,40]}$, and $\langle X_{19}\rangle_{[15,45]}$. When fragment length and extent of overlap are large enough to cover the size of a typical fluctuation for the protein under study, averages computed over different overlapping ensembles yield self-consistent results (as it is in our case, see Figures 3(a2) and (b2)). The average value $\langle X_i\rangle$ is then defined by averaging over all the different fragment ensembles embracing residue i, $\{\,[n_1, n_2] \;\mid\; i \in [n_1, n_2]\}$, as follows:

$$\langle X_i\rangle = \sum_{\{\,[n_1,n_2] \;\mid\; i\in[n_1,n_2]\}} \frac{\langle X_i\rangle_{[n_1,n_2]}\, w(i,[n_1,n_2])}{\mathcal{N}}$$

where $\mathcal{N} = \sum_{\{\,[n_1,n_2] \;\mid\; i\in[n_1,n_2]\}} w(i,[n_1,n_2])$ is the normalization factor. The purpose for the weighting function $w(i,[n_1,n_2])$ is to downplay the finite-size effects introduced by the finite length of each fragment. Since the terminal residues of each fragment are attached to the reference protein structure through the CCD algorithm, the motion of these and a few neighboring residues is artificially restricted, and hence their contribution to the total average needs to be either discarded or strongly reduced. Two different weighting schemes are used to correct for this effect: (i) 5 residues from either end of each fragment are discarded in the calculation of the ensemble averages, that is $w(i,[n_1,n_2]) = 0$ if $min\{|i-n_1|, |i-n_2|\} < 5$, and $w(i,[n_1,n_2]) = 1$ otherwise; (ii) a Gaussian distribution is used to progressively decrease the contribution of the residues closer to the fragments ends, that is $w(i,[n_1,n_2]) = e^{-\frac{1}{2}\left(\frac{\Delta i}{\sigma}\right)^2}$, where $\Delta i = |i-(n_1+n_2)/2|$ measures the distance of residue $i \in [n_1, n_2]$ from the central residue $(n_1+n_2)/2$ in fragment $[n_1, n_2]$. The parameter σ is set to $l/2$.

3.4 Measuring Robustness to Different Approximations

The weighting scheme is one approximation made by PEM. Here is a comprehensive list of all approximations we identify to measure their effects on the equilibrium mobility captured:

(i) The order in which the DOFs are progressively updated in the CCD routine. The associated error is estimated by computing differences between averages obtained from two independently generated ensembles: one where the DOFs are ordered sequentially from the N- to the C- terminus, the other by selecting the DOFs in random order (discussed in detail in the *Supplementary Material*).

(ii) The inaccuracy of the energy force field employed. The associated error is estimated by repeating the ensemble generation with two different force fields, CHARMM[49] and AMBER[50], and measuring the differences between corresponding thermodynamic averages measured over each ensemble.

(iii) The finite-size effects introduced by the definition of fragments and the nature of the CCD algorithm. Differences between averages obtained from the two different weighting schemes described above provide an estimate for the associated error.

(iv) The interleaving procedure used in the minimization of obtained conformations. The associated error is estimated by computing differences between averages obtained from two generated ensembles: one employing the interleaving minimization and the other employing the closure-constrained conjugate gradient descent only.

The errors associated with these approximations are incorporated in the error bars for ensemble averages of NMR data such as order parameters, residual dipolar couplings, and 3-bond scalar couplings. The small error bars (as shown in Figures 3 and 5 in section 4) allow us to conclude these approximations do not significantly affect the equilibrium mobility captured for the proteins employed in this work. In particular, the small size of the error bars indicates that the developed PEM is robust against these approximations. Thus, the results obtained in different fragments can be combined to produce a global picture of fluctuations over an entire protein.

3.5 Implementation Details

Backbone Geometric Exploration: In our implementation of the CCD algorithm, the maximum number $n_{\max}$ of CCD cycles is 500. The closure criterion $\epsilon = 0.001$Å. The $E_{\max}$ employed is empirically valued at 5000kcal/mol.

All-atom Energy Refinement: The maximum number of minimization steps, the frequency of testing whether the convergence criterion has been met, and the actual definition of convergence are all empirically determined quantities that work well for all the proteins used in this work: $N_{\max} = 1000$, $N = 300$, $\eta = 2$ kcal/mol, and $\mu = 20$ kcal/mol. Due to the complexity of approximating the self-motion manifold and our numerical computation of the CHARMM gradient, the steepest descent employed in the closure-constrained backbone refinement to explore motions on the self-motion manifold is limited to 50 steps. In the closure-constrained conjugate gradient descent, in CI2, α-Lac, ubiquitin, and protein G, where interactions between atoms of a fragment and of C_{ref} are strong, $K_{d_i} = 10$. In other systems such as VlsE $K_{d_i} = 100$.

Conjugate Gradient Descent: This algorithm is implemented through the OPTCG procedure in the OPT++ nonlinear optimization package[52]. The pseudo-energy function employed in the closure-constrained conjugate gradient descent and the CHARMM energy function employed in the equilibration of PDB structures are objective nonlinear functions whose first derivatives can be computed analytically. Therefore, they are modeled as NLF1[52] objects in the OPTCG procedure.

Window size and overlap: The window size and overlap employed on applications of PEM to proteins in this work are 30 and 25 residues, respectively. This window size and overlap suffice to obtain full consistency in fluctuations obtained from the analysis of different overlapping windows enclosing each residue (shown in Figures 3(a2) and (b2)).

Employed Packages: Missing atoms of a polypeptide chain are filled in with the PSFGEN[48] package. Given a file that specifies types and charges of atoms in amino acids and a PDB file with the coordinates of the existing atoms of the polypeptide chain, PSFGEN creates a new PDB file where coordinates of the missing atoms are guessed and incorporated in the respective amino acids of the polypeptide chain. The OPT++[52] package is employed for the efficient implementation of the conjugate gradient descent algorithm. The algorithms implemented in OPT++ provide robust and

efficient solutions to nonlinear optimization problems that require expensive function evaluations.

Hardware and Software Setup: The implementation was carried out in `ANSI C/C++` using the `Intel`®8.0 compilers and libraries. The experiments were run on the Rice Terascale Cluster, a 1 TeraFLOP Linux cluster based on `Intel`® `Itanium`®2 processors. Each node has two 64-bit processors running at 900MHz with 1.5MB of L2 data cache and 2GB memory per processor. On such architecture, it takes on average 67 minutes to obtain $1,000$ conformations for a fragment of 30 residues.

4 Results

We present the following results: we first demonstrate how to incorporate loop mobility in the loop modeling problem by applying FEM to the generation of ensembles of relevant loop conformations. Then we present the application of PEM to proteins where equilibrium mobility is due to local fluctuations and validate the obtained fluctuations with available experimental data.

4.1 Generating Equilibrium Ensembles of Missing Loops

Due to their high mobility and low structural conservation, modeling long loops in partially resolved protein structures remains a challenge for structural biology[34]. To first test the accuracy of FEM, we reproduce the native loops in stable proteins, such as CI2, PDB code 1COA[53], and α-Lac, PDB code 1HML[54], respectively[††]. We consider the 12-residue loop between VAL53 and ASP64 in CI2 and the 26-residue loop between LYS51 and THR76 in α-Lac. We use FEM to generate an ensemble of conformations for the considered loop in each protein.

Figure 2(a1) shows the ensemble of generated conformations for the VAL53-ASP64 loop in CI2. Qualitatively, the obtained loop conformations are clustered around the native loop as found in the equilibrated crystal structure of CI2. We quantify the equilibrium mobility of the loop by plotting in Figure 2(a2) the energy profile of the generated ensemble versus the RMSD of the generated loop conformations from the equilibrated native loop conformation. The obtained energy profile is clearly funnel-like, in full agreement with the known role and stability of this loop for the activity of CI2[55,56].

We validate residue fluctuations obtained on the generated ensemble against B factors[53] available

[††]The PDB structures are equilibrated through an energy minimization detailed in the *Supplementary Material*.

for CI2. Fluctuations of each residue are obtained by averaging through the Boltzmann statistics the residue RMSD measured in each loop conformation relative to the native loop conformation as found in the equilibrated crystal structure of CI2. Since fluctuations derived from B factors are different in magnitude from fluctuations obtained over the ensemble generated in this work, we normalize both sets of fluctuations. As shown in Figure 2(a3), the obtained fluctuations are consistent with those derived from the available B factors; the data agree with a Pearson correlation of 96% and q-factor of 28%. This agreement indicates that fluctuations of this loop are mainly local and can be obtained in isolation, even when immobilizing the rest of the protein structure.

The generated ensemble for α-Lac is shown in Figure 2(b1). As expected, the obtained loop conformations are clustered around the native loop of the equilibrated crystal structure. Figure 2(b2) reveals a funneled energy landscape with a global minimum around the native conformation found in the equilibrated crystal structure of α-Lac, similarly to the energy landscape associated with the ensemble of loop conformations generated for CI2.

The fluctuations observed over the generated ensemble for α-Lac are fully consistent with what is obtained from a Monte Carlo simulation guided to agree with hydrogen exchange protection factors[57]. In Figure 2(a3) we compare residue fluctuations measured over the ensemble generated in this work with the fluctuations reported in ref.[57] (data courtesy of M. Vendruscolo). Due to their different magnitudes, fluctuations are normalized in the comparison. A Pearson correlation of 86% and a q-factor of 24% are obtained. Interestingly, the Pearson correlation of the fluctuations obtained from our ensemble with fluctuations derived from B factor data for α-Lac[54] is 63% (data not shown), comparable to the 61% Pearson correlation obtained when comparing fluctuations derived from the B factor data to fluctuations reported in ref.[57]

An additional application of FEM to characterize the mobility of an internal loop at equilibrium is provided in section *Application of FEM to Model Conformational Ensembles of Internal Loops* in the *Supplementary Material*.

The examples described above provide a good testbed for the accuracy of FEM in producing ensembles of native-like loop conformations with associated steep funnel-like energy landscapes for strongly stable proteins. The most interesting application of FEM, however, is the generation of

a large ensemble of loop conformations for proteins with highly flexible loops. In this context, we present here the results obtained when applying this method to generate an ensemble of conformations for the LYS93-GLY112 loop in the crystal structure of VlsE, PDB code 1L8W[58]. This 20-residue loop is missing in the crystal structure due to its high flexibility[58]. Our analysis reveals that there are many geometrically variable conformations relevant for this loop at room temperature. The high conformational heterogeneity of the closed loop conformations can be seen in Figure 2(c1). The heterogeneity of these loop conformations is quantified in Figure 2(c2), where the energy landscape associated with the generated ensemble is plotted as a function of the RMSD from the most stable complete protein conformation obtained through FEM. Figure 2(c2) shows a plateau-like energy landscape, which is very different from the funnel-like landscapes obtained for the loops in CI2 and α-Lac.

To validate the ensemble generated for the missing loop of VlsE, we compare the magnitudes of the structural fluctuations per residue (measured relative to the lowest energy structure obtained) with disorder scores computed from the amino acid sequence of VlsE through the PONDR package[59,60]. We should note that the disorder scores predicted by the PONDR package[59,60] for the loop are all well above 0.5 (the boundary between disorder and order), consistent with the fact that the LYS93-GLY112 loop in VlsE is highly disordered. Since the comparison between fluctuations and disorder scores is between two different quantities of different magnitudes, we normalize both quantities. As shown in Figure 2(c3), the agreement between the residue fluctuations and the PONDR-predicted disorder scores is with a Pearson correlation of 79% and q-factor of 31%. We should note that this comparison is qualitative since the residue fluctuations and the disorder scores represent different quantities. Interestingly, both quantities, as shown in Figure 2(c3), indicate that ILE98 is the most mobile and disordered residue in the missing loop of VlsE.

4.2 Capturing Equilibrium Fluctuations in Ubiquitin and Protein G

We present here results of the application of PEM to characterize the equilibrium ensemble of two proteins: streptococcal protein G[61] (PDB code 1IGD) and human ubiquitin[62] (PDB code 1UBQ). Because of their relatively small sizes (61 residues in protein G and 76 residues in ubiquitin) and

their biological importance[‡‡], these proteins represent an ideal application for PEM. The availability of NMR data for protein G[65–67] and ubiquitin[68–70] makes it possible for us to quantitatively validate the ensembles characterized through PEM.

On both protein G and ubiquitin, windows of length 30 residues with 25-residue overlap suffice to reveal consistent fluctuations measured over ensembles of neighboring overlapping fragments. Figures 3(a1) and (b1) qualitatively show the structural variability of the generated structures for protein G and ubiquitin. The consistency of fluctuations measured over ensembles of neighboring overlapping fragments can be seen in Figures 3(a2) and (b2) where we plot the average RMSD of each residue measured over ensembles of the fragments that encompass that residue.

We validate the ensembles generated for each protein by comparing thermodynamic quantities measured over each ensemble with NMR data that probe the dynamics of each protein. We compare to order parameters (S^2) and residual dipolar couplings (RDCs) for both protein G[65–67] and ubiquitin[68–70]. For ubiquitin, 3-bond scalar couplings (^3J)[69] are also used in our comparison.

Order parameters are measured over the conformational ensembles generated with PEM as outlined in ref.[71]. The magnitudes of the RDCs measured over each ensemble are normalized with respect to those for an amide NH in the same orientation by scaling according to bond lengths and gyromagnetic ratios[72]. 3-bond scalar couplings are measured over the population of rotamers as detailed in ref.[68]

Order parameters provide information on the reorientational averaging of the NH bond. Residual dipolar couplings quantify the fluctuations on the direction of different bond vectors. The 3-bond scalar couplings measure the side-chain population of rotameric states. The difficulty of classic MD simulations in reproducing these data is related to the timescales captured by these parameters: Order parameters extracted from ^{15}N relaxation experiments capture from the picosecond to the nanosecond timescale[1]. RDCs report on averages over longer timescales of up to millisecond range and so can reveal slower protein motions over a very broad timescale[1]. Characterizing side-chain

[‡‡]protein G, a cell surface streptococcal protein, binds immunoglobulin with high affinity and potentially enhances microbial virulence. It is important in labeling and purification of antibodies and the study of protein-protein interactions[63]. Ubiquitin regulates multiple intracellular pathways in eukaryotic cells[64] and is involved in labeling proteins for proteolysis. Its involvement in protein degradation makes it important for anticancer drug discovery.

order parameters and 3-bond scalar couplings can be highly nontrivial since the time scale for the slowest side-chain rotations may be in the millisecond range[73]. In particular, scalar couplings report on rotameric averaging on timescales from few hundredths of a second to picoseconds[74].

4.2.1 Validation of protein G Fluctuations with NMR Measurements

The experimentally available S^2 data for protein G are derived from ^{15}N NMR relaxation experiments[65] and capture the fast dynamics of this protein in the picosecond to nanosecond timescale. For brevity, we will refer to them as fast S^2. Figure 4(a) shows the agreement between the ensemble measured backbone (amide) S^2 and the fast S^2 data of protein G. The Pearson correlation between the two quantities is 73%. We should note that no scaling has been applied to the measured S^2 order parameters to match to the fast S^2 data (no scaling is applied in the comparisons with the experimental data for protein G and ubiquitin). The agreement is better on α-helix and the β_2- and β_3-helix loops (residues $22-48$), indicating that most of the mobility captured by our ensemble for these residues happens on the picosecond to nanosecond timescale. However, the agreement drops on the N- and C- terminal chains and on residues $14-22$ due to a higher heterogeneity reported for these regions from our ensemble. The region between residues $14-22$ incidentally includes the "melting hot spot "[75] loop of residues $14-17$ and the beginning of the β_2-strand, residues $18-22$. The order parameters calculated over our ensemble for residues $14-17$ point to a slower timescale mobility for this region.

To validate the high heterogeneity in this region, we compare our ensemble-averaged S^2 data with order parameters for the NH bond derived in ref.[66] as they provide information on reorientational averaging of the NH bond up to the millisecond timescale. For brevity, we refer to these as slow S^2. Figure 4(b) shows a better agreement between the S^2 data measured over our ensemble and the slow S^2 data derived in ref.[66] as the Pearson correlation improves up to 83%. As Figure 4(b) shows, the agreement between the S^2 data for the residues on the N- and C- terminal chains improves, indicating that the motions in these residues happen in a slower timescale. In addition, while the magnitudes of the calculated S^2 data for residues $14-22$ are higher than those derived in ref.[66], the two profiles for this region of the protein are comparable. This further confirms that the mobility of this important region in protein G happens in a slower timescale.

To further validate slower timescale fluctuations captured by our ensemble for protein G, we compare RDCs measured over our ensemble with five sets of experimental RDC data used in refining the crystal structure[61] to obtain the NMR structure[67] of protein G. In Figure 4(c) we show that the RDCs measured over our ensemble and those experimentally measured in bicelle medium[67] agree with a Pearson correlation of 97% and q-factor of 21%. Naturally, a lower q-factor of 6% is obtained when comparing this experimental RDC data to the RDC-refined NMR structure[67] itself. Comparison of our ensemble-averaged RDC data with experimental RDCs measured over the other four media[67] reveals agreement with Pearson correlation varying from 94% to 98% and q-factor varying from 18% to 24% (data not shown). A complete comparison of the RDC-refined NMR structure reported in ref.[67] with each of the five experimentally measured RDCs reveals a q-factor varying from 5% to 7%, with an average of 6%.

4.2.2 Validation of Ubiquitin Fluctuations with NMR Measurements

In Figure 5(a) we show the agreement between the ensemble measured and the experimentally available backbone (amide S^2) and side-chain (methyl S^2) order parameter data[68,70]. The side-chain S^2 order parameters quantify the contribution of side-chain disorder. They provide indication on the heterogeneity of the population of different rotamer states for a given torsion angle: an extreme value of $S^2 = 1$ indicates no variability, while $S^2 = 0$ indicates a uniform distribution over all allowed rotamers. Figure 5(a) shows a Pearson correlation of 96% and indicates that low S^2 order parameters are found not only for residues in the carboxy-terminal region of ubiquitin, residues from $72 - 76$, but also in residues that form the core of this protein. Fluctuations of each residue can also be seen as residue RMSDs measured over the generated ensemble in Figure 3(b2).

Figure 5(b) shows the agreement between the ensemble averaged RDCs and the experimentally available ones[69]. The RDC parameters measured over the generated ensemble agree with the experimental RDC parameters with a Pearson correlation of 97% and q-factor of 23%. The only better agreement with the experimental RDC parameters comes from the NMR ensemble itself, a Pearson correlation of 99% and q-factor 14%, which is not a surprise since the NMR ensemble reported in ref.[69] is derived from the experimental RDC parameters[69].

Due to the availability of experimental scalar couplings for ubiquitin[69], we also compare measured

ensemble averages of ${}^3J$ with experimental data[69]. Figure 5(c) shows the agreement between the experimental and the ensemble measured ${}^3J_{NC_\gamma}$ and ${}^3J_{CC_\gamma}$, which are the 3-bond scalar couplings between the side-chain gamma carbon and the backbone amide nitrogen and carbonyl carbon, respectively. The ${}^3J$ parameters quantify the side-chain population of rotameric states and are related through the Karplus equation to the probability of occupation of different rotamer states for torsion angles of specific side chains[68]. As outlined in ref.[68], the ensemble of rotameric states can be used to parameterize the Karplus equation. Optimal values to the Karplus parameters A, B, C, δ can be defined to improve the agreement between observed and calculated scalar coupling data. Rather than optimize such parameters, we choose to perform a golden test and use the Karplus equation empirically parameterized for the X-ray structure of human ubiquitin reported in ref.[68] (PEM generates conformations starting from the equilibrated X-ray structure of ubiquitin). Comparing the so measured ensemble averaged scalar coupling data with the ones available from NMR reveals a Pearson correlation of 97%, which indicates that the side chains in the conformations generated by PEM populate the right rotameric states. Such a correlation is higher than the 84% and 89% Pearson correlation obtained when comparing the scalar couplings measured on the ubiquitin crystal structure[62] and NMR ensemble[69], respectively, with experimental scalar coupling data. Such a result indicates that the ensemble averaging of the side-chain dihedrals improves the agreement with experimental scalar coupling data.

4.3 Significance of Agreement with NMR Data

All results presented here have been obtained by using two different force fields: CHARMM22[49] and AMBER94[50]. These force fields have similar functional form but different parameterization strategies. It has been recently shown that MD simulations with these force fields allow to obtain similar structural and dynamical properties of proteins (see ref.[76]). The results obtained in this work are also found to be essentially independent of the choice of CHARMM22 vs. AMBER94. The small differences observed in the results obtained with the two force fields are incorporated in the error bars in Figures 4(a)-(c) and Figures 5(a)-(c). The effect of other approximations used by PEM besides the choice of the force field is also measured as outlined in section 3 and incorporated in the error bars.

It is worth stressing the importance of the recovery of RDCs in both ensembles, (shown in Figure 4(c) for protein G and Figure 5(b) for ubiquitin). While NMR[1] and molecular dynamics[2] simulations can characterize local backbone fluctuations in the picosecond-nanosecond timescale, slower motions in the millisecond-second range, of crucial interest to many functionally important biological processes[77,78], are not well understood. Recovering RDC data that report on slow timescale motions, up to the millisecond range, is an important result and confirms the validity of PEM in capturing equilibrium mobility in proteins.

In addition, the correct prediction of NMR data related to side-chain motion, such as the methyl S^2 order parameters (Figure 5(a)), and 3-bond scalar couplings ${}^3J$ (Figure 5(c)), is a significant result. The NMR ensemble available for ubiquitin[69] correlates with a Pearson correlation of 62% with the experimentally available S^2 order parameters, significantly lower than the Pearson correlation of 96% obtained with PEM. In addition, it has been previously reported that a 6ns MD simulations on ubiquitin performed in explicit solvent and reported in ref.[79] cannot capture the heterogeneity of the native state of the protein as given in the experimental S^2 order parameters[68] (the Pearson correlation with the experimental S^2 order parameters is 62%). The only other effort we know of that is successful in recovering the NMR data for human ubiquitin, presented in ref.[79], guides replica exchange MD simulations to generate ubiquitin conformations that correlate well with NOE derived distances[69] and S^2 order parameters[68] and reports Pearson correlations of no lower than 96% with experimental S^2, RDCs, and scalar couplings.

Finally, the recovery in this work of NMR data related to side-chain dynamics, scalar couplings and S^2 order parameters, is an important result since it has been estimated that the time scale for the slowest side-chain rotations may be about milliseconds[73]. As a consequence, the equilibrium distribution of side-chain conformers cannot be observed directly in MD simulations[76]. Since different conformations are generated independently in our ensembles, different low energy conformers for a given side chain can be sampled even if they are separated by a large barrier, which would hinder the transition from one to the other in MD simulations. Indeed, a closer look at our ensemble of ubiquitin structures reveals that 88% of the allowed side-chain rotamers are populated, although some are found with much smaller frequency than others (as expected in the human ubiquitin native ensemble - see ref.[68,79]). The successful recovery of these side-chain NMR data in our ensemble

(Figures 5(a)-(c)) further corroborates the validity of the proposed PEM in properly characterizing equilibrium local fluctuations.

5 Discussion and Conclusion

Capturing equilibrium mobility in proteins is important for understanding biological function. We propose a method to address the mobility of missing loops in protein structures. The method generates an ensemble of physical loop conformations on which thermodynamic quantities can be measured for validation with corresponding data from experiment and simulation. Furthermore, we extend this method to capture the equilibrium mobility of an entire protein.

Designing methods to obtain an ensemble of conformations available to a protein at equilibrium is a novel contribution of this work. The FEM proposed to model loop mobility at equilibrium makes use of an efficient robotics-inspired exploration to sample the conformational space available to a missing fragment that fits with a given protein structure. This exploration allows FEM to explore the space of arbitrarily long fragments, an advantage over database and ab-initio methods[20–30,34].

The multi-scale approach employed in this work allows to efficiently model protein fragments as kinematic chains. In addition, the use of all-atom force fields allows to accurately estimate conformational energies. A statistical mechanics formulation then provides a natural way to associate a weight to each obtained conformation and as a result allows to obtain an equilibrium conformational ensemble. This is an obvious advantage over existing exploration methods applied to proteins[32,33,40]. The extension of FEM into PEM is a novel approach to obtain equilibrium fluctuations of an entire protein by combining equilibrium fluctuations of protein fragments.

When applied to stable proteins such as CI2 and α-Lac, the proposed FEM recovers the native loops of these proteins. The generated ensembles are clustered around the native loops, and the associated energy landscapes are funnel-like. Fluctuations measured over each ensemble are fully consistent with experimental data and existing simulations. A novel application of our method on VlsE with a missing loop of 20 residues generates an ensemble whose conformational heterogeneity is consistent with the high disorder of the missing loop. These results point to an immediate future application of the proposed FEM where consideration of the crystal environment as in ref.[80] will allow to model the effects of crystal packing on loop mobility.

Applications of PEM on ubiquitin and protein G reveal fluctuations that correlate very well with order parameter, residual dipolar coupling, and 3-bond scalar coupling NMR data. The proposed PEM fully characterizes the equilibrium mobility in proteins such as ubiquitin and protein G, where mobility is not due to concerted motions. Because this method explores the mobility of a protein one fragment at a time, it is not immediately clear whether it can capture concerted motions. We are currently investigating this issue and extending the proposed methods into a more general approach. We are also investigating ways to improve the efficiency of the search for low-energy conformations in the proposed methods by, for instance, searching for optimal side-chain configurations in backbone-dependent rotamer libraries such as those provided in SCW3RL 3.0[81].

An additional consideration for future work is the inclusion of solvent effects on the modeled equilibrium fluctuations. While we do not employ the ensembles obtained in this work to make inferences about the relationship between structure and function, modeling water may allow us to answer important questions on the role of water in functional motions[82]. Investigation of more contemporary force fields is another obvious direction of future work. The agreement shown in this work between the ensembles obtained when using CHARMM22[49] vs. AMBER94[50] is similar to the one obtained in ref.[76], where different force fields of similar functional form are shown to behave comparably in MD simulations. Evidence of differences between force fields, however, on other physical conditions, such as modeling of peptide unfolding[83], indicates that protein modeling with an exhaustive set of force fields is worth investigating.

Since the recovery of NMR data probing the dynamics of proteins is generally a challenge for even long MD simulations, the successful prediction for protein G and ubiquitin in this work is a particularly significant result. The nontrivial recovery of these NMR data suggests that the methods we propose can provide detailed information on the equilibrium flexibility of proteins and so help us better understand the interplay between flexibility and function.

Acknowledgments

This work was supported by grants from NSF (CC Career CHE-0349303, LK ITR-0205671, LK and CC CCF-0523908 and CNS-0454333), ATP (003604-0010-2003), the Robert A. Welch Foundation (CC Norman Hackermann Young Investigator award, and C-1570), and the Sloan Foundation (LK).

AS is supported by a training fellowship from the Nanobiology Training Program of the W. M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NIH Grant No.1 R90 DK71504-01). The Rice Terascale Cluster used in this work is supported by NSF under Grant EIA-0216467, Intel, and Hewlett Packard.

We acknowledge Hernan Stamati for contributions to initial stages of this project. We thank Dr. Kresten Lindorff-Larsen, Dr. Michele Vendruscolo, Dr. Jean-Claude Latombe, and Dr. Kevin MacKenzie for their comments, and members of Kavraki's and Clementi's groups for stimulating discussions.

References

1. Kay LE. NMR studies of protein structure and dynamics. J Magn Reson 2005;173(2):193–207.

2. Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci USA 2005; 102(19):6679–6685.

3. Peters GH, Frimurer TM, Olsen OH. Molecular dynamics simulations of protein-tyrosine phosphatase 1B. I. Ligand-induced changes in the protein motions. Biophys J 1999;77(1):505–515.

4. Schnell JR, Dyson HJ, Wright PE. Structure, dynamics, and catalytic function of dihydrofolate reductase. Annu Rev Biophys and Biomolec Struct 2004;33(1):119–140.

5. Smith GR, Sternberng MJE, Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 2005;347(5):1077–1101.

6. Dunbrack Jr RL. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. Proteins: Struct Funct Genet 1999;37(S3):81–87.

7. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modeling. J Mol Biol 1997;267(2):352–367.

8. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys and Biomolec Struct 2001;30(1):173–189.

9. Moult J, Fidelis K, Zemla A, Hubbard T. Critical assessment of methods of protein structure prediction (CASP) round V. Proteins: Struct Funct Genet 2003;53(S6):334–339.

10. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motion on proteins. Science 1991; 254(5038):1598–1603.

11. Lee A, Streinu I, Brock O. A methodology for efficiently sampling the conformation space of molecular structures. J Phys Biol 2005;2(4):108–S115.

12. Mamonova T, Hespenheide B, Straub R, Thorpe MF, Kurnikova M. Protein flexibility using constraints from molecular dynamics simulations. J Phys Biol 2005;2(4):137–147.

13. Thorpe MF, Ming L. Macromolecular flexibility. Phil Mag 2004;84:1323–31137.

14. Cortes J, Simeon T, de Angulo R, Guieysse D, Remaud-Simeon M, Tran V. A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 2005;21(S1):116–125.

15. Amato NM, Dill KA, Song G. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. J Com Biol 2002;10(3-4):239–255.

16. Apayadin MS, Brutlag DL, Guestrin C, Hsu D, Latombe JC. Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. J Com Biol 2003;10(3-4):257–281.

17. LaValle SM, Finn PW, Kavraki LE, Latombe JC. A randomized kinematics-based approach to pharmacophore-constrained conformational search and database screening. J Comput Chem 2000; 21(9):731–747.

18. Manocha D, Zhu Y. Kinematic manipulation of molecular chains subject to rigid constraints. In: Proc Int Conf Intell Sys Mol Biol (ISMB), volume 2 1994; pp. 285–293.

19. Singh AP, Latombe JC, Brutlag DL. A motion planning approach to flexible ligand binding. In: Proc Int Conf Intell Sys Mol Biol (ISMB), volume 7 1999; pp. 252–261.

20. Bruccoleri RE, Karplus M. Conformational sampling using high temperature molecular dynamics. Biopolymers 1990;29(14):1847–1862.

21. Summers NL, Karplus M. Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro - non-Pro mutations. J Mol Biol 1990;216(4):991–1016.

22. Finkelstein AV, Reva BA. Search for the stable state of a short chain in a molecular field. Protein Eng 1992;5(7):617–624.

23. McGarrah DB, Judson RS. Analysis of the genetic algorithm method of molecular conformation determination. J Comput Chem 1993;14(11):1385–1395.

24. Zheng Q, Rosenfeld R, Vajda S, DeLisi C. Determining protein loop conformation using scaling-relaxation techniques. Protein Sci 1993;2(8):1242–1248.

25. Zheng Q, Rosenfeld R, DeLisi C, Kyle DJ. Multiple copy sampling in protein loop modeling: computational efficiency and sensitivity to dihedral perturbations. Protein Sci 1994;3(3):493–506.

26. van Vlijmen HWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267(4):975–1001.

27. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40(1):135–144.

28. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9(9):1753–1773.

29. Tossato CE, Bindewald E, Hesser J, Maenner R. A divide and conquer approach to fast loop modeling. Protein Eng 2002;15(4):279–286.

30. Du PC, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the protein databank? An update. Protein Eng 2003;16(6):407–414.

31. Canutescu AA, Dunbrack Jr RL. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 2003;12(5):963–972.

32. Lotan I, van den Bedem H, Deacon AM, Latombe JC. Computing protein structures from electron density maps: the mising loop problem. In: Erdman M, Hsu D, Overmars M, van der Stappen F, editors, Algorithmic Foundations of Robotics VI. Springer STAR Series 2005; pp. 345–360.

33. van den Bedem H, Lotan I, Latombe JC, Deacon AM. Real-space protein-model completion: an inverse-kinematics approach. Acta Crystallogr 2005;D61(1):2–13.

34. Kolodny R, Guibas L, Levitt M, Koehl P. Inverse kinematics in biology: the protein loop closure problem. Int J Robot Res (IJRR) 2005;24(2-3):151–163.

35. Berman HM, Westbrook J, Feng Z, Gilliand G, Bhat TN, Weissig H, N SI, Bourne PE. The Protein Data Bank. Nucl Acids Res 2000;28(1):235–242.

36. Kavraki LE, Svetska P, Latombe JC, Overmars M. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE T Robotic Autom 1996;12(4):566–580.

37. Yakey J, LaValle SM, Kavraki LE. Randomized path planning for linkages with closed kinematic chains. IEEE T Robotic Autom 2001;17(6):951–959.

38. Han L, Amato NM. A kinematics-based probabilistic roadmap method for closed chain systems. In: Donald BR, Lynch KM, Rus D, editors, Algorithmic and Computational Robotics: New Directions. MA: AK Peters, Wellesley 2001; pp. 233–246.

39. Craig J. Introduction to robotics: mechanics and control. 2 edition. Boston, MA: Addison-Wesley 1989, 450 pp.

40. Cortes J, Simeon T, Remauld-Simeon M, Tran V. Geometric algorithms for the conformational analysis of long protein loops. J Comput Chem 2004;25(7):956–967.

41. Chirikjian GS. General methods for computing hyper-redundant manipulator inverse kinematics. In: Proc IEEE/RSJ Int Conf Intell Robot Sys (IROS), volume 2 1993; pp. 1067 – 1073.

42. Manocha D, Canny J. Efficient inverse kinematics for general 6R manipulators. IEEE T Robotic Autom 1994;10(5):648–657.

43. Wedemeyer WJ, Scheraga HJ. Exact analytical loop closure in proteins using polynomial equations. J Comput Chem 1999;20(8):819–844.

44. Coutsias E, Seok C, Jacobson CM, Dill K. A kinematic view of loop closure. J Comput Chem 2004; 25(4):510–528.

45. Zhang M, White RA, Wang L, Goldman R, Kavraki LE, Hasset B. Improving conformational searches by geometric screening. Bioinformatics 2005;21(5):624–630.

46. Fine RM, Wang HJ, Shenkin PS, Yarmush DL, Levinthal C. Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins: Struct Funct Genet 1986;1(4):342–362.

47. Wang LT, Chen CC. A combined optimization method for solving the inverse kinematics problem of mechanical manipulators. IEEE T Robotic Autom 1991;7(4):489–499.

48. Gullingsrud J, Phillips J. PSFGEN User's Guide. Technical report, University of Illinois at Urbana-Champaign 2002. http://www.ks.uiuc.edu/Research/vmd/plugins/psfgen/.

49. MacKerell JAD, Bashford D, Bellot M, L DJR, Evanseck JD, J FM, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, K LFT, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III WE, Roux B, Schlenkrich B, Smith JC, Stote RH, Straub J, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102(18):3586–3616.

50. Wendy DC, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1994;117(19):5179–5197.

51. Burdick JW. On the inverse kinematics of redundant manipulators: characterization of the self-motion manifolds. In: Proc IEEE Int Conf Robot Autom (ICRA), volume 1 1989; pp. 264–270.

52. Meza JC. OPT++: An object-oriented class library for nonlinear optimization. Technical Report SAND94-8225, Sandia National Laboratories 1994. http://csmr.ca.sandia.gov/opt++/.

53. Jackson SE, Moracci M, elMasry N, Johnson CM, Fersht AR. Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. Biochemistry 1993;32(42):11259–11269.

54. Ren J, Stuart DI, Acharya KR. Alpha-lactalbumin possesses a distinct zinc binding site. J Biol Chem 1993;268(26):19292–19298.

55. Li A, Daggett V. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. Proc Natl Acad Sci USA 1994;91(22):10430–10434.

56. Jackson SE, Fersht AR. Contribution of residues in the reactive site loop of chymotrypsin inhibitor 2 to protein stability and activity. Biochemistry 1994;33(46):13880–13887.

57. Vendruscolo M, Pacci E, Dobson C, Karplus M. Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. J Am Chem Soc 2003;125(51):15686–15687.

58. Eicken C, Sharma V, Klabunde T, Lawrenz MB, Hardham JM, Norris SJ, Sacchettini JC. Crystal structure of lyme disease variable surface antigen VlsE of borrelia burgdorferi. J Biol Chem 2002; 277(24):21691–21696.

59. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. Genome Informatics 1999;10:30–40.

60. Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Sequence complexity of disordered protein. Proteins: Struct Funct Genet 2001;42(1):38–48.

61. Derrick JP, Wigley DB. The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. J Mol Biol 1994;243(5):906–918.

62. Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 1987; 194(3):531–544.

63. Sjőbring C, Bjőrck L, Kaster W. Streptococcal protein G-gene structure and protein binding properties. J Biol Chem 1991;266(1):399–405.

64. Pickart CM. Back to the future with ubiquitin. Cell 2004;116(2):181–190.

65. Hall JB, Fushman D. Characterization of the overall and local dynamics of a protein with intermediate rotational anisotropy: Differentiating between conformational exchange and anisotropic diffusion in the B3 domain of protein G. J Biomol NMR 2003;27(3):261–275.

66. Bouvignes G, Bernadó P, Meier S, Cho K, S G, Brueschweiler R. Identification of slow correlated motions in proteins using residual dipolar couplings and hydrogen-bond scalar couplings. Proc Natl Acad Sci USA 2005;102(39):13885–13890.

67. Ulmer TS, Ramirez BE, Delaglio F, Bax A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc 2003;125(30):9179–9191.

68. Chou JJ, Case DA, Bax A. Insights into the mobility of methyl-bearing side chains in proteins from $^3J_{CC}$ and $^3J_{CN}$ couplings. J Am Chem Soc 2003;125(29):8959–8966.

69. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 1998;120(27):6836–6837.

70. Tjandra N, Feller SE, Pastor RW, Bax A. Rotational diffusion anisotropy of human ubiquitin from ^{15}N NMR relaxation. J Am Chem Soc 1995;117(50):12562–12566.

71. Best RB, Vendruscolo M. Determination of ensembles of structures consistent with NMR order parameters. J Am Chem Soc 2004;126(26):8090–8091.

72. Tjandra N, Bax A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 1997;278(5340):1111–1114.

73. Ming D, Brueschweiler R. Prediction of methyl-side chain dynamics in proteins. J Biomol NMR 2004; 29(3):363–368.

74. Bax A, Vuister GW, Grzesiek S, Delaglio F, Wang AC, Tschudin R, Zhu G. Measurement of homo- and heteronuclear j couplings from quantitative j correlation. Meth Enzymology 1994;239:79–105.

75. Ding K, Louis JM, M GA. Insights into conformation and dynamics of protein GB1 during folding and unfolding by NMR. J Mol Biol 2004;335(5):1299–1307.

76. Price DJ, Brooks CLI. Modern protein force fields behave comparably in molecular dynamics simulations. J Comput Chem 2002;23(11):1045–1057.

77. Tousignant A, Pelletier JN. Protein motions promote catalysis. Chem Biol 2004;11(8):1037–1042.

78. Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opinion Struct Biol 2003; 13(6):748–757.

79. Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 2005;433(7022):128–132.

80. Jacobson PJ, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins: Struct Funct Bioinf 2004;55(2):351–367.

81. Canutescu AA, Shelenkov AA, Dunbrack Jr RL. A graph-theory algorithm for rapid protein side chain prediction. Protein Sci 2003; 12(9):2001–2014.

82. Mattos C. Protein-water interactions in a dynamic world. Trends Biochem Sci 2002; 27(4):203–208.

83. Hu H, Elstner M, Hermans J. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine dipeptides (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. Proteins: Struct Funct Bioinf 2003; 50(3):451–463.

84. Humphrey W, Dalke A, Schulten K. VMD - Visual Molecular Dynamics. J Molec Graphics 1996; 14(1):33–38. http://www.ks.uiuc.edu/Research/vmd/.

(a) (b) (c)

Figure 1: (a) The mobile anchors in two different polypeptide chains for the CI2 VAL53-ASP64 loop fragment, drawn in grey, are not attached to the stationary anchors drawn in black. (b) Mobile anchor n_1 is attached to its corresponding stationary anchor through rigid body transformations. (c) Rotations of the dihedral bonds of the fragment steer the other mobile anchor n_2 towards its target pose in the stationary anchor.

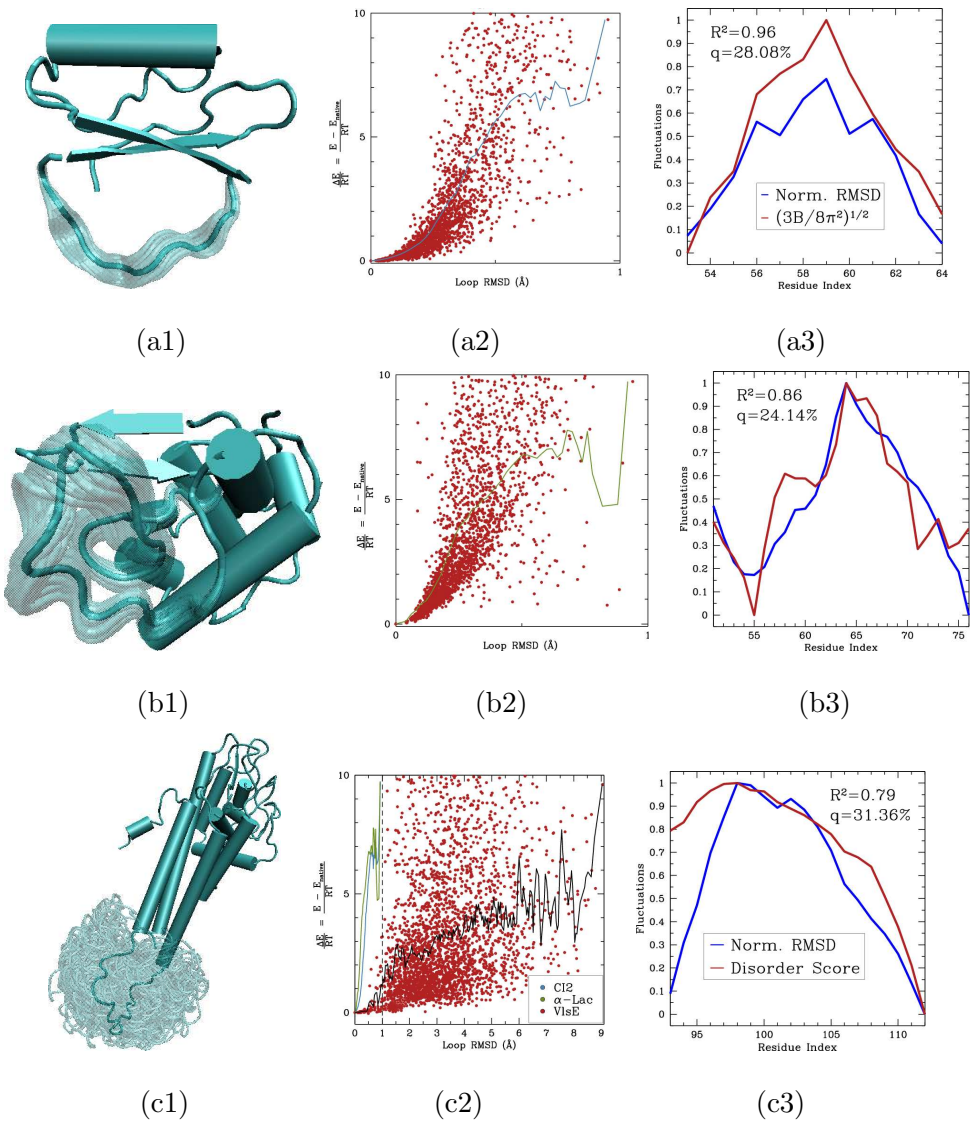
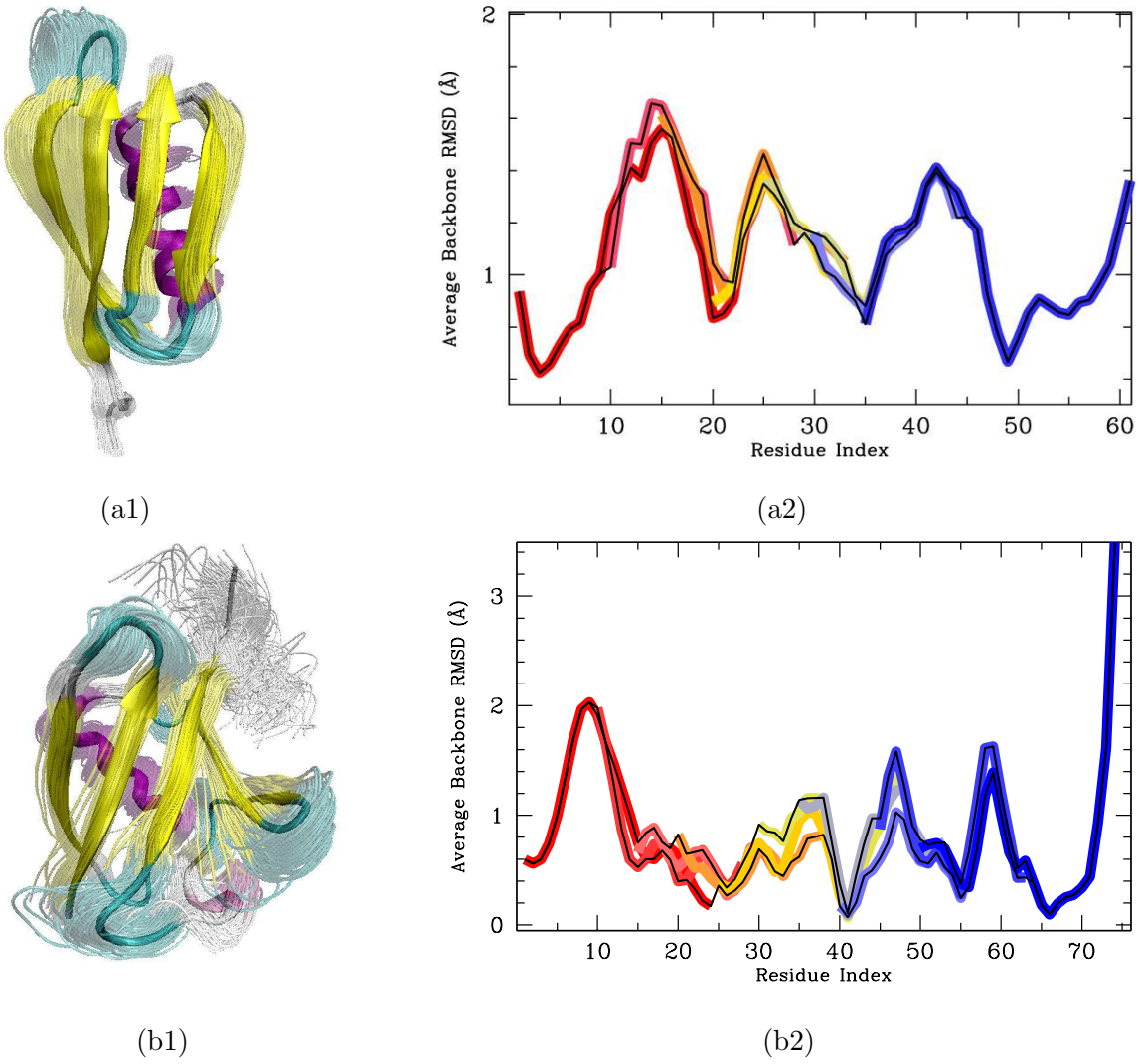


Figure 2: (a1), (b1), (c1) $5,000$ transparent loop conformations vs. opaque reference structure (equilibrated native structure for CI2 and α-Lac and lowest energy structure for VIsE). Obtained conformations are rendered with the VMD 1.8.3 software[84]. (a2), (b2), (c2) Energy landscapes associated with generated ensembles are shown by plotting the energetic difference vs. the RMSD of each conformation relative to a reference structure. Energy landscapes are shown only for conformations with energy less than 10 RT units away from the reference structure (each ensemble is reduced to 2499, 2022, and 2755 conformations). An average energy profile is computed by distributing conformations in bins every 0.001 Å away from reference structure and measuring the energy of each bin as an average over its conformations. Average energy profiles obtained for CI2 and α-Lac are very steep compared to the flat average energy profile of VIsE. (a3), (b3), (c3) Obtained fluctuations vs. B factor-derived fluctuations for the CI2 loop, fluctuations in ref.[57] for the α-Lac loop, and disorder scores for the VIsE loop.



(a1) (a2)

(b1) (b2)

Figure 3: (a1) and (b1) Obtained native ensembles for protein G and ubiquitin, respectively. (a2) and (b2) Average RMSD per residue obtained by combining the local fluctuations of all the different regions. Results for different regions are shown in different colors, from red to blue as a window of 30 residues slides from the N- to the C- terminus of the protein. The black lines mark the highest and lowest rmsd values recorded from all the different windows embracing each given residue, and provide an estimate for the uncertainty of the procedure. Two consecutive 30-residues windows have an overlap of 25 residues. The results corresponding to the first and last 5 residues of each fragment are discarded as they are biased by the finite size of the window.

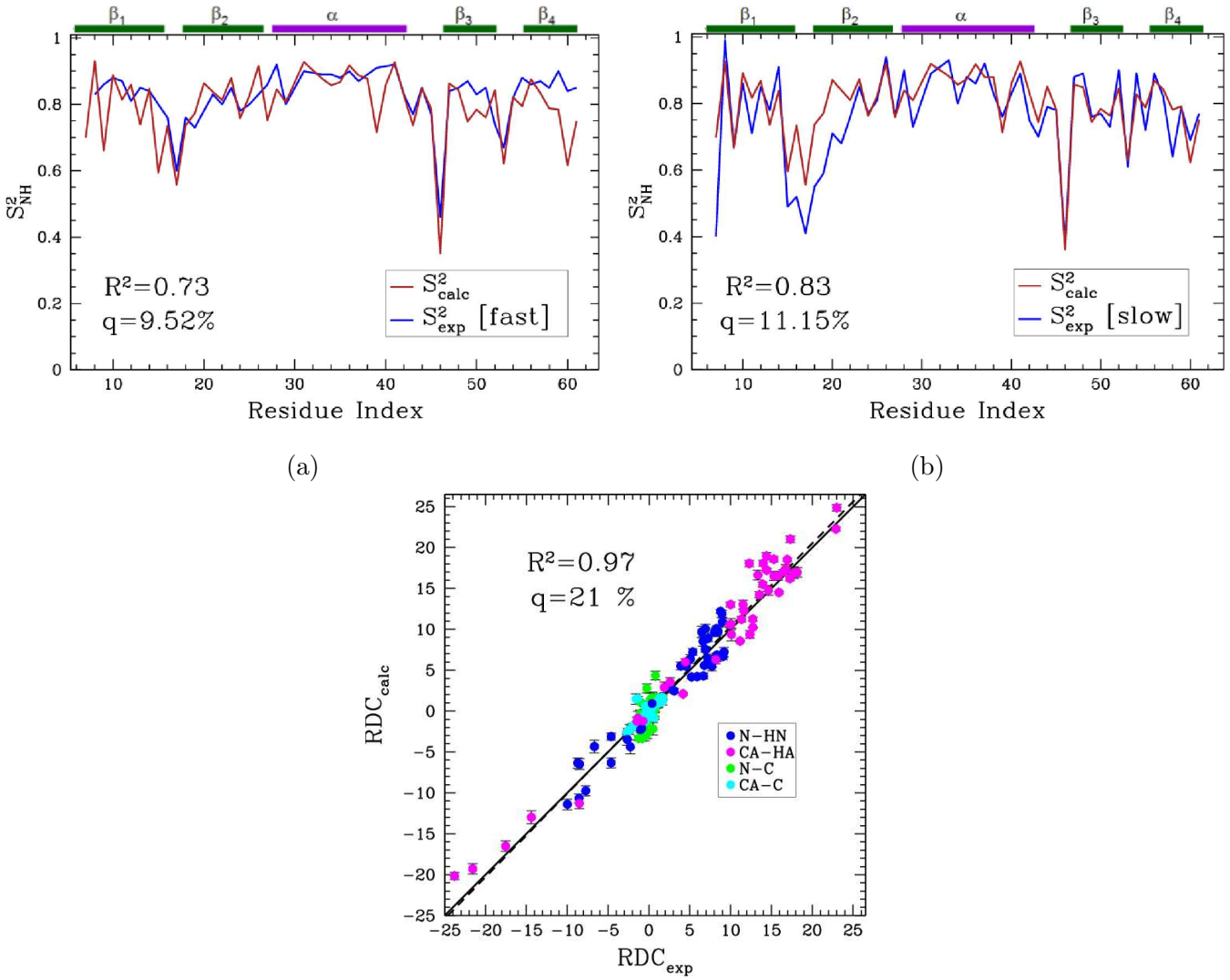


Figure 4: Comparison of NMR data with thermodynamics data measured over the generated ensemble for protein G. (a) Comparison of S^2 backbone (amide) order parameters measured over our ensemble (S^2_{calc}) with fast S^2_{NH} data obtained from NMR relaxation measurements (S^2_{exp}). (b) Comparison of S^2 backbone (amide) order parameters measured over our ensemble (S^2_{calc}) with slow S^2_{NH} data obtained from NMR relaxation measurements (S^2_{exp}). (c) Comparison of residual dipolar coupling (RDC) parameters as obtained in our ensemble (RDC_{calc}, on the y-axis), and from NMR relaxation measurements (RDC_{exp}, on the x-axis). Results for different bond types are shown in different colors. (a)-(c) The dashed black line indicates the linear least squares regression fit on the two sets of data, while the continuous line represents the identity line.

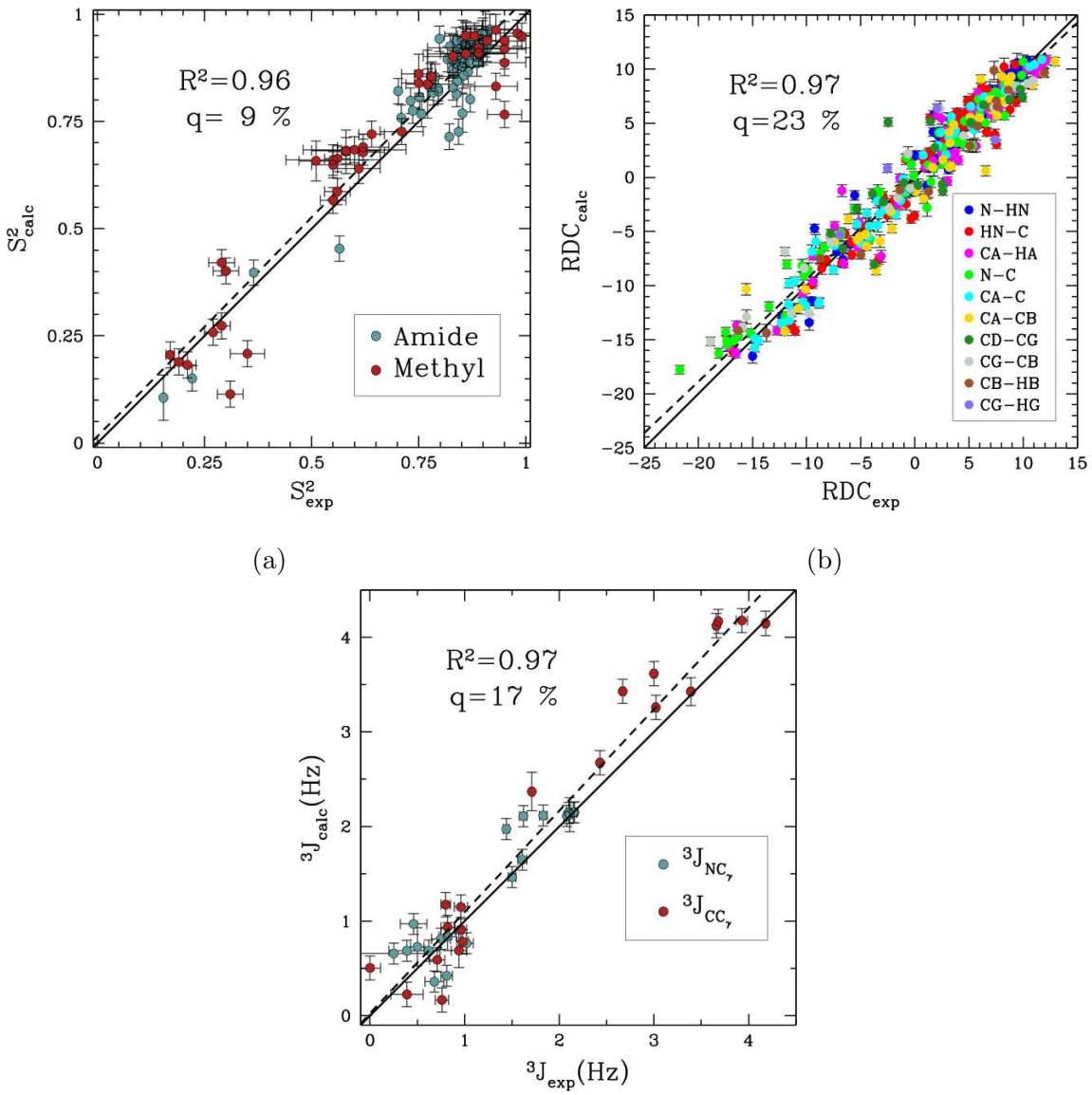


Figure 5: Comparison of NMR data with thermodynamics data measured over the generated ensemble for ubiquitin. (a) Comparison of S^2 order parameters for backbone (amide S^2) and side chains (methyl S^2), as obtained in our ensemble (S^2_{calc}, on the y-axis), and from NMR relaxation measurements (S^2_{exp}, on the x-axis). (b) Comparison of residual dipolar coupling (RDC) parameters as obtained in our ensemble (RDC_{calc}, on the y-axis), and from NMR relaxation measurements (RDC_{exp}, on the x-axis). Results for different bond types are shown in different colors. (c) Comparison of 3-bond scalar coupling parameters ${}^3J_{NC_\gamma}$ and ${}^3J_{CC_\gamma}$ as obtained in our ensemble (${}^3J_{calc}$, on the y-axis) and as extracted from NMR relaxation experiments (${}^3J_{exp}$, on the x-axis). (a)-(c) The dashed black line indicates the linear least squares regression fit on the two sets of data, while the continuous line represents the identity line.