Algorithmica manuscript No.
(will be inserted by the editor)

Sampling Conformation Space to Model Equilibrium Fluctuations in Proteins

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

Received: date / Revised version: date

Abstract This paper proposes the Protein Ensemble Method (PEM) to model equilibrium fluctuations in proteins where fragments of the protein polypeptide chain can move independently of one another. PEM models global equilibrium fluctuations of a polypeptide chain by combining local fluctuations of consecutive overlapping fragments of the chain. Local fluctuations are computed by a probabilistic exploration that exploits analogies between proteins and robots. All generated conformations are subjected to energy minimization and then are weighted according to a Boltzmann distribution. Using the theory of statistical mechanics the Boltzmann-weighted fluctuations corresponding to each fragment are combined to obtain fluctuations for the entire protein. The agreement obtained between PEM-modeled fluctuations, wet-lab experiment and guided simulation measurements, indicates that PEM is able to reproduce with high accuracy protein equilibrium fluctuations that occur over a broad range of timescales.

Key words Sampling conformations – equilibrium fluctuations – proteins – robotics – inverse kinematics – statistical mechanics.

1 Introduction

In flexible biomolecules such as proteins, biological function often relates with the ability of a protein to change shape as needed, for instance, to accommodate other molecules for binding [1, 2]. Upon binding, a protein may assume different low-energy conformations [3]. Understanding protein function requires characterizing the entire conformation space available to a protein at equilibrium (under physiological conditions) [4].

⋆ Corresponding author

Experimental techniques such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) report few of the conformations available to a protein at equilibrium. NMR experiments additionally report statistical averages over all equilibrium conformations [5]. While measuring fluctuations that occur over a wide range of timescales, from picoseconds to milliseconds [5], NMR experiments do not provide detailed information on all the equilibrium conformations behind the measurements. Simulation techniques such as Molecular Dynamics (MD) and Monte Carlo [6, 7] sample the conformation space one trajectory at a time and consequently are oftentimes limited to modeling only up to nanosecond equilibrium fluctuations [8–13]. Since events such as binding may occur beyond the nanosecond timescale, this limitation is serious [8–13].

The Protein Ensemble method (PEM) we propose in this work complements existing experimental and simulation techniques. PEM explores the equilibrium conformation space of the whole protein. Unlike existing simulation techniques, PEM does not follow trajectories in conformation space but samples conformations independently of one another.

PEM is based on the premise that, in proteins where fragments of the polypeptide chain do not move in concert with one-another, global equilibrium fluctuations of the polypeptide chain can be obtained by combining local equilibrium fluctuations of fragments of the chain. These fragments are defined consecutively and with overlap by sliding a window over the polypeptide chain. PEM measures equilibrium fluctuations of amino acids of each fragment as Boltzmann-weighted averages over the sampled space of low-energy conformations of each fragment. The theory of statistical mechanics [14] is employed to transform the collection of sampled conformations of each fragment into a Boltzmann ensemble of conformations. PEM exploits analogies between robot kinematic chains and protein polypeptide chains [15, 16] to sample conformations of a fragment similarly to sampling configurations of a kinematic chain.

PEM samples conformations of a fragment through a probabilistic space exploration that is computationally effective because the number of parameters needed to represent the conformation of a fragment of the polypeptide chain is smaller than for the entire chain. PEM then employs an optimization-based inverse kinematics method to map the sampled space to a lower dimensional space of conformations that satisfy the kinematic constraints imposed on the ends of a fragment by the rest of the chain.

The reduced dimensionality of the resulting space makes it computationally feasible to address energetic considerations on the conformations of this space. PEM employs an energy minimization procedure to minimize the energy of each kinematically constrained conformation of a fragment. The minimization procedure interleaves exploring the self-motion manifold of the redundant DOFs of a fragment with a conjugate gradient descent on a pseudo-energy landscape. Each low-energy conformation obtained is weighted by the Boltzmann probability that measures its feasibility at equilibrium. Such weighting allows to measure equilibrium fluctuations of a fragment on the obtained conformations.

The method presented in this work indicates that one computationally effective strategy to model global equilibrium fluctuations of a protein is to combine

local equilibrium fluctuations of consecutive overlapping protein fragments. This strategy is appealing because fluctuations of different fragments can be obtained in parallel. The strategy is well suited for proteins with non-concerted fluctuations, that is, where equilibrium fluctuations of a fragment can be obtained while the rest of the polypeptide chain is unperturbed.

Focusing on proteins with non-concerted motions is however of very broad interest. There is no evidence of any correlation between global physico-chemical properties such as stability or contact order [17] and the nature, local or correlated, of protein fluctuations. Moreover, despite the limited information on protein structures and motions available in current databases [18] and literature, proteins with non-concerted motions represent a significant portion of proteins with known structure [19, 20]. For the proteins studied in this article, our results show that PEM-modeled fluctuations are fully consistent with multiple timescale measurements obtained from NMR wet-lab experiments and guided simulation techniques. Thus, for the examples considered, PEM can be employed to provide a microscopic level of understanding of protein function.

The rest of this article is organized as follows. In Sect. 2 we summarize related work. We devote Sect. 3 to a thorough comparison and discussion of the advantages of PEM over existing simulation techniques. Sect. 3 also provides biophysical background and rationale behind our design of PEM. Details and analysis of PEM are related in Sect. 4. Sect. 5 shows that PEM-obtained fluctuations of the SH3 domain of the Fyn tyrosine kinase (SH3) and α-lactalbumin (α-Lac) agree very well with NMR measurements and guided simulations. In Sect. 6 we lay the ground work for methods that can be employed to assess the accuracy of PEM-obtained fluctuations when experimental or guided simulation data are not available for comparison. We conclude in Sect. 7 with a discussion.

2 Related Work

We first summarize protein modeling work that exploits analogies between protein polypeptide chains and robot kinematic chains. A survey of simulation techniques that model equilibrium fluctuations follows. The provided survey is not meant to be comprehensive but instead focuses on simulation techniques that allow us to place the proposed PEM in context. Since PEM explores the space of kinematically constrained conformations, a discussion of probabilistic space exploration and inverse kinematics methods is also included.

2.1 Background Work in Protein Modeling

A protein molecule consists of repeated blocks of atoms known as amino acids. An amino acid has an alpha carbon (C_α) atom connected to a hydrogen atom, an amino group, a carboxylic group, and a group of atoms known as a sidechain. Consecutive peptide bonds between the amino nitrogen and the carboxylic carbon link amino acids in a polypeptide chain, as shown in Fig. 1(a). Amino acids are numbered from the N- to the C-terminus which refer to the amino and carboxyl

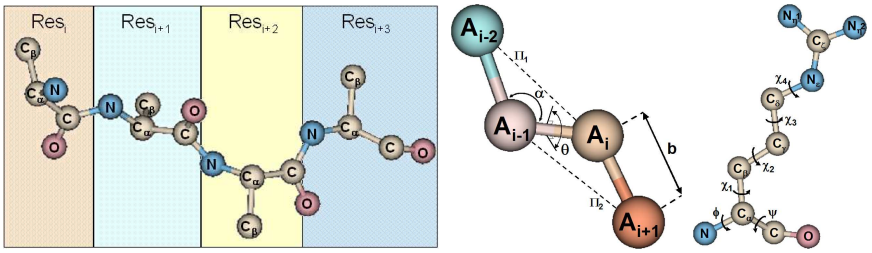


Fig. 1 (a) Polypeptide chain with four amino acids. C_β is the only sidechain atom shown. (b) b refers to the bond length and α to the bond angle. The dihedral angle θ is the angle between planes π_1 defined by the first and second bond and π_2 defined by the second and third bond. Rotation by θ changes positions of atom A_{i+1} and others down the chain. (c) The rigidity of the peptide bond leaves each amino acid with two backbone dihedral angles, ϕ, ψ, for a total of at most 6 dihedral angles. All figures are generated with MOLMOL [21].

not involved in peptide bonds. The backbone is what remains after stripping the sidechains off a polypeptide chain (removing C_β in Fig. 1(a)).

A conformation C that uniquely describes the 3D structure of a protein with N atoms may be represented as a vector $\langle A_{1x}, A_{1y}, A_{1z}, \ldots, A_{Nx}, A_{Ny}, A_{Nz} \rangle$, where A_{ix}, A_{iy}, A_{iz} are atom A_i coordinates. The parameters needed to represent C, in this case the atom coordinates, are often referred to as degrees of freedom (DOFs). We note that most of the $3N$ DOFs in this representation are redundant. Atom positions can be measured using only bond lengths, bond angles, and dihedral angles, which are illustrated in Fig. 1(b). The idealized geometry model [22] allows to fix bond lengths and bond angles to idealized equilibrium values and employ only dihedral angles as DOFs. As illustrated in Fig. 1(c), there are 2 backbone DOFs, the ϕ and ψ dihedral angles, and at most 4 sidechain DOFs per amino acid. On average, the total number of DOFs is $3N/7$ [23].

Geometric Modeling We model a polypeptide chain as a kinematic chain with revolute joints [15, 16] by employing the idealized geometry model. As in forward kinematics [24] where a joint rotation changes positions of following links, rotation about a bond by a dihedral angle changes positions of following atoms. We propagate rotations down a polypeptide chain as in [25] to compute atom positions.

Energetic Modeling We consider atomic interactions through all-atom empirical forcefields such as CHARMM [26] and AMBER [27] which allow to sum over all favorable and unfavorable interactions to calculate the energy of a conformation. State-of-the-art protein folding theories are based on the description of a protein's energy landscape as a multi-dimensional funnel [28–33], the global minimum of which corresponds to a stable conformation assumed by the protein at equilibrium. The equilibrium conformation space of a protein corresponds to the region around the global minimum, which can be populated by the protein upon thermal motions.

2.2 Survey of Simulation Techniques

Current simulation techniques to sample conformation space are either systematic or random searches [13]. MD simulations [4, 13, 34, 35] systematically update atom coordinates of a conformation to obtain a new one by numerically solving Newton's equations of motions. The occurrence of a conformation obtained with a constant temperature MD simulation is proportional to the Boltzmann probability. Since the solution accuracy demands a small timestep in the order of femtoseconds, obtaining a physical trajectory of conformations is computationally demanding [13,36]. Moreover, thoroughly sampling conformation space may require many trajectories. The sampling of rare events such as crossing local maxima of the energy landscape adds to the computational cost of sampling conformation space in a sequential fashion. Thus, in a reasonable amount of time, MD simulations sample a small sub-space of the conformation space available to a protein and are often limited to exploring events that occur within nanoseconds [8–13].

Rather than solving Newton's equations of motions, random search techniques such as Monte Carlo [7, 13] conduct a biased probabilistic walk in conformation space to obtain a sequence of conformations. The biased probabilistic walk ensures through the Metropolis criterion [37] that a conformation is obtained with frequency proportional to its Boltzmann probability. While sometimes computationally more efficient than MD simulations, Monte Carlo simulations also obtain conformations sequentially. Hence they also spend considerable time sampling rare events such as crossing maxima in the energy landscape. Extensions to enhance sampling include methods such as importance [38] and umbrella sampling [39], replica Monte Carlo [40], jump walking [41], multicanonical ensemble [42], entropic sampling [43], weighted histograms [44], local elevation [45], parallel tempering/replica exchange [46], smart walking [47], multicanonical jump walking [48], conformational flooding [49], local energy flattening [50], activation relaxation [51], Markov state models [52], and guided simulation techniques [53–55] that use experimental measurements to guide trajectories to relevant regions of conformation space.

The PEM we propose in this work classifies as a random search that transforms a non-Boltzmann collection of randomly sampled conformations into a Boltzmann ensemble by weighting each conformation with its Boltzmann probability. Rather than obtaining conformations sequentially, PEM probes the energy landscape through a probabilistic exploration that samples conformations independently of one another.

2.3 Probabilistic Space Exploration and Inverse Kinematics Methods

Probabilistic Space Exploration Analogies between robot kinematic chains and protein polypeptide chains [15, 16] allow to use probabilistic space exploration methods to explore the conformation space of a protein [16, 56, 57]. The introduction of the Probabilistic RoadMap (PRM) [58, 59] method in the robotics community enabled the efficient exploration of high-dimensional configuration spaces. In

the context of computational biology, instead of sampling conformations in a sequential fashion, the probabilistic exploration in PRM probes conformation space to sample conformations. Such exploration offers an advantage over combinatorial methods [60, 61] as it allows to efficiently sample large conformation spaces of arbitrarily long polypeptide chains [16, 56, 57].

Probabilistic Space Exploration with Kinematic Constraints Conformations of a polypeptide chain are often kinematically constrained, e.g. by the bond network of a protein [62]. Conformations can first be sampled without considering the kinematic constraints, later enforcing the constraints with a gradient descent [63]. Conformations may be subjected to attractive forces that pull the end effector of the chain, the robot hand or gripper, to its target position and orientation [64]. Maintaining kinematic constraints can be integrated in the sampling process by solving the constraints on 6 DOF sub-chains of sampled conformations [65]. Due to their thorough sampling these methods have an advantage over database methods [66, 67] when applied to protein loops [68].

Inverse Kinematics Methods Satisfying kinematic constraints on the end-effector involves maintaining the end-effector in a specified pose, i.e., a particular position and orientation. Inverse kinematics (IK) [24] asks what DOF values will result in a configuration where the end-effector assumes the target pose. Methods to solve the IK problem can be divided into exact and optimization-based. As far as exact IK methods are concerned, a tight upper bound of 16 solutions has been established for the IK problem for 6R kinematic chains (chains with 6 revolute DOFs) operating in a 3D workspace [69]. Exact IK methods can only enumerate solutions for these chains. One such efficient method [70] has been applied to short molecular chains [15]. Other exact IK methods that deal with molecular chains include [71–75]. IK methods for hyper-redundant kinematic chains are based on curve approximation [76]. In [74] the IK problem is solved for 6 not necessarily consecutive DOFs. The 6-DOF limitation recently extended to 9 DOFs [75]. Currently, only optimization-based IK methods can address the IK problem for kinematic chains with an arbitrary number of DOFs. Methods like random tweak [77] and cyclic coordinate descent (CCD) [78] iteratively solve a system of equations related to kinematic constraints. Unlike random tweak CCD does not compute an inverse or pseudoinverse of a Jacobian matrix and so is computationally stable. Its linear time complexity on the number of DOFs makes it a method of choice for solving the IK problem for polypeptide chains [79–81]. The employment of CCD in the method we propose is inpired by the work in [80, 81].

3 Comparison with Existing Techniques and PEM Biophysical Rationale

The PEM proposed in this work employs a probabilistic space exploration with kinematic constraints. In PEM conformations are obtained independently of one another which allows to model equilibrium fluctuations with no inherent timescale limitations. This is an advantage over existing simulation techniques which, due to their exploration of protein conformation space one trajectory at a time, are limited to modeling equilibrium fluctuations up to the nanosecond timescale [8–13].

A comparison with existing simulation techniques, presented in the following for both accuracy and running time, highlights the advantages of the proposed PEM. The purpose of comparing running times is mainly to illustrate the orders of magnitude difference between PEM and existing simulation techniques since running times of simulation studies are reported in different machines and by different authors.

Applications of PEM to SH3 and α-Lac in this work and to other proteins in [82] show that PEM-obtained fluctuations agree very well with available experimental data over multiple timescales. Pearson correlations no lower than 0.80 are achieved between equilibrium fluctuations obtained by PEM after no more than 164 CPU hours on a current processor and experimental and guided simulation measurements (details on the accuracy of the results obtained by PEM and the running times can be found in Sect. 5). On the other hand, achieving these same high correlations with existing simulation techniques is either possible through simulations that require orders of magnitudes longer CPU time (year-long) [83] or through simulations that shorten their running time to a few months or a few weeks by incorporating experimental data to guide MD or Monte Carlo trajectories to relevant regions of conformation space [53–55].

The high computational time demand of existing simulation techniques limits a direct comparison between these techniques and PEM to a few well-studied proteins. On α-Lac, a protein studied in this article, obtaining equilibrium fluctuations beyond the ns timescale remains challenging for existing simulation techniques. To the best of our knowledge, the only simulation study that overcomes the timescale limitations on α-Lac is a Monte Carlo simulation that resorts to employing a coarse-grained representation of this protein and guides trajectories by incorporating available experimental data [53]. With no a priori knowledge of experimental data and at the same time employing an all-atom representation of protein conformations, in 164 CPU hours PEM obtains equilibrium fluctuations of α-Lac that occur over a wide range of timescales. As detailed in Sect. 5, the ensemble of conformations obtained by PEM for α-Lac agrees very well with the ensemble of conformations obtained in [53].

Another reference protein system for comparisons is ubiquitin, where 6ns of an MD simulation in explicit solvent are needed to obtain a correlation of 0.62 with NMR data [55]. While running times are not reported, our experience estimates that 2ns of simulation time on an AMD Athlon 1900MP machine require one week of CPU time. Longer CPU times are needed to achieve higher accuracy: 80ns, estimated to about one year of CPU time, are needed to obtain a 0.96 correlation with NMR order parameters [83]. The only successful simulation study to our knowledge to obtain good agreement with experimental data (correlation of 0.96) in a few months (22.5ns) guides MD trajectories to relevant regions of conformation space with NMR data [55]. While this result is very significant [84], the required a priori knowledge of high-quality experimental data limits the predictive power of guided simulation techniques. It is noteworthy to emphasize that with no a priori knowledge of experimental data, equilibrium fluctuations obtained by PEM after 120 hours of CPU time on ubiquitin agree with available NMR data with correlations no lower than 0.95 [82].

As the comparisons on α-Lac and ubiquitin illustrate, the demanding computational time of existing simulation techniques makes it hard to replicate simulation studies. On the other hand, PEM's reasonable running time easily allows for replication of the results and applications to many proteins.

It would be a tremendous advance, equivalent to decades of work in biophysics, to be able to model equilibrium fluctuations in any protein. As a first step in this direction, this article sets a very precise goal and focuses on obtaining equilibrium fluctuations in proteins where fluctuations of fragments of the polypeptide chain are uncorrelated. To obtain such fluctuations, PEM employs a first-order approximation that is a powerful algorithmic approach well-rooted in biophysics, particularly in the context of protein folding [85–88]. In protein folding, the enumeration of all configurations of a protein (where each amino acid is considered either in an ordered or a disordered state) is often addressed through a first-order approximation which groups all ordered amino acids on one single continuous stretch of the protein sequence. Considering one single continuous stretch of the protein sequence at a time (or one fragment at a time) is known as the "single sequence approximation." The single sequence approximation was first proposed in the context of the helix-coil theory [85,86] and lately has been shown sufficient in enumerating folding propensities of amino acids of many different proteins [87, 88].

PEM uses the single sequence approximation in a novel context; the method samples conformations of a fragment while the rest of the polypeptide chain is unperturbed in order to obtain detailed atomistic structural information about a protein's conformations at equilibrium. The applicability of the single sequence approximation in this context is justified in proteins where there are no correlated motions between fragments of the polypeptide chain that are far in sequence and where, as a consequence, fluctuations of one fragment can be obtained independently of another. As discussed in detail in Sect. 4, in the absence of correlated fluctuations, PEM constructs fluctuations of the entire polypeptide chain in a multiscale fashion by combining together the fluctuations obtained for fragments covering the chain. An initial treatment of equilibrium fluctuations in the context of even correlated fluctuations through the employment of higher-order approximations is presented in Sect. 6.

4 PEM Algorithm

In Sect. 4.1 we show how PEM defines fragments on a protein polypeptide chain and then combines fluctuations measured over the equilibrium conformation space of each fragment to model global equilibrium fluctuations of a polypeptide chain. In Sect. 4.2 we describe the PEM exploration of the equilibrium conformation space of a fragment. We analyze this exploration in Sect. 4.3.

4.1 Modeling Global Equilibrium Fluctuations Using Local Fluctuations

As shown in pseudocode in Algorithm 1, PEM takes as input an experimentally determined conformation C_{PDB} from the Protein Data Bank [18]. Since C_{PDB}

is an average over protein conformations at equilibrium, PEM initially minimizes the energy of C_{PDB} with a conjugate gradient descent on the energy landscape, detailed in Sect. 4.2.2, to obtain a conformation C_{ref} whose energy E_{ref} is assumed to correspond to the global minimum of the energy landscape. PEM employs C_{ref} as a reference conformation to sample low-energy conformations near the global minimum.

Algorithm 1 PEM (C_{PDB}, l, δl, dl, w)

Input:
C_{PDB}: protein conformation obtained from the PDB
l: length of window sliding over polypeptide chain of protein
δl: overlap between consecutive windows
dl: size of increment to l and δl
w: function to weight fluctuation of an amino acid of a fragment
Output: Equilibrium fluctuations $\langle X_i \rangle$ of each amino acid i

1: $C_{\mathrm{ref}} \leftarrow$ energetically refine C_{PDB}
2: $P \leftarrow$ protein polypeptide chain comprising amino acids 1 to N
3: Slide over P a window of length l amino acids with overlap of δl amino acids between consecutive windows to define fragments $[n_1, n_2]$
4: **for** each fragment $[n_1, n_2]$ **do**
5: $\quad \Omega_{[n_1,n_2]} \leftarrow$ ensemble of sampled low-energy conformations of fragment $[n_1, n_2]$
6: $\quad$ associate $e^{-(E(C)-E_{\mathrm{ref}})/(RT_0)}$ to each $C \in \Omega_{[n_1,n_2]}$ to obtain Boltzmann ensemble
7: $\quad Z \leftarrow \sum_{C \in \Omega_{[n_1,n_2]}} e^{-(E(C)-E_{\mathrm{ref}})/(RT_0)}$ $\quad \triangleright$partition function-normalization factor
8: $\quad \langle X_i \rangle_{[n_1,n_2]} \leftarrow \frac{1}{Z} \sum_{C \in \Omega_{[n_1,n_2]}} e^{-(E(C)-E_{\mathrm{ref}})/(RT_0)}\, X_i(C)$ for amino acid $i \in [n_1, n_2]$
9: **for** each amino acid $i \in P$ **do**
10: $\quad \mathcal{N}_i \leftarrow \sum_{\{\,[n_1,n_2]\,:\, i \in [n_1,n_2]\}} w(i, [n_1, n_2])$ $\quad \triangleright$normalization factor
11: $\quad \langle X_i \rangle \leftarrow \frac{1}{\mathcal{N}_i} \sum_{\{\,[n_1,n_2]\,:\, i \in [n_1,n_2]\}} \langle X_i \rangle_{[n_1,n_2]}\, w(i, [n_1, n_2])$
12: $\quad \{\langle X_i \rangle_{\min}, \langle X_i \rangle_{\max}\} \leftarrow \{\min, \max\}_{\{\,[n_1,n_2]\,:\, i \in [n_1,n_2]\}} \langle X_i \rangle_{[n_1,n_2]}$
13: $\quad$ **if** $\langle X_i \rangle_{\max} - \langle X_i \rangle_{\min} \geq \langle X_i \rangle_{\min}$ **then**
14: $\quad\quad l \leftarrow l + dl$ and $\delta l \leftarrow \delta l + dl$
15: $\quad\quad$ **goto** line 3

4.1.1 Splitting a Polypeptide Chain into Consecutive Overlapping Fragments As shown in line 3 of Algorithm 1, PEM slides a window of length l amino acids over the polypeptide chain P to split the chain into consecutive fragments. The window is slid so that neighboring fragments overlap significantly with one another in $\delta l \approx l$ amino acids (by definition, $\delta l < l$). As illustrated in Fig. 2(a), sliding a window of length 30 with overlap of 25 amino acids defines 19 fragments on the 123 amino acid chain of α-Lac. We denote a fragment encompassing amino acids n_1 to n_2 as $[n_1, n_2]$.

4.1.2 Modeling Local Equilibrium Fluctuations of a Fragment PEM samples low-energy conformations of a fragment $[n_1, n_2]$ while keeping the rest of the polypeptide chain as in C_{ref}. This introduces kinematic constraints on amino acids

n_1 and n_2 of each fragment conformation. Minimizing unfavorable energetic interactions between atoms introduces energetic constrains on fragment conformations.

As shown in line 5 of Algorithm 1, PEM samples the space of kinematically and energetically constrained conformations of each fragment $[n_1, n_2]$ as described in detail in Sect. 4.2 to obtain an ensemble $\Omega_{[n_1,n_2]}$ of low-energy conformations. Since each fragment is shorter than the entire chain P, obtaining $\Omega_{[n_1,n_2]}$ involves exploring a lower dimensional space. Fig. 2(a) shows such ensembles for fragments defined on the polypeptide chain of α-Lac. As shown in line 6 of Algorithm 1, the theory of statistical mechanics [14] is employed to transform the sampled ensemble $\Omega_{[n_1,n_2]}$ into a Boltzmann ensemble of conformations by weighting each conformation C of $\Omega_{[n_1,n_2]}$ with its Boltzmann probability $e^{-(E(C)-E_{\rm ref})/RT_0}$, where $E(C)$ refers to the energy of C, $E(C)-E_{\rm ref}$ to the difference in energy of C from the reference energy $E_{\rm ref}$, R to the gas constant, and T_0 to room temperature (300 K).

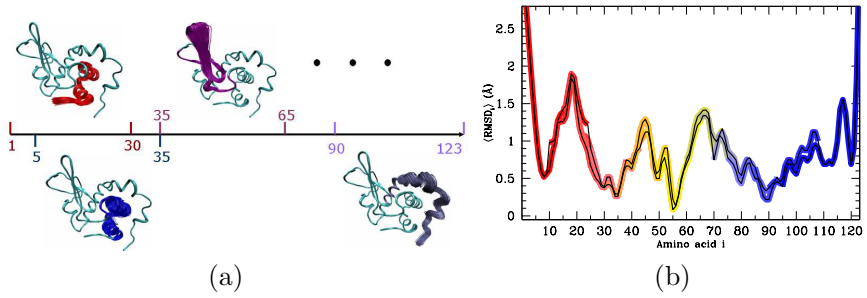
Let $X_i(C)$ be a measurement of the fluctuation of an amino acid i around $C_{\rm ref}$ as witnessed by a conformation C. An example of $X_i(C)$ is the Root-Mean-Squared-Deviation (RMSD) of an amino acid i from $C_{\rm ref}$:

$$\mathrm{RMSD}_i(C) = \sqrt{\frac{1}{\#\,\mathrm{atoms\,in\,i}} \sum_{\mathrm{atom\,j}\in\,\mathrm{i}} \parallel \mathbf{p_j}(C) - \mathbf{p_j}(C_{\rm ref}) \parallel^2},$$

where $\mathbf{p_j}$ is the position of atom j, and $\|\cdot\|$ is the L_2 norm. Other choices for $X_i(C)$ include the deviation from $C_{\rm ref}$ of the orientation of a particular bond vector in amino acid i (order parameters [5], presented in detail in Sect. 5, constitute another choice for $X_i(C)$ employed by PEM). The transformation of $\Omega_{[n_1,n_2]}$ into a Boltzmann ensemble allows to measure a statistical average of $X_i(C)$ over all conformations $C \in \Omega_{[n_1,n_2]}$. As shown in line 8 of Algorithm 1, we sum over all $X_i(C)$, weighting each by the Boltzmann probability of the corresponding conformation $C \in \Omega_{[n_1,n_2]}$, to obtain a Boltzmann-weighted average $\langle X_i \rangle_{[n_1,n_2]}$. This average quantifies the fluctuation of amino acid i as witnessed by the $\Omega_{[n_1,n_2]}$ ensemble of conformations available to fragment $[n_1, n_2]$ at equilibrium.

The average $\langle X_i \rangle_{[n_1,n_2]}$ measured over ensemble $\Omega_{[n_1,n_2]}$ can change with the addition of sampled conformations to the ensemble. To determine a termination condition for sampling, we measure whether, after adding conformations to ensemble $\Omega_{[n_1,n_2]}$, there are any changes in ensemble-averaged measurements such as $\langle \mathrm{RMSD}_i \rangle_{[n_1,n2]}$. When such measurements converge, the sampling of low-energy conformations of fragment $[n_1, n_2]$ terminates as no new information is obtained about equilibrium fluctuations of the fragment.

4.1.3 Measuring Global Equilibrium Fluctuations over Fragment Ensembles
Modeling global equilibrium fluctuations of a polypeptide chain involves quantifying the fluctuation of any amino acid i of the chain as witnessed by the available equilibrium conformation space. As the fluctuation of an amino acid i is a statistical average over the available conformations of the chain, we denote it by $\langle X_i \rangle$. PEM estimates the global fluctuation $\langle X_i \rangle$ of an amino acid i by combining local fluctuations $\langle X_i \rangle_{[n_1,n_2]}$. As shown in line 11 of Algorithm 1, PEM estimates



(a) (b)

Fig. 2 (a) Sliding a window of length 30 and overlap of 25 amino acids on the 123 amino acid chain of α-Lac defines 19 fragments, starting with $[1, 30]$ and ending with $[90, 123]$. An ensemble of low-energy conformations is sampled for each fragment through the exploration detailed in Sect. 4.2. Each ensemble is shown in different colors while the rest of $C_{\rm ref}$ is in cyan. Conformations are drawn with VMD [89]. (b) $\langle {\rm RMSD}_i \rangle_{[n_1,n_2]}$ values, measured as in line 8 of Algorithm 1, are drawn in different colors for different fragments $[n_1, n_2]$. Values for the first and last 5 amino acids of each fragment are discarded. $\langle {\rm RMSD}_i \rangle_{\rm min}$ and $\langle {\rm RMSD}_i \rangle_{\rm max}$, measured as in line 12 of Algorithm 1, are drawn in black.

$\langle X_i \rangle$ as a weighted average over all fluctuations $\langle X_i \rangle_{[n_1,n_2]}$ measured over the sampled ensembles of the fragments $[n_1, n_2]$ overlapping in i. As illustrated for α-Lac in Fig. 2, $\langle X_{19} \rangle$ is averaged over $\langle X_{19} \rangle_{[1,30]}$, $\langle X_{19} \rangle_{[5,35]}$, $\langle X_{19} \rangle_{[10,40]}$, and $\langle X_{19} \rangle_{[15,45]}$. Due to the method employed by PEM to satisfy the kinematic constraints on amino acids n_1 and n_2 of a fragment $[n_1, n_2]$, amino acids i close to n_1 or n_2 do not deviate significantly from their configurations in $C_{\rm ref}$ in the sampled ensemble $\Omega_{[n_1,n_2]}$. Their fluctuations are consequently low and not representative of equilibrium conditions. To take this into account, their contribution to the global equilibrium fluctuation $\langle X_i \rangle$ is downplayed through a weighting function $w(i, [n_1, n_2])$.

An example of a weighting function that downplays fluctuations of the first and last 5 amino acids of each fragment 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. In Fig. 2(b) we show measured $\langle {\rm RMSD}_i \rangle$ for each amino acid i in α-Lac using this weighting function. Fig. 2(b) shows that $\langle {\rm RMSD}_i \rangle_{[n_1,n_2]}$ values measured over ensembles of fragments that encompass amino acid i are similar, as indicated by the small difference between $\langle {\rm RMSD}_i \rangle_{\rm max}$ and $\langle {\rm RMSD}_i \rangle_{\rm min}$, where $\langle {\rm RMSD}_i \rangle_{\rm max}$ and $\langle {\rm RMSD}_i \rangle_{\rm min}$ are measured as shown in line 12 of Algorithm 1. A large difference between $\langle {\rm RMSD}_i \rangle_{\rm min}$ and $\langle {\rm RMSD}_i \rangle_{\rm max}$ would indicate that the length of the window limits fluctuations, in which case, as shown in lines 13-15 of Algorithm 1, window length and overlap are incremented by dl amino acids.

Since the rest of the polypeptide chain remains unperturbed as in $C_{\rm ref}$ while PEM explores equilibrium fluctuations of a fragment, concerted fluctuations between atoms in a fragment and atoms in the rest of the polypeptide chain cannot be captured. Extensions to capture such fluctuations are discussed in Sect. 5.

4.2 Sampling the Equilibrium Conformation Space of a Fragment

PEM models equilibrium fluctuations of a protein polypeptide chain by combining equilibrium fluctuations measured over sampled ensembles of fragments defined consecutively and with overlap over the polypeptide chain. PEM obtains an ensemble $\Omega_{[n_1,n_2]}$ of low-energy conformations of a fragment $[n_1, n_2]$ in two phases. First, the kinematic constraints that the rest of the polypeptide chain imposes on amino acids n_1 and n_2 are exploited to sample a kinematically constrained conformation space as detailed in Sect. 4.2.1. Second, the sampled space is mapped to a sub-space of low-energy conformations as detailed in Sect. 4.2.2.

4.2.1 Probabilistic Exploration with Kinematic Constraints Modeling fluctuations of amino acids of a fragment $[n_1, n_2]$ while keeping the rest of the polypeptide chain as in $C_{\rm ref}$ introduces kinematic constraints on the poses of n_1 and n_2. Keeping the pose of n_1 or n_2 as in $C_{\rm ref}$ involves satisfying 6 constraints: 3 positional constraints for the coordinates of the $\rm C_\alpha$ atom and 3 orientational constraints so that the axes of a local frame at $\rm C_\alpha$ align with the N and C backbone atoms of the amino acid. No constraints are introduced for the sidechain atoms since their positions can change without affecting the rest of the polypeptide chain.

PEM samples kinematically constrained conformations of $[n_1, n_2]$ as shown in pseudocode in Algorithm 2. As shown in line 1 of Algorithm 2, PEM models $[n_1, n_2]$ as a kinematic chain whose base is at n_1 and end-effector at n_2. This analogy allows PEM to sample conformations for $[n_1, n_2]$ through a probabilistic exploration. Conformations are first sampled without considering the constraints. Since exploring different configurations for the sidechains of $[n_1, n_2]$ does not affect the conformation of the rest of the polypeptide chain, sidechains are initially kept in their configurations as in $C_{\rm ref}$. This effectively reduces the dimensionality of the explored conformation space since only the ϕ, ψ dihedral angles starting at amino acid $n_1{+}1$ and ending at $n_2{-}1$ are employed as DOFs. As shown in line 3 of Algorithm 2, values for these DOFs are sampled uniformly at random in $[-\pi, \pi]$.

Rotations by the sampled angles do not change the atom positions of n_1 but violate the kinematic constraint on n_2. Thus each sampled conformation C is subjected to an optimization-based inverse kinematics method, CCD [78], as shown in line 7 of Algorithm 2. We implement CCD as in [79] and apply it to a conformation as in [80, 81] to satisfy the kinematic constraint on n_2. Given a particular permutation of DOFs σ, CCD analytically finds for one DOF at a time the value that minimizes the distance between the poses of n_2 in C and $C_{\rm ref}$. As an iterative method, CCD proceeds in cycles. Each cycle iterates over all DOFs according to the permutation σ of the DOFs until n_2 reaches a pose within an ϵ-neighborhood of the target pose in $C_{\rm ref}$. As shown in line 5 of Algorithm 2, the number of cycles is limited to $n_{\max}$.

Only conformations with no self-collisions between atoms are passed on to the energy minimization procedure. For each conformation values for the sidechain dihedral angles of the fragment are sampled uniformly at random in $[-\pi, \pi]$ until the resulting conformation is free of collisions. Since a high energy indicates the

Algorithm 2 ExploreWithConstraints (C_{ref}, $[n_1, n_2]$, $n_{\max}$, ϵ, σ)

Input:
C_{ref}: conformation corresponding to global minimum of energy landscape
$[n_1, n_2]$: fragment $[n_1, n_2]$ for which to explore conformations
$n_{\max}$: maximum number of CCD cycles
ϵ: criterion for evaluating satisfaction of kinematic constraint on n_2
σ: permutation of DOFs of $[n_1, n_2]$
Output: Conformation C that satisfies kinematic constraint on n_2

$K \leftarrow$ kinematic chain modeling $[n_1, n_2]$
$B \leftarrow$ DOFs of K corresponding to backbone dihedral angles of $[n_1, n_2]$
$\theta|_B \leftarrow$ DOF values sampled uniformly at random in $[-\pi, \pi]^{|B|}$
$C \leftarrow$ apply rotations by $\theta|_B$ dihedral angles
for $n \leftarrow 1$ to $n_{\max}$ **do**
  $B_\sigma \leftarrow$ permutation of DOFs B
  $\bar{C} \leftarrow$ CCD(B_σ, pose of n_2 in C, target pose of n_2 in C_{ref})
  $d \leftarrow$ Euclidean distance between pose of n_2 in $\bar{C}$ and target pose of n_2 in C_{ref}
  if $d \leq \epsilon$ **then**
    exit **for** loop

presence of collisions, as an approximation, a conformation is deemed free of self-collisions if its energy is below a threshold MAX_ENERGY value.

4.2.2 Energy Minimization The minimization procedure interleaves two techniques, an exploration of the self-motion manifold of the redundant DOFs of a fragment with a conjugate gradient descent (the exploration of the self motion manifold is not employed in the minimization of C_{PDB}). To minimize unfavorable interactions these techniques change positions of a fragment's atoms while keeping all other atoms as in C_{ref}, since C_{ref} is already a low-energy conformation.

As local searches, both techniques can yield local minima of the energy landscape. To escape such minima, the minimization procedure interleaves them as follows: If the improvement in energy is less than a cutoff value η after N minimization steps, the procedure determines a local minimum has been reached and switches techniques. The total number of minimization steps is limited to $N_{\max}$. The minimization terminates earlier if the improvement in energy over N steps is less than a convergence value μ. In order to add a resulting conformation C to an ensemble $\Omega_{[n_1,n_2]}$, the difference between its energy $E(C)$ and E_{ref} needs to be small, since the Boltzmann probability of C determines the extent to which the measurement $X_i(C)$ contributes to the ensemble average $\langle X_i \rangle_{[n_1,n_2]}$. A cutoff of 10^{-15} for the Boltzmann probability of a conformation C, for instance, implies that C is added to the ensemble $\Omega_{[n_1,n_2]}$ if its energy $E(C)$ is no higher than 20 kcal/mol from E_{ref}.

Exploring the Self-motion Manifold Due to the 6 constraints on n_2, the remaining $2\,(n_2 - n_1 - 1) - 6$ of the q DOFs of the fragment are redundant. They define a sub-space, the self-motion manifold [90], which we explore for fluctuations of backbone atoms to minimize the energy of a conformation and yet maintain the pose of n_2. We approximate the manifold with its tangent space as in [81] and

obtain an instantaneous change in q of $\dot{q} = J^{\ddagger}(q)\dot{x} + N(q)N^T(q)g(q)$ [90], where $J^{\ddagger}(q)$ is the pseudo-inverse of a $6 \times m$ Jacobian matrix relating the linear and angular velocities of a frame x attached to n_2, $N(q)$ is an orthonormal basis for the null-space, and $g(q)$ is the gradient of the energy function. The constraints on n_2 force $\dot{x} = 0$. Projecting $g(q)$ on the null space yields a motion $\dot{q}$ in dihedral space that minimizes the energy function while keeping n_2 in its pose. We explore the manifold through a steepest descent that at each step updates $J(q)$ as in [91] and $N(q)$ to compute $\dot{q}$. A singular value decomposition (SVD) [92] yields $J(q) = U\Sigma V^T$, where the vectors of V corresponding to zero-valued singular values provide $N(q)$. Due to these computational requirements, we limit the number of steps of the descent.

Conjugate Gradient Descent We design a pseudo-energy function $E = E_{\text{forcefield}} + \sum_{atom\, i \notin \text{fragment}} K_{d_i} \parallel \mathbf{p_i}(C) - \mathbf{p_i}(C_{\text{ref}}) \parallel^2$, where $\mathbf{p_i}$ is the position of atom i. A conjugate gradient descent on this landscape minimizes the energy of a conformation (first term) while limiting fluctuations of atoms outside a fragment (second term). The empirically determined damping constant K_{d_i} allows to maintain crucial interactions between atoms in C during the minimization ($K_{d_i} = 0$ for the minimization of C_{PDB}).

The energy minimization procedure maps a kinematically constrained conformation space to a sub-space of low-energy conformations. Therefore, to assess the PEM exploration of the equilibrium conformation space of a fragment, we investigate the PEM coverage of the space of kinematically constrained conformations.

4.3 Analysis of the PEM Exploration of Conformation Space

Modeling a fragment as a kinematic chain and using CCD allows PEM to map the uniformly sampled space $\mathcal{C}$ of chain configurations to the space $\bar{\mathcal{C}}$ of IK solutions, configurations that satisfy the end-effector kinematic constraints. To sufficiently explore the sub-space of low-energy conformations to which the energy minimization procedure maps $\bar{\mathcal{C}}$, PEM needs to provide a good coverage of $\bar{\mathcal{C}}$.

The solution space $\bar{\mathcal{C}}$ can be described by a system of multi-variable nonlinear polynomial equations that relate the chain DOFs to the end-effector constraints [24]. $\bar{\mathcal{C}}$ may contain components of different dimensions such as isolated solutions, solution curves, and solution surfaces [93]. A notion of coverage of $\bar{\mathcal{C}}$ can be given through that of dispersion [94] which measures the largest portion of $\bar{\mathcal{C}}$ where PEM samples no configurations. A good coverage of $\bar{\mathcal{C}}$ involves minimizing dispersion in each component of $\bar{\mathcal{C}}$. Covering each component uniformly, as provided through the notion of discrepancy [94], might be desirable as well.

The question whether applying CCD to $\mathcal{C}$ provides a good coverage of each component of $\bar{\mathcal{C}}$ remains challenging and open to theoretical analysis. Answering this question is further complicated by the not yet understood dependence of the $\bar{\mathcal{C}}$ exploration on the σ permutation of the chain DOFs employed by CCD. Demonstration of an inadequate coverage of $\bar{\mathcal{C}}$ does not necessarily mean that the exploration of the equilibrium conformation space of a polypeptide chain is insufficient. The reason is that not all components of $\bar{\mathcal{C}}$ may be accessible to a protein.

It has been shown that certain equilibrium conformations may be kinetically inaccessible, i.e., unreachable within biological timescales [95].

In light of these open questions, to provide insight into the coverage of $\bar{\mathcal{C}}$, we analyze experimentally the PEM exploration of $\bar{\mathcal{C}}$ for kinematic chains with increasing number of DOFs. We start with 6R chains where the upper bound of 16 IK solutions [69] allows us to directly compare these solutions to those obtained by PEM when mapping $\mathcal{C}$ with CCD. On kinematic chains with more than 6 DOFs where the dimensionality of $\bar{\mathcal{C}}$ does not allow a direct comparison, we analyze solutions obtained when applying CCD to neighborhoods of configurations in $\mathcal{C}$. To investigate how the σ permutation of DOFs employed by CCD affects the exploration of $\bar{\mathcal{C}}$ we repeat each experiment with three obvious choices for σ; counting from the base to the end-effector (N- to C-terminus) of the chain we define: (i) the random permutation, where the order of DOFs changes randomly in each CCD cycle; (ii) the identity permutation, where the value for DOF i is found before the one for DOF $i+1$; and (iii) the reverse permutation, which refers to the reverse of the identity permutation.

We first determine whether for 6R chains mapping $\mathcal{C}$ with CCD allows to sample all $\bar{\mathcal{C}}$. We do so on a comprehensive list of 20 IK problems for which all IK solutions, obtained with a polynomial continuation method, are documented [96]. For each problem we compare its IK solutions to the solutions sampled by PEM. Two configurations are deemed close if their geodesic distance in $SO(2)^n$ normalized by the number n of DOFs is no more than 0.1 radians. Sampled solutions of a problem with i IK solutions are discretized into i bins, each bin corresponding to an IK solution. A sampled solution goes into a particular bin if it is closest to the IK solution associated with that bin. Solutions are sampled until no bin is empty. We find that for each problem, for each choice of σ, all bins are filled after sampling a maximum of 100 solutions. The maximum distance between a sampled solution and its closest IK solution is no more than 0.02 radians. We conclude that for 6R chains, for each choice of σ, CCD allows to obtain all isolated IK solutions of $\bar{\mathcal{C}}$.

For redundant kinematic chains (more DOFs than constraints) the solution space $\bar{\mathcal{C}}$ is not discrete but can consist of components of different dimensions [93]. Thus we assess the ability of PEM to sample different regions of $\bar{\mathcal{C}}$ for redundant chains by analyzing the solutions obtained when applying CCD to neighborhoods of a configuration A sampled uniformly at random in $\mathcal{C}$. We sample 1000000 neighbor configurations of A uniformly at random from neighborhoods of radii $\{1^\circ, 5^\circ, 10^\circ\}$ deviation per DOF. We apply CCD to A and its neighbor configurations so the end-effector reaches a target pose that we randomly sample in $SE(3)$. We analyze how different the solutions to which CCD maps a neighborhood of A are from configuration B to which A maps under CCD. We do so for 100 instances of configurations A for different choices of permutations σ on chains of 30, 50, and 100 DOFs.

We first measure the probability that CCD maps a neighbor configuration of A to B. Table 1 shows that when employing CCD to increasing perturbations of a random configuration A, the probability of obtaining the IK solution B to which A maps decreases rapidly. In general the probability gets respectively smaller when employing CCD with the identity, reverse, and random permutation. In fact, when

	[30 DOFs]			[50 DOFs]			[100 DOFs]		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
1°	0.6115	7.0175	5.8640	0.2241	7.7586	5.9319	0.2279	6.2142	5.0963
5°	0.0045	0.0988	0.0395	0.0010	0.0036	0.0018	0.0010	0.0010	0.0010
10°	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010

Table 1 A configuration A sampled uniformly at random from $\mathcal{C}$ maps with CCD to an IK solution $B \in \bar{\mathcal{C}}$. Table shows how many (in %) of 1000000 neighbor configurations of A sampled uniformly at random map with CCD to B. Rows show results obtained when neighbor configurations of A are sampled from neighborhoods of radii $\{1^\circ, 5^\circ, 10^\circ\}$. Columns show results obtained for kinematic chains of 30, 50, and 100 DOFs when CCD employs three different choices of the σ permutation of DOFs. (i)-(iii) refer to the random, identity, and reverse permutations, respectively. Results are averaged over 100 instances of A.

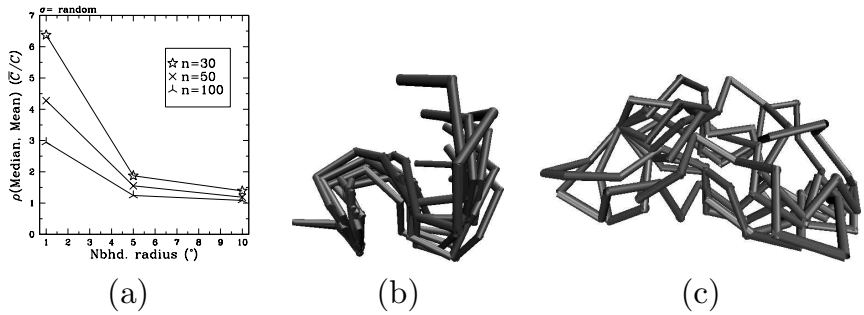


Fig. 3 (a) A distribution of configurations in $\mathcal{C}$ is obtained by sampling uniformly at random 1000000 configurations of kinematic chains of $30, 50, 100$ DOFs from neighborhoods of radii $\{1^\circ, 5^\circ, 10^\circ\}$ around a configuration A sampled uniformly at random in $\mathcal{C}$. Mapping this distribution with CCD yields a distribution of configurations in $\bar{\mathcal{C}}$. For each distribution we measure the distance between the mean and median configurations through ρ, the geodesic distance in $SO(2)^n$ normalized by the number n of DOFs. We plot the ratio of the distance corresponding to the distribution in $\bar{\mathcal{C}}$ over that corresponding to the distribution in $\mathcal{C}$ averaged over 100 instances of A. (b) Configurations of a chain with 12 DOFs are sampled uniformly at random from a 10° radius neighborhood that maps with CCD to the configurations shown in (c). (a)-(c) Results shown are obtained when CCD employs the random permutation of DOFs.

employing CCD with the random permutation on a chain of 50 DOFs, the probability of obtaining the same IK solution B drops quickly to 0.0001% when increasing the perturbation to 10°. The decrease in the probability of obtaining the same solution in $\bar{\mathcal{C}}$ upon increasing neighborhood radii in $\mathcal{C}$ indicates that the sampling of solutions is not limited to a particular region of $\bar{\mathcal{C}}$.

We now compare the obtained distribution of solutions in $\bar{\mathcal{C}}$ with the distribution of the sampled neighbor configurations. For each distribution we measure the distance between the mean and median configurations. The median configuration of the neighborhood of A corresponds to the median distance between A and its neighbor configurations. The median configuration of the obtained distribution

in $\bar{\mathcal{C}}$ corresponds to the median distance between B and obtained solutions in $\bar{\mathcal{C}}$. Fig. 3(a) shows that for each of the chains the distance between the mean and median configurations in the distribution of sampled solutions is persistently larger than in the distribution of neighbor configurations of A. The difference between the two distributions gets smaller as neighbor configurations of A get more diverse with the increase of neighborhood radius and number of DOFs. Fig. 3(b)-(c) illustrates how small perturbations around a configuration A can map to a diverse set of solutions. Similar results are obtained for permutations other than the random. While the observed diversity is not desirable when kinematically constrained chains need to follow a particular trajectory [97], this very feature of CCD allows in this work to obtain different solutions and so explore different regions of $\bar{\mathcal{C}}$.

This analysis shows that CCD allows PEM to explore different configurations of $\bar{\mathcal{C}}$ for redundant chains. While the question whether applying CCD to $\mathcal{C}$ allows to cover all components of $\bar{\mathcal{C}}$ remains, in practice, the PEM exploration of $\bar{\mathcal{C}}$ is sufficient to model equilibrium fluctuations in proteins. As shown in Sect. 5, a good agreement is obtained between $\langle X_i \rangle$ measurements computed over the sub-space of low-energy conformations to which $\bar{\mathcal{C}}$ maps under the energy minimization procedure and measurements provided from experiments or guided simulations.

5 Results

We here assess the PEM exploration of the equilibrium conformation space of a protein polypeptide chain by directly comparing PEM-modeled fluctuations of the chain with experimental and guided simulation measurements of equilibrium fluctuations that occur over a broad range of timescales.

5.1 Implementation Details

In our implementation of CCD $n_{\max}$ is set to 500 and $\epsilon = 0.001$ Å. The `MAX_ENERGY` cutoff is 5000 kcal/mol. In the energy minimization procedure, $N_{\max}$ is set to 1000, $N = 300$, $\mu = 2$ kcal/mol, and $\eta = 20$ kcal/mol. The steepest descent employed to explore the self-motion manifold is limited to 50 steps. The damping constant K_{d_i} in the pseudo-energy function is 10 for all proteins in this work. Initial values for the window length l and overlap δl are set to 20 and 15 amino acids, respectively. If the PEM-obtained fluctuations appear biased by the selection of the window length, as detailed in Sect. 4.1.3, both l and δl are incremented by 5 amino acids. Even though theoretical maximum values for the parameters l and δl can reach the entire chain length N, we recommend $20 \leq l \leq 40$ to maintain accuracy and efficiency. Convergence of PEM-obtained fluctuations for the proteins presented in this work is attained on $l = 30$ and $\delta l = 25$ amino acids. In the exploration of the self-motion manifold we numerically compute $\dot{q}$ through the implementation of finite differences in the OPT++ nonlinear optimization package [98] modeling the energy function as an FDNLF1 object. The conjugate gradient descent is implemented through the OPTCG procedure in the same package modeling the pseudo-energy function as an NLF1 object since its

gradient can be computed analytically. PEM is implemented in ANSI C/C++ using Intel®8.0 compilers and libraries. All experiments were run on the Rice TeraCluster of 900 MHz Intel® Itanium2® processors.

5.2 Modeling Global Equilibrium Fluctuations with PEM

We apply PEM to model equilibrium fluctuations of the SH3 domain of the Fyn tyrosine kinase [99], PDB code 1NYF, and α-Lac [100], PDB code 1HML. PEM obtains around $12,000$ conformations for each of the 6 fragments of the SH3 polypeptide chain in a total of 80 hours of computation time. For α-Lac, PEM obtains around $10,000$ conformations for each of its 19 fragments in a total of 164 hours of computation time. Equilibrium fluctuations of these proteins, measured over the obtained ensembles as described in Sect. 4.1.3, are compared with NMR and guided simulation measurements. For SH3, the comparison shown here is with NMR order parameter (S^2) data, whereas for α-Lac the comparison shown is with RMSD values obtained from a guided simulation. Comparisons of modeled equilibrium fluctuations with other NMR measurements are available for more proteins in [82] (work in [82] also shows that PEM-modeled equilibrium fluctuations are robust against different energy functions such as CHARMM [26] or AMBER [27], weighting functions, permutations of DOFs employed by CCD, and interleaving schemes in the energy minimization procedure).

We choose the comparison with available S^2 data for SH3 because these data quantify the degree of heterogeneity in equilibrium fluctuations of bond vectors over multiple timescales [5]. $S^2=1$ indicates no heterogeneity, whereas $S^2=0$ indicates a uniform distribution over all allowed vectors. Amide S^2 data measure fluctuations that occur within nanoseconds, whereas methyl S^2 data measure fluctuations that may take up to milliseconds [101]. Consequently, while obtaining an ensemble that agrees with NMR methyl S^2 data remains a challenge for simulation techniques [8–13], the comparison with these data of the PEM-obtained ensemble of SH3 is a direct way to assess whether the PEM exploration of conformation space is sufficient to model equilibrium fluctuations with no timescale limitations. We measure S^2 data for a bond as in [54] by averaging over the distribution of vectors observed for the bond in the obtained ensemble.

SH3 is an important protein to understand cancer at a cellular level. Fig. 4(a1) shows the ensemble of conformations obtained with PEM for all the fragments defined on the 56 amino acid polypeptide chain of this protein. In Fig. 4(a2) we compare S^2 data measured over SH3 equilibrium conformations sampled with PEM to NMR S^2 data [102]. The data agree with a Pearson correlation of 0.93. The good agreement, in particular with the NMR methyl S^2 data, indicates that PEM-modeled equilibrium fluctuations are not limited to fast timescales but fully capture SH3 fluctuations observed at equilibrium.

α-Lac is involved in the synthesis of lactose. Fig. 4(a2) shows the ensemble of conformations obtained with PEM for all the fragments defined on the 123 amino acid polypeptide chain of this protein. In Fig. 4(b2) we plot RMSD values obtained with PEM for each amino acid of α-Lac vs. RMSD values measured

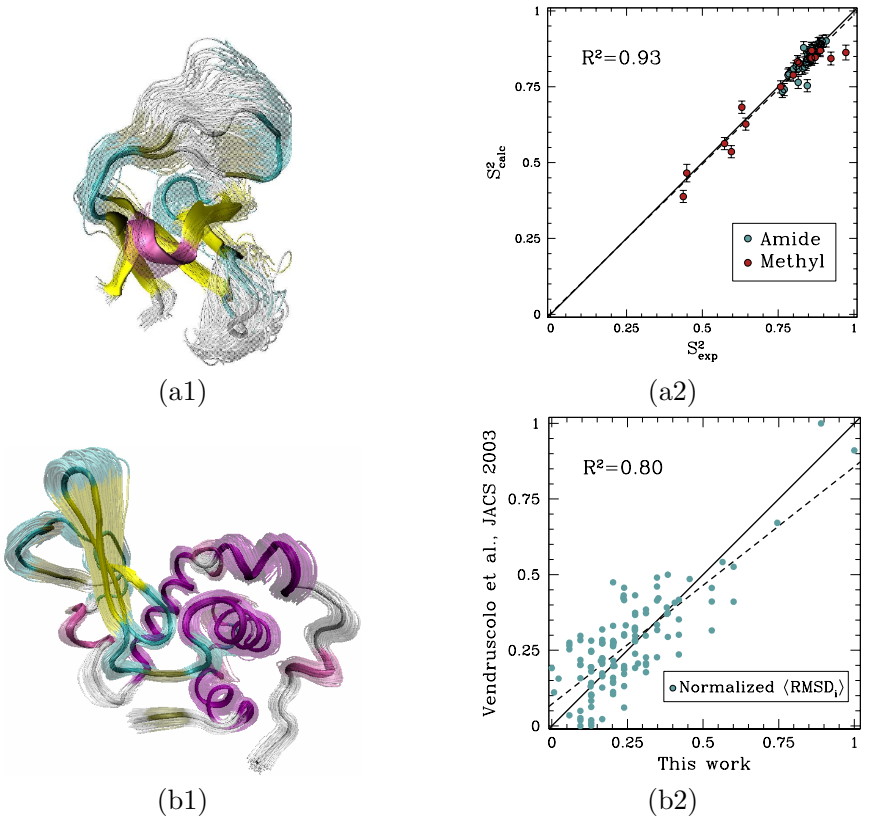


Fig. 4 (a1)-(b1) C_{ref} is shown opaque with PEM-obtained fragment conformations shown transparent for SH3 and α-Lac, respectively. Conformations are drawn with VMD [89]. (a2) Comparison of SH3 PEM-obtained S^2 data (S^2_{calc}) with NMR S^2 data (S^2_{exp}). (b2) Comparison of RMSD values obtained with PEM for each amino acid of α-Lac with RMSD values measured on a Monte-Carlo ensemble [53]. Both data sets are normalized to account for their different magnitudes. (a2)-(b2) The dashed black line indicates the linear least squares regression fit of the data sets, and the continuous line is the identity line.

over an ensemble published in [53]. Since the guided simulation in [53] uses experimental data such as hydrogen exchange protection factors [19], the published ensemble [53] includes fluctuations that occur beyond nanoseconds. The RMSD values agree with a Pearson correlation of 0.80. This correlation indicates that the α-Lac equilibrium fluctuations modeled with PEM include fluctuations that occur at timescales slower than nanoseconds.

6 The Accuracy of Modeled Fluctuations and Higher-order Approximations

PEM is a first-order method that samples conformations of a fragment while the rest of the polypeptide chain is unperturbed. The results presented in Sect. 5 indi-

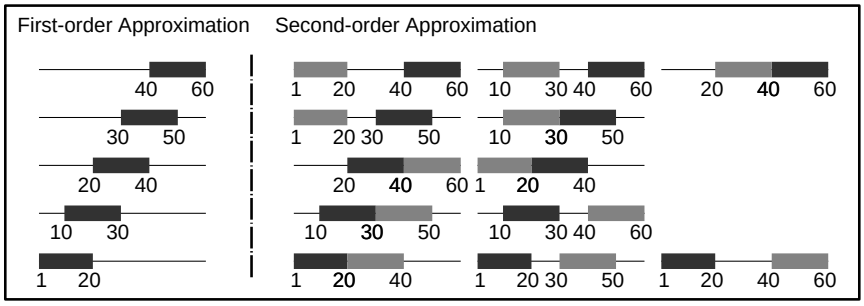


Fig. 5 Under the first-order approximation employed by PEM, shown in the left panel, a window slides over a polypeptide chain. This is illustrated by black windows of length $l = 20$ and overlap $\delta l = 10$ on a polypeptide chain of $N = 60$ amino acids. The second-order approximation is shown on the right panel. All possible ordered pairs of non-intersecting windows with length l and overlap δl are considered. In this case, conformations are first obtained by PEM for the fragments defined by the windows drawn in black. With each so-obtained conformation as initial reference structures, final conformations are then obtained by applying PEM to the fragments defined by the windows drawn in gray.

cate that PEM allows to model with remarkable accuracy equilibrium fluctuations in proteins with non-concerted motions. The applicability of PEM to such proteins is important because proteins with non-concerted motions represent a significant fraction of proteins with known structure [19, 20]. Moreover, the comparison of PEM-modeled fluctuations with experimental data can be used as a framework to test whether local fluctuations are sufficient to explain experimental data.

In the absence of experimental data to which to compare PEM-modeled fluctuations, higher-order approximations are needed to test the accuracy of PEM-obtained measurements independently of experimental data and possibly detect the presence of concerted motions. Since in most proteins concerted motions involve typically no more than two fragments of the polypeptide chain at a time [19], employing a second-order approximation may be sufficient in detecting the presence of correlated fluctuations at equilibrium.

The second-order approximation would involve two windows sliding over a polypeptide chain. All possible ordered pairs of non-intersecting windows can be easily enumerated, as illustrated in Fig. 5 for windows drawn in black and gray. For each ordered pair of non-intersecting windows, conformations with PEM-obtained fluctuations of the fragment defined by the window drawn in black would be used as initial $C_{\rm ref}$ structures to obtain conformations with additional fluctuations modeled with PEM for the fragment defined by the window drawn in gray. Note that in Fig. 5 all possible ordered pairs of non-intersecting windows are considered since the decision of which window to use to obtain initial structures may affect the ensemble of final conformations generated.

While details of the implementation of the second-order approximation go beyond the scope of this work, the pseudocode in Algorithm 3 provides a high-level glimpse on how to obtain equilibrium conformations of a protein when consid-

ering all pairs of fragments that can be defined over a polypeptide chain. The measurement of average quantities $\langle X_i \rangle$ for each amino acid i over the obtained conformations can be addressed similarly as in Sect. 4.1.3. Technical details on the implementation of the second-order approximation are currently under investigation.

Algorithm 3 Second-order Model(C_{ref}, l, δl)

Input:
C_{ref}: reference protein conformation
l: length of window sliding over polypeptide chain of protein
δl: overlap between consecutive windows
Output: Ensemble of low-energy protein conformations Ω

1: $\Omega \leftarrow \emptyset$
2: $P \leftarrow$ protein polypeptide chain comprising amino acids 1 to N
3: Slide over P a window A of length l with overlap of δl to define fragments $[n_1, n_2]$
4: **for** each fragment $[n_1, n_2]$ defined by A **do**
5: slide over P a window B of of length l with overlap of δl to define fragments $[m_1, m_2]$
6: **if** $[n_1, n_2] \cap [m_1, m_2] == \emptyset$ **then**
7: $C_{[n_1,n_2]} \leftarrow$ low-energy conformation of $[n_1, n_2]$ with rest of P fixed as in C_{ref}
8: **for** each fragment $[m_1, m_2]$ defined by B **do**
9: $C_{[n_1,n_2],[m_1,m_2]} \leftarrow$ low-energy conformation of $[m_1, m_2]$ with rest of P fixed as in $C_{[n_1,n_2]}$
10: $\Omega \leftarrow \Omega \cup C_{[n_1,n_2],[m_1,m_2]}$

If the measurements obtained with the second-order approximation agree with those obtained by PEM, then there is strong evidence that the second-order approximation does not significantly change the ensemble of conformations obtained with PEM. That is, the equilibrium fluctuations of the protein under investigation are inherently local and can be modeled by PEM with high accuracy. If however the equilibrium fluctuations generated with PEM do not agree with those obtained with the second-order approximation, then there is evidence that correlated fluctuations may be present. We are currently working to introduce higher-order approximations to model equilibrium fluctuations in a more general framework, that will incorporate also concerted motions.

7 Discussion

We have presented PEM, a novel method that combines a robotics-inspired probabilistic exploration with the theory of statistical mechanics to model protein equilibrium fluctuations. PEM employs the computationally feasible approach of modeling global equilibrium fluctuations of a protein polypeptide chain by combining local equilibrium fluctuations of consecutive overlapping fragments of the chain.

PEM-modeled fluctuations agree very well with NMR and guided simulation measurements of equilibrium fluctuations that span multiple timescales. These results and our analysis of the PEM exploration of the kinematically constrained

conformation space of a fragment indicate that sampling conformations independently of one another allows PEM to sufficiently explore the conformation space available to a protein at equilibrium. Unlike guided simulation techniques, PEM does not use experimental measurements for a sufficient exploration. Thus PEM complements both experimental and simulation techniques as it provides a detailed and extensive view of the equilibrium conformation space available to a protein.

More than 90% of PEM's computation time is spent in the energy minimization procedure. This is due to two reasons. First, the all-atom energy function employed to compute the energy of a conformation is of quadratic complexity in the number of atoms of a protein. Second, the 20 kcal/mol cutoff employed for the energetic difference of an equilibrium conformation from the reference energy and the ruggedness of the energy landscape require a high number of minimization steps. This computation cost begs the need for more efficient energy computations and energy minimization techniques that still maintain the physico-chemical details needed to relate computation and theory with wet-lab experiments.

Currently PEM allows to model only non-concerted equilibrium fluctuations because it samples conformations of a fragment while the rest of the polypeptide chain is unperturbed. As a first-order approximation method, PEM provides a robust starting point to consider higher-order approximations that will allow to capture even concerted motions at equilibrium. We are investigating such higher-order approximation strategies of defining fragments and combining fragment fluctuations so as to model concerted motions in proteins. We are also considering uses of PEM in conjunction with MD and Monte Carlo for a finer sampling of particular regions of the sampled conformation space.

Because proteins with non-concerted equilibrium fluctuations constitute a significant portion of proteins with known structure [19, 20], the proposed PEM is an important first step toward understanding how equilibrium fluctuations affect the ability of a protein to interact with other biomolecules. It is our hope that PEM will become a valuable tool in understanding the microscopic principles that drive macroscopic events such as biological function.

Acknowledgements This work was supported by grants from NSF (CC Career CHE-0349303, LEK ITR-0205671, LEK GM078988, LEK and CC CCF-0523908, and CNS-0454333), the Robert A. Welch Foundation (CC Norman Hackermann Young Investigator award, and C-1570), and the Sloan Foundation (LEK). AS is partly 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.

References

1. J. R. Schnell, H. J. Dyson, and P. E. Wright, Structure, dynamics, and catalytic function of dihydrofolate reductase, Annu. Rev. Biophys. and Biomolec. Struct., 33 (2004), 119–140.

2. L. Sun and Z. J. Chen, The novel functions of ubiquitination in signaling, Curr. Opinion Struct. Biol., 16 (2004), 119–126.
3. G. R. Smith, M. J. E. Sternberng, and P. A. Bates, 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., 347 (2005), 1077–1101.
4. M. Karplus and J. Kuriyan, Molecular dynamics and protein function, Proc. Natl. Acad. Sci. USA, 102 (2005), 6679–6685.
5. L. E. Kay, NMR studies of protein structure and dynamics, J. Magn. Reson., 173 (2005), 193–207.
6. W. F. van Gunsteren and H. J. C. Berendsen, Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry, Angew. Chem. Int., 29 (1990), 992–1023.
7. D. R. Ripoll, J. A. Vila, and H. A. Scheraga, Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH, J. Mol. Biol., 339 (2004), 915–925.
8. V. Daggett. Long timescale simulations, Curr. Opin. Struct. Biol., 10 (2002), 160–164.
9. D. J. Price and C. L. Brooks, Modern protein force fields behave comparably in molecular dynamics simulations, J. Comput. Chem., 23 (2002), 1045–1057.
10. T. Hansson, C. Oostenbrink, and W. F. van Gunsteren, Molecular dynamics simulations, Curr. Opin. Struct. Biol., 12 (2002), 190–196.
11. B. Hess, Convergence of sampling in protein simulations, Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys., 65 (2002), 1–10.
12. K. Tai, Conformational sampling for the impatient, Biophys. Chem., 107 (2004), 213–220.
13. W. F. van Gunsteren, D. Bakowies, R. Baron. I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glattli, P. H. Hunenberger, M. A. Kastenholtz, C. Ostenbrink, M. Schenk, D. Trzesniak, N. F. van der Vegt, and H. B. Yu, Biomolecular modeling: Goals, problems, perspectives, Angew. Chem. Int. Ed. Engl., 45 (2006), 4064–4092.
14. D. Chandler, Introduction to modern statistical mechanics, Oxford University Press, New York, NY, 1987.
15. D. Manocha and Y. Zhu, Kinematic manipulation of molecular chains subject to rigid constraints, in Proceedings of the Second Int Conf Intell Sys Mol Biol (ISMB) (R. B. Altman, D. L. Brutlag, P. D. Karp, R. H. Lathrop, and D. B. Searls, eds.), AAAI, Stanford, CA, 1994, pp. 285–293.
16. A. P. Singh, J. C. Latombe, and D. L. Brutlag, A motion planning approach to flexible ligand binding, in Proceedings of the Seventh Int Conf Intell Sys Mol Biol (ISMB) (R. Schneider, P. Bork, D. L. Brutlag, J. I. Glasgow, H.-W. Mewes, and R. Zimmer, eds.), AAAI, Heidelberg, Germany, 1999, pp. 252–261.
17. K. W. Plaxco, K. T. Simmons, and D. Baker, Contact order, transition state. placement, and the refolding rates of single domain proteins, J. Mol. Biol., 277 (1998), 985–994.
18. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, Nucl. Acids Res., 28 (2000), 235–242.
19. A. R. Fersht, Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding, 3rd edition, W. H. Freeman and Co., New York, NY, 1999.
20. J. Bhalla, G. B. Storchan, C. M. MacCarthy, V. N. Unversky, and O. Tcherkasskaya, Local flexibility in molecular function paradigm, Molecular & Cellular Proteomics, 5 (2006), 1212–1223.
21. R. Koradi, M. Billeter, and K. Wuthrich, MOLMOL - a program for display and analysis of macromolecular structures, J. Mol. Graph., 14 (1996), 51–55, http://www.bruker-biospin.de/NMR/nmrsoftw/prodinfo/molmol/.

22. R. A. Engh and R. Huber, Accurate bond and angle parameters for X-ray protein structure refinement, Acta Crystallogr., A47 (1991), 392–400.
23. R. Abayagan, M. Totrov, and D. Kuznetsov, ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation, J. Comput. Chem., 15 (1994), 488–506.
24. J. Craig, Introduction to robotics: mechanics and control, 2 edition, Addison-Wesley, Boston, MA, 1989.
25. M. Zhang and L. E. Kavraki, A New Method for Fast and Accurate Derivation of Molecular Conformations, J. Chem. Inf. Comput. Sci., 42 (2002), 64–70.
26. 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, J. Comput. Chem., 4 (1983), 187–217.
27. D. C. Wendy, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, J. Am. Chem. Soc., 117 (1994), 5179–5197.
28. N. D. Socci, J. N. Onuchic, and P. G. Wolynes, Protein folding mechanisms and the multidimensional folding funnel, Prot: Struct. Funct. and Genet., 32 (1998), 136–158.
29. J. D. Bryngelson, and P. G. Wolynes, Spin Glasses and the Statistical Mechanics of Protein Folding, Proc. Natl. Acad. Sci., 84 (1987), 7524–7528.
30. P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Navigating the folding routes, Science, 267 (1995), 1619–1620.
31. P. Das, S. Matysiak, and C. Clementi, Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes, Proc. Natl. Acad. Sci., 102 (2005), 10141–10146
32. S. Matysiak and C. Clementi, Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: How far can a minimalist model go?, J. Mol. Biol., 343 (2004), 235–248.
33. S. Matysiak and C. Clementi, Minimalist protein model as a diagnostic tool for misfolding and aggregation, J. Mol. Biol., 363 (2006), 297–308.
34. J. Norberg and L. Nilsson, Advances in biomolecular simulations: methodology and recent applications, Q. Rev. Biophys., 36 (2003), 257–306.
35. S. A. Adcock and J. A. McCammon, Molecular dynamics: Survey of methods for simulating the activity of proteins, Chem. Rev., 106 (2006), 1589–1615.
36. R. Elber, Long-timescale simulation methods, Curr. Opin. Struct. Biol., 15 (2005), 151–156.
37. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys., 21 (1953), 1087–1092.
38. M. H. Kalos, D. Levesque, and L. Valleau, Helium at zero temperature with hard-sphere and other forces, Phys. Rev. A, 9 (1974), 2178–2195.
39. G. M. Torrie and J. P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling, J. Comput. Phys., 23 (1977), 187–199.
40. R. H. Swendsen and J. S. Wang, Replica Monte Carlo simulation of spin-glasses, Phys. Rev. Lett., 57 (1986), 2607–2609.
41. D. D. Frantz, D. L. Freeman, and D. J. D., Reducing quasi-ergodic behavior in Monte Carlo simulations by j-walking, J. Chem. Phys., 93 (1990), 2769–2784.
42. B. A. Berg and T. Neuhaus, Multicanonical ensemble: A new approach to simulate first-order phase transitions, Phys. Rev. Lett., 68 (1992), 9–12.
43. Y. Lee, New Monte Carlo algorithm: entropic sampling, Phys. Rev. Lett., 71 (1993), 211–214.

44. S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules: I. The method, J. Comput. Chem., 13 (1993), 1011–1021.
45. T. Huber, A. E. Torda, and W. F. van Gunsteren, Local elevation: a method for improving the searching properties of molecular dynamics simulation, J. Comput. Aided Mol. Design, 8 (1994), 695–708.
46. U. H. E. Hansmann, Parallel tempering algorithm for conformational studies of biological molecules, Chem. Phys. Lett., 281 (1997), 140–150.
47. R. Zhou and B. J. Berne, Smart walking: a new method for Boltzmann sampling of protein conformations, J. Chem. Phys., 107 (1997), 9185–9196.
48. H. Xu and B. B. J., Multicanonical jump walking: a method for efficiently sampling rough energy landscapes, J. Chem. Phys., 110 (1999), 10299–10306.
49. B. G. Schulze, H. Grubmueller, and J. D. Evanseck, Functional significance of hierarchical tiers in carbonmonoxy myoglobin: conformational substates and transitions studied by conformational flooding simulations, J. Am. Chem. Soc., 122 (2000), 8700–8711.
50. Y. Zhang, D. Kihara, and J. Skolnick, Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding, Proteins: Struct. Funct. Bioinf., 48 (2002), 192–201.
51. R. Malek and N. Mousseau, Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique, Phys. Rev. E, 62 (2000), 7723–7728.
52. N. Singhal, C. D. Snow, and V. S. Pande, Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin, J. Chem. Phys., 121 (2004), 415–425.
53. M. Vendruscolo, E. Pacci, C. Dobson, and M. Karplus, Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange, J. Am. Chem. Soc., 125 (2003), 15686–15687.
54. R. B. Best and M. Vendruscolo, Determination of ensembles of structures consistent with NMR order parameters, J. Am. Chem. Soc., 126 (2004), 8090–8091.
55. K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, and M. Vendruscolo, Simultaneous determination of protein structure and dynamics, Nature, 433 (2005), 128–132.
56. M. S. Apaydin, A. P. Singh, D. L. Brutlag, and J. C. Latombe, Capturing molecular energy landscapes with probabilistic conformational roadmaps, in IEEE Int Conf Robot Autom (ICRA), IEEE, Seoul, Korea, 2001, pp. 932–939.
57. M. S. Apaydin, D. L. Brutlag, C. Guestrin, D. Hsu, J. C. Latombe, and C. Varma. Stochastic Roadmap Simulation: An Efficient Representation and Algorithm for Analyzing Molecular Motion, J. Comp. Biol., 10 (2003), 257–281.
58. L. E. Kavraki, P. Švetska, J.-C. Latombe, and M. Overmars, Probabilistic roadmaps for path planning in high-dimensional configuration spaces, IEEE T. Robotic. Autom., 12 (1996), 566–580.
59. H. Choset, K. M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. E. Kavraki, and S. Thrun, Principles of Robot Motion: Theory, Algorithms, and Implementations, MIT Press, Cambridge, Massachusetts, 2005.
60. M. Dudek and H. J. Scheraga, Protein structure prediction using a combination of sequence homology and global energy minimization. I. Global energy minimization of surface loops, J. Comput. Chem., 11 (1990), 121–151.
61. C. M. Deane and T. L. Blundell, A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins, Proteins: Struct. Funct. Bioinf., 40 (2000), 135–144.

62. D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, Protein flexibility predictions using graph theory, Proteins: Struct. Funct. Bioinf., 44 (2001), 150–165.
63. J. Yakey, S. M. LaValle, and L. E. Kavraki, Randomized path planning for linkages with closed kinematic chains, IEEE T. Robotic. Autom., 17 (2001), 951–959.
64. A. Lee, I. Streinu, and O. Brock, A methodology for efficiently sampling the conformation space of molecular structures, J. Phys. Biol., 2 (2005), 108–S115.
65. L. Han and N. M. Amato, A kinematics-based probabilistic roadmap method for closed chain systems, in Algorithmic and Computational Robotics: New Directions (B. R. Donald, K. M. Lynch, and D. Rus, eds.), AK Peters, Wellesley, MA, 2001, pp. 233–246.
66. S. C. Tossato, E. Bindewald, J. Hesser, and R. Maenner, A divide and conquer approach to fast loop modeling, Protein Eng., 15 (2002), 279–286.
67. P. C. Du, M. Andrec, and R. M. Levy, Have we seen all structures corresponding to short protein fragments in the protein databank? An update., Protein Eng., 16 (2003), 407–414.
68. J. Cortes, T. Simeon, R. de Angulo, D. Guieysse, M. Remaud-Simeon, and V. Tran, A path planning approach for computing large-amplitude motions of flexible molecules, Bioinformatics, 21 (2005), 116–125.
69. E. J. F. Primrose, On the input-output equation of the general 7R- mechanism, Mech. Mach. Theory, 21 (1986), 509–510.
70. D. Manocha and J. Canny, Efficient inverse kinematics for general 6R manipulators, IEEE T. Robotic. Autom., 10 (1994), 648–657.
71. N. Go and H. J. Scheraga, Ring closure and local conformational deformations of chain molecules, Macromolecules, 3 (1970), 178–187.
72. D. Manocha, Y. Zhu, and W. Wright, Conformational analysis of molecular chains using nano-kinematics, Comput. Appl. Biosci., 11 (1995), 71–86.
73. W. J. Wedemeyer and H. J. Scheraga, Exact analytical loop closure in proteins using polynomial equations, J. Comput. Chem., 20 (1999), 819–844.
74. E. Coutsias, C. Seok, C. M. Jacobson, and K. Dill, A kinematic view of loop closure, J. Comput. Chem., 25 (2004), 510–528.
75. M. Zhang, R. A. White, L. Wang, R. Goldman, L. E. Kavraki, and B. Hasset, Improving conformational searches by geometric screening, Bioinformatics, 21 (2005), 624–630.
76. G. S. Chirikjian, General methods for computing hyper-redundant manipulator inverse kinematics, in IEEE/RSJ Int Conf Intell Robot Sys (IROS), IEEE, Yokohama, Japan, 1993, pp. 1067 – 1073.
77. R. M. Fine, H. J. Wang, P. S. Shenkin, D. L. Yarmush, and C. Levinthal, Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations, Proteins: Struct. Funct. Genet., 1 (1986), 342–362.
78. L. T. Wang and C. C. Chen, A combined optimization method for solving the inverse kinematics problem of mechanical manipulators, IEEE T. Robotic. Autom., 7 (1991), 489–499.
79. A. A. Canutescu and R. L. Dunbrack Jr., Cyclic coordinate descent: A robotics algorithm for protein loop closure, Protein Sci., 12 (2003), 963–972.
80. I. Lotan, H. van den Bedem, A. M. Deacon, and J.-C. Latombe, Computing protein structures from electron density maps: the mising loop problem, in Algorithmic Foundations of Robotics VI (M. Erdman, D. Hsu, M. Overmars, and F. van der Stappen, eds.), Springer STAR Series, 2005, pp. 345–360.
81. H. van den Bedem, I. Lotan, J.-C. Latombe, and A. M. Deacon, Real-space protein-model completion: an inverse-kinematics approach, Acta Crystallogr., D61 (2005), 2–13.

82. A. Shehu, C. Clementi, and L. E. Kavraki, Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations, Proteins: Struct. Funct. Bioinf., 65 (2006), 164–179.
83. 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: Struct. Funct. Bioinf., 65 (2006), 712–725.
84. S. Borman, Protein structure wed to dynamics: Technique determines structure and motions of native proteins simultaneously, Chemical And Engineering News, 83 (2005), 12.
85. J. A. Schellman, The factors affecting the stability of hydrogen-bounded polypeptide structures in solution, J. Phys. Chem., 62 (1958), 1485–1494.
86. V. Munoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Folding dynamics and mechanism of beta-hairpin formation, Nature, 390 (1997), 196–199.
87. V. J. Hilser and E. Freire, Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors, J. Mol. Biol., 262 (1996), 756–772.
88. J. O. Wrabl, S. A. Larson, and V. J. Hilser, Thermodynamic propensities of amino acids in the native state ensemble: Implications for fold recognition, Protein Sci., 10 (2001), 1032–1045.
89. W. Humphrey, A. Dalke, and K. Schulten, VMD - Visual Molecular Dynamics, J. Mol. Graph., 14 (1996), 33–38, http://www.ks.uiuc.edu/Research/vmd/.
90. J. W. Burdick, On the inverse kinematics of redundant manipulators: characterization of the self-motion manifolds, in IEEE Int Conf Robot Autom (ICRA), IEEE, Scottsdale, AZ, 1989, pp. 264–270.
91. K. Chang and O. Khatib, Operational space dynamics: efficient algorithms for modeling and control of branching mechanisms, in IEEE Int Conf Robot Autom (ICRA), IEEE, San Francisco, CA, 2000, pp. 850–856.
92. E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, LAPACK Users' Guide, third edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.
93. A. J. Sommese, J. Verschelde, and C. W. Wampler, Advances in polynomial continuation for solving problems in kinematics, J. Mech. Design, 126 (2004), 262–268.
94. H. Choset, K. M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. E. Kavraki, and S. Thrun, Principles of robot motion, 1 edition, MIT Press, Cambridge, MA, 2005.
95. D. Baker and D. A. Agard, Kinetics versus thermodynamics in protein folding, Biochemistry, 33 (1994), 7505–7509.
96. C. W. Wampler and A. Morgan, Solving the 6R inverse position problem using a generic-case solution methodology, Mech. Mach. Theory, 26 (1989), 91–106.
97. O. Brock and O. Khatib, Elastic strips: A framework for motion generation in human environments, Int. J. Robot. Res. (IJRR), 21 (2002), 1031–1052.
98. J. C. Meza, OPT++: An object-oriented class library for nonlinear optimization, Technical Report SAND94-8225, Sandia National Laboratories (1994), http://csmr.ca.sandia.gov/opt++/.
99. C. J. Morton, D. J. Pugh, E. L. Brown, and D. A. Renzoni, Solution structure and peptide binding of the SH3 domain from human Fyn, Structure with Folding and Design, 4 (1996), 705–714.
100. J. Ren, D. I. Stuart, and K. R. Acharya, Alpha-lactalbumin possesses a distinct zinc binding site, J. Biol. Chem., 268 (1993), 19292–19298.
101. D. Ming and R. Brueschweiler, Prediction of methyl-side chain dynamics in proteins, J. Biomol. NMR, 29 (2004), 363–368.

102. A. Mittermaier and L. E. Kay, The response of internal dynamics to hydrophobic core mutations in the SH3 domain from the Fyn tyrosine kinase, Protein Sci., 13 (2004), 1088–1099.