Appeared in Annual Reviews of Analytical Chemistry, Volume 5, p.273-291 (2012)
DOI:10.1146/annurev-anchem-062011-143024

COMPUTATIONAL MODELS OF PROTEIN KINEMATICS AND DYNAMICS: BEYOND SIMULATION

Bryant Gipson(1)
David Hsu(2)
Lydia E. Kavraki(3)
Jean-Claude Latombe(4)

(1) Computer Science Department, Rice University, Houston, TX 77005, USA.
Email: bryant.gipson@rice.edu
(2) Computer Science Department, National University of Singapore, Republic of Singapore.
Email: dyhsu@comp.nus.edu.sg
(3) Computer Science Department and Bioengineering Department, Rice University, Houston, TX 77005, USA.
Email: kavraki@rice.edu
(4) Computer Science Department, Stanford, CA 94305, USA.
Email: latombe@cs.stanford.edu

Table of Contents

Abstract

Physics-based simulation represents a powerful method for investigating the time-varying behavior of dynamic protein systems at high spatial and temporal resolution. Such simulations can be prohibitively difficult or lengthy, however, for large proteins or in probing the lower resolution, long-timescale behaviors of proteins generally. Importantly, not all questions about a protein system require full space and time resolution to produce an informative answer. For instance, by avoiding the simulation of uncorrelated, high-frequency atomic movements, a larger domain-level picture of protein dynamics can be revealed. The purpose of this review is to highlight the growing body of complementary work that goes beyond simulation. In particular, the review focuses on methods that address kinematics and dynamics, as well as on methods that address larger organizational questions and are capable of quickly yielding useful information about the long-timescale behavior of a protein.

Keywords

Protein motion, kino-geometric methods, kinematic models, conformation sampling, graph-based methods

1. INTRODUCTION

Proteins are involved in many biological processes including, to name a few, metabolism, signal transmission, storage of energy, defense against intruders, and muscle buildup. The ability to carry out these functions simultaneously depends on the possible conformational changes of the folded protein and on the dynamics of these deformations. Complete understanding of protein function therefore requires an understanding of the dynamic behavior of a protein, in addition to its static structural features. Physics-based simulations (1-3) offer a direct method to study proteins by describing physical interactions among atoms and numerically solving the associated equations of motion. They constitute a central investigatory tool in molecular and structural biology, allowing analysis in areas that are difficult, expensive or unfeasible to probe experimentally. The purpose of this paper, however, is to highlight the growing body of work that goes beyond simulation. Such methods attempt to quickly answer questions about protein kino-dynamics, as well as larger organizational questions, by generating information about the long-timescale behavior of a protein (4, 5).

Proteins are sequential assemblies of amino acids (a few dozen to several hundred), called residues, joined by peptide bonds and range from hundreds to tens of thousands of atoms in size. Under normal physiological conditions, a protein usually folds into a compact, yet flexible structure. This is referred to as the protein's folded state and is defined by a three-dimensional (3-D) arrangement of secondary structure elements (helices and strands connected by loops). Though this structure is generally not fully rigid, its main features and overall shape are uniquely determined by the protein's amino acid sequence. It is widely accepted that the function of a folded protein is highly dependent on its structure and its ability to deform (6, 7).

For example, in structure-based drug design one must take protein flexibility into consideration in order to correctly predict the interaction between a protein and a potential drug molecule (8, 9). Knowledge of the folded state is also useful for testing energy functions (10), gaining insights into free energy and key determinants of protein stability (11, 12), and modeling structural heterogeneity from NMR, cryo-EM, and X-ray crystallography data (13, 14). The ability to predict the folding motion of a protein of a given sequence also has important potential applications in the design of new proteins (15) and the discovery of cures for neurodegenerative diseases (16). However, for a given protein, only a small number of folded conformations can be determined experimentally.

As of August 2011, the most popular experimental method, X-ray crystallography, has been used to determine 65,195 of the 74,732 protein structures deposited in the Protein DataBank (PDB) (17). This method provides relatively good resolution data and is applicable to large proteins, but requires the creation of a high-quality crystal of the protein of interest, an operation that may not be feasible for some proteins. Additionally, a crystallographic experiment only allows the determination of a single conformation. Software techniques (13, 14, 18, 19) and/or multiple experiments with independently created crystals may produce distinct folded conformations, but these are often produced in too small a number to adequately characterize the flexibility of the folded protein. The next two most widely used experimental methods, NMR spectrometry (9,014 entries in the PDB) and cryo-EM (373 entries), allow the observation of a protein in solution and make it possible to determine several conformations. However, despite recent progress (20), cryo-EM still often produces relatively low-resolution results, yielding ambiguous conformational models, and NMR can only be applied to small proteins.

Computational physics-based methods offer a clear advantage for understanding protein flexibility, as they can characterize dynamic systems and require little prior knowledge. In this context, Molecular Dynamics (MD) simulation models physical interactions among atoms by a potential function and solves Newton's, Lagrange's or Langevin's equations of motion (1). Unfortunately, the solutions to these systems are complicated (6, 21): not only is the potential function made up of many terms, but the equations of motion must also be solved at a time step (on the order of the femtosecond) much shorter than that of atomic fluctuations, in order to reduce cumulative integration errors. MD simulation is thus a computationally intensive process. Modern computers can generate roughly a few nanoseconds of simulation in a day for a medium-size protein—a timescale insufficient for capturing most biologically relevant transitions and events. Distributed computing (22) and specialized architectures (23, 24) speed up MD simulation, with no loss of accuracy, but computational time remains an issue and furthermore the sheer size of data generated becomes a greater hurdle complicating biological insights. One may also achieve faster simulation by using coarser representations (e.g., by grouping atoms together) and approximate or heuristic potentials, but the resulting methods, which include, among others, "coarse-grained" force fields (25), multi-scale modeling (26), improved sampling (27), replica exchange (28), normal mode analysis (29-31), elastic network models (32-34) and Monte Carlo sampling (35, 36), are less accurate, and still produce staggering amounts of data.

Physics-based simulation offers high-resolution spatial and time-dependent information about the conformational neighborhoods of a subset of protein states. While this information is critical to answer some biological questions, structural biologists and bioengineers deal with an increasing diversity of problems that often require computational tools to quickly generate compact, pertinent data that may be obtained from lower-resolution representations of the conformational landscape. For example, a pharmaceutical engineer may want to quickly screen a large database of ligands to identify those which have a reasonable chance to bind to a protein and select "leads" for a new drug. Alternatively, a biologist may want to explore the conformation space of a folded protein to find low-potential conformations, or to simply characterize the range of feasible deformations of a protein. These goals may be better achieved by different methods, in particular, by deliberately avoiding the modeling of fast-frequency motions, which are responsible for the high computational complexity of physics-based simulation methods. Then a compromise is made between accuracy and speed or storage requirements. If higher accuracy is eventually desired, the results of these methods can also be used as a launching point for physics-based simulations. The purpose of this paper is to review non-simulation methods aimed at quickly generating useful information about the long-timescale behavior of a protein. Specifically, the paper consists of three main sections that address the following representation and algorithmic issues:

1. Section 2 reviews a simplified representation of the kinematics of a protein, called the linkage model, which is used by several methods discussed in the other sections of the paper. The linkage model naturally eliminates atomic fluctuations by enforcing distance and angular constraints among covalently bonded atoms. These constraints drastically reduce the number of degrees of freedom (the number of variables required to describe a system) of a protein, which makes it easier for other procedures to explore the

conformation space. However, as atoms can no longer move independently of one another, manipulating this representation raises a challenging question: how can one change atom positions without breaking the constraints? This question is often referred to as the "inverse kinematics" problem, and Section 2 also reviews methods developed to solve it.

2. Section 3 considers conformational sampling, that is: given a set of constraints provided by a kinematic model, how can valid (i.e., biologically relevant) conformations be generated? This question is of fundamental interest because of the tight relationship between protein conformation and function (37-39). Section 3 focuses on the use of geometric constraints in reducing the computational complexity of generating valid protein conformations. It can be shown that such geometric constraints implicitly encode dominant energy terms. Their use therefore produces a two-fold benefit, in that geometric constraints are also present in a favorable format that yields efficient algorithms. Section 3 considers loop sampling, as well as the protein conformational sampling problem generally.

3. Regardless of the method used to find novel protein conformations, a set of valid conformations by itself provides no comparative information about the relationships between protein states. Section 4 describes two broad organizational frameworks designed to answer questions about the collective properties of protein conformation space: probabilistic roadmaps (40), which characterize the local connectivity of a space; and Markov Models (41), which describe probabilistic and long-timescale characteristics of the behavior of a protein. Both methods are complementary and designed to answer large-scale questions concerning "ensemble" properties of proteins (e.g., folding rate, mean first-passage time, and probability of folding, to name a few) without performing explicit physics-based simulation.

2. KINEMATIC MODELING OF A PROTEIN

2.1. Kinematic Linkage Model

A straightforward representation of a protein conformation is a list of the 3-D coordinates of the atom centers in a reference coordinate frame. This representation yields a conformation space of dimensionality 3*n*, where *n* is the number of atoms in the protein. As this representation makes it possible to study protein motion at all timescales, it is not surprising that it is used by most MD simulators.

However, once high frequencies have been smoothed out over picoseconds timescales (42), one may observe that lengths of covalent bonds, angles between adjacent covalent bonds, and dihedral angles around non-rotatable bonds (double, partially double, and peptide bonds) remain almost constant (43). This observation allows one to model a protein's long-term kinetics by a kinematic linkage (44) where—in kinematics terminology (45)—atoms or small groups of atoms are "links" and rotatable bonds are "joints". These joints constitute the degrees of freedom (DOFs) of the model and are typically parameterized by dihedral angles (also called internal coordinates) as depicted in Figure 1a. The resulting model is illustrated in Figure 1b: a kinematic linkage that consists of a long chain—the protein main-chain—in which each residue contributes two DOFs (the so-called ϕ and ψ angles around the N—C_α and C_α—C bonds, respectively) and short side-chains, each with 0 to 6 DOFs (the χ angles). By keeping bond lengths and angles fixed, the linkage model provides a conformational representation that naturally eliminates uncorrelated high-frequency atomic fluctuations and emphasizes "slow" DOFs. In this model, each conformation *c* is defined by the values of the ϕ, ψ and χ angles, and can be seen as a representative of the small region spanned by uncorrelated atomic fluctuations around *c* in the higher-dimensional conformation space parameterized by the 3-D coordinates of all atoms. The dimensionality of the conformation space of the linkage model is upper-bounded by $(2+k)\times p$, where *p* is the number of residues and *k* is the maximum number of χ angles in a side-chain. For most proteins, $(2+k)\times p$ is much smaller than $3n$. Some works have extended MD simulation to the linkage model (46, 47) to reduce the number of variables and increase the integration time step. However, this approach introduces additional computational costs due to complicated intrinsic properties of dihedral angle dynamics.

2.2. Inverse Kinematics Problem

In some respects, however, the linkage model is more difficult to manipulate, as atomic positions can no longer be independently modified. This raises the following inverse kinematics (IK) problem: find conformations of protein fragments that are geometrically consistent with the rest of the main-chain conformation (48).

More formally (44), consider a given conformation *c* of some protein *P*. Let *F* be an inner fragment of *p* consecutive residues in *P*, one can attach two Cartesian coordinate frames Ω_1 and Ω_2 respectively to the N and C termini of *F* (Figure 2a). *F* is said to be in a closed conformation when the pose Π_{cl} (position and orientation) of Ω_2 relative to Ω_1 is fully determined by the conformation *c* of *P*. In general, arbitrary choices of the values of the ϕ and ψ angles in *F* produce poses of Ω_2 relative to Ω_1 that will differ from Π_{cl} (Figure 2b). Conformations of *F* that are not geometrically consistent with the rest of *P* are said to be open. Thus, the IK problem is to determine the values of the ϕ and ψ angles in *F* that result in a closed conformation of *F*.

It is well known from the fields of Kinematics and Robotics (45, 49, 50) that, while the space of all conformations of *F*'s main chain has dimensionality $n = 2p$ (the total number of ϕ and ψ angles in *F*), the subspace $\textit{Closed}(\Pi_{\text{cl}})$ of closed conformations of *F* for a given pose Π_{cl} has dimensionality $n\text{–}6$, except for critical values of Π_{cl} that form a subset of zero measure in the 6-D space $\mathbf{R}^3\times SO(3)$ of all the poses of Ω_2 relative to $\Omega_{1.}$ Here, **R** is the set of the real numbers and SO(3) is the Special Orthogonal Group of 3-D rotations. So, in general, given a pose Π_{cl} of Ω_2 relative to Ω_1, *F* may admit closed conformations only if $n \geq 6$, i.e., if it consists of at least 3 residues. If $n = 6$, the number of IK solutions is finite and varies between 0 and 16 (51-53). If *F* consists of more than 3 residues, the number of IK solutions is in general (i.e., except for critical values of Π_{cl}) either 0 or infinite; in the second case, it is possible to deform the fragment continuously without breaking closure.

2.3. Inverse Kinematics Methods

Analytical IK methods have been proposed for 3-residue fragments in (48, 54). In (54) the problem is reduced to solving a transcendental equation, while the polynomial formulation described in (48) makes it possible to accurately enumerate all possible solutions. The method applies to any fragment of 3 or more residues, in which the ϕ and ψ angles of only 3 (possibly non-consecutive) residues are allowed to vary. Another polynomial formulation is proposed in (55), but the polynomial equations are solved with a subdivision algorithm, which yields approximate solutions. Along a related line of research, the structure of the IK map over $\mathbf{R}^3 \times SO(3)$ is studied in (56), showing that the critical poses of Ω_2 relative to Ω_1 decompose $\mathbf{R}^3 \times SO(3)$ into regular regions, such that over each such region the number of IK solutions is constant. This decomposition leads to a constructive proof of the existence of a region where the theoretical maximum of 16 solutions is attained. This region may not be accessible in practice, however, as it may correspond to high-energy conformations with clashes among side-chains.

When the protein fragment F contains $p > 3$ residues and all ϕ and ψ angles in F are allowed to vary, the IK problem may have an infinite number of solutions and no analytical method is known to compute them. The solutions then span a $(2p-6)$-D space $Closed(\Pi_{cl})$, in which F can deform continuously without breaking closure. Several methods have been proposed to sample conformations in $Closed(\Pi_{cl})$. The RLG method proposed in (57) first picks $p-3$ pairs of ϕ and ψ angles in F at random and then uses an IK method like the one in (48) to determine the remaining 6 angles. However, RLG considers only position accessibility and ignores orientation accessibility, meaning angular values may be selected that do not allow Ω_2 to eventually reach Π_{cl}. By running RLG repeatedly with different values of the $p-3$ pairs of ϕ and ψ angles, one can sample multiple conformations in $Closed(\Pi_{cl})$.

Another approach to sample conformations in $Closed(\Pi_{cl})$ is to use an iterative optimization method. The general idea is to iteratively modify all the ϕ and ψ angles in F in order to reduce the distance between the current pose of Ω_2 and its desired pose Π_{cl}. The popular cyclic coordinate descent (CCD), initially proposed in (58), is applied in (59) by defining the N- and C-anchors as the two fixed residues of the protein that bracket the deforming fragment F on its N- and C-termini, respectively. A fictitious residue M is added at the C-terminus of F. Given any initial conformation of F (picked at random or otherwise), the CCD method iteratively modifies the ϕ and ψ angles in F until M matches the fixed C-anchor. To do this, it minimizes the sum $S = ||N^M N||^2 + ||C_\alpha{}^M C_\alpha||^2 + ||C^M C||^2$, where $||X^M X||$ (X = N, C_α, or C) is the Euclidean distance between the X atom of M and the X atom of the C-anchor. CCD considers each of the ϕ and ψ angles in the fragment in some sequence and resets its value to the one that minimizes S. This value can also be computed analytically (59). CCD iterates until S has been reduced below a small threshold, but convergence is not guaranteed.

2.4. Incorporating Additional Distance Constraints

It is sometimes useful to constrain the linkage model further in order to maintain certain features. For instance, hydrogen bonds (H-bonds) are known to play a key role in both the formation and stabilization of protein structures (60-62). H-bonds involving atoms from residues that are close along the protein main-chain stabilize secondary structure elements, while H-bonds between atoms from distant residues stabilize the protein’s tertiary structure and shape loops and other features that often participate in functional sites. To prevent strong H-bonds from breaking during linkage deformation, one may constrain the linkage model by adding distance equality constraints to the model presented in Section 2.1.

The effect of these constraints is to rigidify atom groups. The method developed in (63-66) derives a distance constraint graph from both the linkage model and the geometry of the H-bonds that must not be broken. The nodes of the graph are the atoms in the protein and each edge represents an equality distance constraint. For instance, a constant angle between two consecutive bonds A—B and B—C leads to an edge between the nodes representing the atoms A and C. An individual H-bond yields three distance constraints. The constraint graph is then processed by a 3-D variant of an algorithm, known as the pebble game (67, 68), to identify all the groups of atoms made rigid by the graph edges. This algorithm is based on the Laman’s theorem initially developed to study the rigidity of planar structures made of bars connected by hinges (69). The result yields a new kinematic linkage model of the protein in which each link is now a rigid group of atoms. Every pair of adjacent links shares exactly two atoms connected by a rotatable covalent bond or an H-bond. Only the dihedral angles around these shared bonds are variable in the new linkage, allowing for less mobility than the original non-constrained linkage. But such a model may contain closed kinematic cycles, up to several dozen, some of which may share dihedral angles. The values of the angles in the cycles can no longer be chosen independently of one another (as will be discussed in Section 3.3).

3. GEOMETRIC CONFORMATION SAMPLING

3.1. Goal

The goal of geometric conformation sampling—as opposed to physics-based sampling, a review of which can be found in (70)—is to explore the range of deformations of a protein (usually a folded one) taking only kinematic and geometric constraints into account. For this, most geometric methods use the kinematic linkage model of Section 2. This model is usually augmented by inequality inter-atomic distance constraints (or volume exclusion constraints) preventing large overlaps (or clashes) between atoms. By modeling each atom as a hard sphere, with van der Waals radii reduced by a multiplication factor of .7 to .8, these distance constraints can be preserved by forbidding any two spheres to overlap. A brute-force algorithm to detect violation of this constraint (by comparing every pair of atoms) runs in time quadratic in the number of atoms. However, the “grid” method analyzed in (71) and used in many implementations only takes linear time. It consists of indexing all atom centers in a 3-D grid of small cubes and only checking pairs of atoms whose centers fall in the same cube or in neighboring cubes. A conformation that satisfies the volume exclusion constraints is said to be clash-free. The attractiveness of a geometric approach derives from the fact that geometric constraints have a favorable format that yields efficient algorithms. They do not require explicit potential functions, which in some cases are difficult to provide (for instance, when a protein may interact with yet unknown molecules). They also make it possible to sample broadly distributed accessible conformations. They do not, however, address the problem of recognizing functional conformations in the generated distribution. If a potential function or structure-based function prediction software (72) is available, sampled conformations may then be filtered in a post-processing phase. Alternatively, results from geometric conformation sampling may serve as the launching point for local physics-based simulation, producing high-resolution time-resolved information from the output of broad low-resolution exploration.

Geometric conformation sampling may apply to an entire protein or, instead, be restricted to a fragment of a protein, typically a flexible loop. In the following, we first consider loop sampling, then protein sampling.

3.2. Loop Conformation Sampling

Loop/fragment conformation sampling has a wide range of applications, for example, to predict deformations that allow ligand binding (73), interpret noisy regions in electron density maps (74), fill gaps in homology modeling (75, 76), create fragment moves in Monte Carlo simulations (77), and tweak main-chain positions for energy optimization (78). Although loop sampling involves relatively few variable dihedral angles, it is still a challenging problem as it requires dealing with two potentially conflicting constraints: a valid loop conformation must both be clash-free and closed (see Section 2.2) in order to be consistent with the rest of the protein (assumed rigid). Basic strategies such as CCD (see Section 2.3) can be employed here, but recent literature offer alternatives tailored to proteins. The loop conformation sampler is mainly characterized by the strategy it uses to achieve these two constraints.

RAPPER (79) iteratively builds up a loop conformation from its N terminus toward its C terminus. At each step, it selects the values of the ϕ and ψ angles in each successive residue at random from a precomputed table of residue-specific values derived from a large collection of diverse protein structures. It also checks that the added residue does not clash with the rest of the protein or the portion of the loop built so far, and that the residue's C_α atom is not further away from the loop's C anchor than a certain threshold that would prevent loop closure. When a complete conformation has been generated, there remains a potentially large gap between the loop's last residue and its anchor on the protein. RAPPER runs an iterative minimization procedure to close this gap, checking volume exclusion at each iteration.

RLG (57) successively samples closed conformations using the RLG IK method reviewed in Section 2.3 and rejects each sampled conformation that is not clash-free. The rejection ratio tends to be high, since clash-free conformations usually span a small subset of the closed conformation space.

The method in (80) and LoopTK (81) decompose a loop into three fragments, independently sample clash-free conformations of the two fragments rooted at the N and C anchors, and close the loop with the middle fragment. LoopTK uses SCWRL3 (82) side chains and includes an efficient method to deform any sampled conformation *c* and generate more conformations around it. This method consists of computing the tangent space of the closed conformation space at *c*, a technique often used in Robotics (83), and moving by small increments in that space. LoopTK has been used to determine loops with up to 25 residues and its combination with a functional site prediction program (72) made it possible to generate and recognize calcium-binding loop conformations.

Finally, it should be noted that some procedures sample loop conformations using libraries of fragments obtained from previously solved structures (84-86). However, they do not check that sampled conformations satisfy the volume-exclusion constraints.

3.3. Protein Conformation Sampling

Sampling entire protein conformations is more complicated than loop sampling, as it involves many more variable dihedral angles. Most methods surveyed below assume that a folded conformation of a protein is given and explore the protein's folded state (or a subset of it) by sampling new conformations obtained by deforming previously sampled conformations (initially, the given folded conformation).

ROCK (for Rigidity Optimized Conformational Kinetics) (66) transforms covalent bonds, H-bonds (with potential energy less than a given threshold) and hydrophobic contacts into equality distance constraints between atoms (see Section 2.4). Using the pebble game algorithm (68) it identifies rigid groups of atoms. The resulting kinematic model of the protein is made of rigid groups connected by variable dihedral angles around rotatable bonds. It usually contains many closed cycles. To sample new conformations, ROCK performs a random walk starting at the given conformation. At each step, it perturbs variable dihedral angles not contained in any cycle at random. It also perturbs at random all variable dihedral angles in each cycle, except 6, which are then solved using an IK procedure. As it closes cycles sequentially, the closure of each cycle results in breaking the previously treated cycles with which it shares variable dihedral angles. Once all cycles have been treated, ROCK uses a minimization procedure to reduce to zero a gap function measuring cycle break-up. Due to conflicting cycle closure constraints, this function can have local minima, hence the minimization process may get trapped into a local minimum. If all cycles are successfully closed, the resulting conformation is checked for atomic clashes.

FRODA (for Framework Rigidity Optimized Dynamic Algorithm) (63, 65) performs the same rigidity analysis as in ROCK. It also performs a random walk but, it differs in the way it samples each new conformation. The positions of all the atoms are first independently perturbed at random. Then iterative optimization is used to fit the relative positions of the atoms in every rigid group *R* back to the geometric template associated with *R*, while avoiding clashes between atoms from different groups. This has the indirect effect of achieving cycle closure. Experiments with FRODA show that each step of the random walk is 100 to 1000 times faster than that of ROCK. However, FRODA's steps may be small, as the process of fitting back atoms to templates often tends to partially cancel out the initial deformation. In addition, the method is not well suited for generating deformations in which large groups of atoms perform correlated moves. The sampling strategies of both ROCK and FRODA can be biased to sample a sequence of conformations between two given protein states and therefore determine pathways between these conformations (65).

KGS (for Kino-Geometric Sampling) (87) performs the same rigidity analysis as ROCK and FRODA, but uses a different sampling strategy and a different method to deform a conformation into a new one. Random walks used by ROCK and FRODA (in their unbiased mode) have an inherently slow diffusion rate and hence are slow to explore a folded state. Instead, KGS uses a diffusive strategy that guides exploration toward less visited space (88). In addition, its deformation method aims at keeping all cycles closed to avoid having to close them back later. It consists of computing the tangent space of the space of conformations where all cycles are closed (83) and moving in that space. This procedure requires non-trivial computations due to the large potential number of interdependent cycles, but allows the sampler to make relatively big deformation steps. In particular, KGS has been able to successfully explore the folded states of Cyanovirin-N, a potent HIV-inactivating protein, and the periplasmatic L-lysine/L-arginine/L-ornithine protein (LAO) (89). Each of these two proteins has two distinct sub-states (PDB ID: 2EZM and 1L5E for Cyanovirin-N, and 2LAO and 1LAF for LAO). Transition from one state to the other involves a hinge and a twist motion between two domains.

The Protein Ensemble Method (PEM) (90) accepts as input a 3-D protein structure, e.g., taken from the PDB. It computes an ensemble of conformations that collectively characterize the mobility of the entire protein at equilibrium. This is done by generating and combining ensembles of conformations for consecutive overlapping fragments (sequences of consecutive amino acids). PEM finds geometrically feasible conformations of each fragment using CCD (see Section 2.3). The approach blends geometric exploration of conformation space with a statistical mechanics formulation to generate an ensemble of physical conformations on which thermodynamic quantities can be measured as ensemble averages. It has been developed for proteins that do not exhibit correlated motion and has been validated on proteins for which ensemble data exists from NMR experiments.

In (91) new conformations are sampled in the context of a Graph-Based model (see Section 4). The procedure starts with a given set of valid conformations (possibly containing only a single conformation) and generates new reasonable conformations by expanding from the original ones. In essence, a tree of conformations is generated with some notion of succession/propagation of one conformation from another. The way propagation, and hence exploration is done, is guided by low-dimension projections of the conformations generated so far. These projections are spatially partitioned into cells and a given projection is selected relative to a weighting scheme that favors larger, less dense cells in order to promote conformation exploration. The conformation associated with this projection then serves as the starting point for expansion of the exploration. The expansion first applies a series of random perturbations, essentially a short random walk, to the known valid conformation and then applies a selection filter (based on energy) to the result. If the resulting conformation is valid, it is added to the set of valid conformations, otherwise it is discarded. In either case the process repeats from the beginning to generate new low-energy conformations and to characterize the energy landscape of the protein.

In the case where a protein structure is not completely known, Rosetta (92) performs a fragment-level construction, using template fragments drawn from libraries of known motifs from homologous and other structures. Conformations resulting from this construction are then optionally post-modified with a Monte Carlo search or other randomized optimization designed to expand the range of the search space. All resulting conformations are then energetically minimized. While computationally intensive, this method has recently been used to produce detailed maps of the energy landscapes of a number of protein domains (93).

Finally, for many of the approaches described above, such as (90, 91), ensuring that conformations are drawn from a representative sampling of the free-energy landscape of a protein system (while avoiding oversampling) is of critical importance, both for good coverage and for speed. Dimension reduction—the approximate low-dimensional representation of high-dimensional systems—can be useful in efficiently guiding several algorithms to representative or unique regions of a conformation space. In (94) the free-energy landscape of DecaAlanine is characterized in 2-D by applying principal component analysis (PCA) directly on dihedral angles under a Cartesian transform. In (95) the free energy landscape of an SH3 domain was characterized in 2- and 3-D (reduced from an original 171), with very low residual error, using non-linear dimension reduction (ScIMaP algorithm). With both methods, conformations with similar features were shown to aggregate into well-separated minima in the lower dimensional representations. Such representations could be used to heuristically guide a sampling scheme as it progresses, while periodically updating results to include newly generated conformations.

4. GRAPH-BASED MODELS OF PROTEIN MOTION

4.1. Introduction

Conformation sampling provides information on the accessible conformation space. But it does not describe conformational changes over time. Here, we review methods that take a set of conformations as input and build a directed graph modeling the long-timescale motion behavior of a protein. The input conformations may have been sampled using geometric or potential-based methods. The nodes of the computed graph represent individual conformations of a protein, or groups of conformations. Its arcs represent transitions between them. The goal is to capture a huge number of possible long-timescale motion paths into a compact and explicit representation that can then be analyzed by efficient computational tools. In particular, graph-based methods make it possible to compute ensemble properties—such as folding rate, mean-first passage time, transition state ensemble, P_{fold} values (96), dominant ordering on secondary structure formation—that characterize protein behavior over a myriad motion paths without performing any explicit simulation.

There has recently been a surge of interest in graph-based models. This trend started with the adaptation of probabilistic roadmaps developed for robot motion planning (40) to represent molecular motion. Then roadmaps evolved into point-based Markov models, and more recently into cell-based and hidden Markov models. We review this line of work below. This review is derived in part from (4).

4.2. Roadmaps

In a classical robot motion planning problem, a robot must move among obstacles without colliding with any of them. A configuration[1] of the robot is said to be valid if the robot at this configuration does not collide with any obstacle. It is usually prohibitively expensive to compute the space of valid configurations of a robot (the robot's valid space), but there exists efficient techniques to check if a given configuration or a given motion path is valid. Probabilistic RoadMap (PRM) planning exploits this observation by computing an approximate representation of the valid space in the form of an undirected graph, the probabilistic roadmap (40). Each node

[1] The word "configuration" for robots has the same meaning as "conformation" for molecules. A configuration of a robot uniquely determines the position of every point on this robot.

of the roadmap corresponds to a valid robot configuration sampled randomly from the robot configuration space, and each edge between two nodes represents a simple valid path between the corresponding configurations (usually, a linear interpolation between them). A PRM planner constructs a roadmap until it connects a start to a goal configuration. Under assumptions that are generally satisfied in practice, the probability that PRM planning finds a motion path between two configurations converges to 1 exponentially in the number of nodes of the roadmap (97). In other words, a probabilistic roadmap provides a good approximation of the connectivity of the space of valid configurations. PRM planning and its variants are currently the most widely used approach to plan the motions of complex articulated robots.

The PRM approach was adapted to model and analyze the motion of a flexible ligand binding with a protein assumed rigid (98). The adaptation relies on an analogy between valid (non-valid) configurations for robots and low-energy (high-energy) conformations for molecules. However, while the configuration space of a robot is cleanly divided between valid and non-valid configurations, the energy landscape over the conformation space of a molecule or a group of molecules does not provide such a clear-cut division. Moreover, while in robotics one is interested in finding one reasonably good motion path, in biology one is interested in characterizing the behavior of a molecule over a representative set of motion paths. To address these differences, the method in (98) proceeds as follows. It attaches a Cartesian frame, *P*, to the protein (assumed rigid) and another one, *L*, to a rigid group of three atoms in the flexible ligand. It defines the conformation of the ligand by 6 parameters representing the position and orientation of *L* relative *P*, plus *p* dihedral angles around the ligand's rotatable bonds. It then samples at random many conformations of the ligand such that the origin of *L* is within some predefined distance from the protein. Each sampled conformation *c* is retained as a node of the roadmap with the following probability distribution:

$$P(c \text{ is retained}) = \begin{cases} 0 & \text{if } E(c) \geq E_{\max} \\ \frac{E_{\max}-E(c)}{E_{\max}-E_{\min}} & \text{if } E_{\min} < E(c) < E_{\max} \\ 1 & \text{if } E(c) \leq E_{\min} \end{cases} \tag{1}$$

where $E(c)$ is the potential energy of the ligand consisting of van der Waals and electrostatic terms, and $E_{\max}$ and $E_{\min}$ are input thresholds. So, the method leads to a greater density of nodes in the low-energy regions of the ligand's conformation space. Next, each node is connected to its k nearest neighbors by a linear-interpolation path. The path between two nodes c and c' is discretized into a sequence of conformations $c_0 = c$, c_1, ..., c_i, ..., $c_s = c'$, such that in any two successive conformations c_i and c_{i+1} no two corresponding atoms are further apart than 1Å. It is accepted only if all the discretized conformations along the path have energy less than a maximum energy threshold. If the path is accepted, the roadmap nodes c and c' are connected to each other by two roadmap arcs of opposite directions. The arc from c to c' is labeled by a weight $w(c \rightarrow c')$ measuring the energetic difficulty of traversing the path from c to c'. For any 3 successive conformations c_{i-1} , c_i and c_{i+1}, with potential values E_{i-1}, E_i, and E_{i+1}, the following equation is used to estimate the probability that the ligand at conformation c_i will move next to c_{i+1}:

$$P(c_i \rightarrow c_{i+1}) = \frac{e^{-(E_{i+1}-E_i)/kT}}{e^{-(E_{i+1}-E_i)/kT} + e^{-(E_{i-1}-E_i)/kT}}$$

where k is the Boltzmann constant and T the absolute temperature. The weight $w(c \rightarrow c')$ is computed as:

$$w(c \rightarrow c') = -\sum_{i=0}^{s-1} \log[P(c_i \rightarrow c_{i+1})].$$

Similarly, the arc from c' to c is labeled by $w(c' \rightarrow c) = -\sum_{i=1}^{s} \log[P(c_{i+1} \rightarrow c_i)]$. So, the roadmap represents a distribution of plausible paths of the ligand through the space surrounding the receptor protein.

Once constructed a roadmap is used in (98) to predict an active binding site from a given collection of potential binding sites, all with low potential energies. This is done by computing the N (where $N \approx 100$) most favorable paths in the roadmap that enter each site from distant conformations and the N most favorable paths that leave each site. It was observed on several protein-ligand complexes that the active binding site is often not the one with the lowest potential energy, but the one for which both the entering and the leaving paths have the highest weights on average. This result suggests the presence of an energy barrier around the active site.

This method was extended in (99) to protein folding, in order to predict the dominant order of secondary structure formation. The protein is modeled using the linkage model of Section 2.1 with fixed χ angles (i.e., rigid residues) and a roadmap is computed by sampling conformations in this model. A key difference with the method of (98) is the sampling strategy. Here, the strategy creates a wavefront of conformations expanding from the given folded conformation. Each new conformation c is obtained by perturbing every ϕ and ψ angle in a previously sampled conformation using a Gaussian distribution. It is retained as a new node of the roadmap with the probability distribution defined in Equation (1), where E is now an energy function that rejects conformations containing collisions among side-chains and favors hydrogen and disulfide bonds in secondary structure elements, as well as hydrophobic interactions. The nodes of the roadmap are sorted into bins based on the number of native contacts, where a native contact is defined as a pair of residues whose C_α atoms are less than 7Å apart in the folded conformation. The sampling strategy fills the bins starting with the bin with all native contacts. Once a bin contains a least a certain number of nodes, sampling is performed around conformations in that bin to fill bins with fewer native contacts. Hence, the density of roadmap nodes over the conformation space is a decreasing function of the distance from the input folded conformation.

The method in (99) then computes the N best paths to the folded conformation from conformations in the zero-native-contact bin. Along each path, the appearance time for a secondary structure element is measured as the mean appearance time for all of its contacts. The predicted secondary structure formation order is the order with the greatest frequency over all paths. The method was tested on a set of 14 proteins ranging from 56 to 110 residues in size. It correctly predicted the order of secondary structure formation in all cases where laboratory data was available.

This work is extended in (100) to analyze proteins for which laboratory experiments show that secondary structures form in different dominant orders. In (101), a new sampling strategy is proposed based on rigidity analysis (see Section 2.4). This strategy scales up better to large proteins than the previous bin-based strategy.

4.3. Point-Based Markov Models

To capture the stochasticity of molecular motion, the roadmap model was transformed in (102) into a Markov model by treating each roadmap node as a state and assigning each arc $c \rightarrow c'$ a transition probability $P(c \rightarrow c')$ derived from the energetic difference between the conformations *c* and *c*' and inspired from the Metropolis criterion. A self-transition is added to each node with probability such that all transition probabilities at this node add up to 1. The resulting graph is treated as Markov model in the following sense: the probability of transitioning from *c* to *c*' is a constant that does not depend on the protein's history before reaching *c*. It is called a point-based Markov model (PMM), as each state represents a single conformation.

In principle, a PMM makes it possible to perform a random walk similar to a Monte-Carlo simulation. However, the most interesting feature of a PMM is that it allows the computation of ensemble properties without performing any explicit simulation or computing any specific path, by using a technique known as first-step analysis. In (102) this technique was used to efficiently compute the P_{fold} value, a theoretical measure on the progress of protein folding (96). Let *F* (resp. *U*) denote the set of nodes that correspond to conformations that are considered folded (resp. unfolded). The value $P_{\text{fold}}(c)$ at any node *c* is the probability that from *c* the protein will reach *F* before *U*. By definition $P_{\text{fold}}(c) = 1$ if $c \in F$ and 0 if $c \in U$. Computing $P_{\text{fold}}(c)$ at each other node using simulation would require performing many runs from *c*. Instead, with first-step analysis, one can write the following equation, which corresponds to performing a single simulation step for many simulation runs all at once:

$$P_{\text{fold}}(c) = \sum_{c' \in F} P(c \rightarrow c') \times 1 + \sum_{c' \in U} P(c \rightarrow c') \times 0 + \sum_{c' \notin F \cup U} P(c \rightarrow c') \times P_{\text{fold}}(c').$$

This leads to a sparse system of linear equations, one for each node not in $F \cup U$. A linear system solver computes the P_{fold} values at all nodes simultaneously. This computation takes *all* paths encoded in the roadmap into account. The method was applied to a monomer of repressor of primer (PDB ID: 1ROP) and engrailed homeo-domain (1HDD). A simplified kinematic model and the H-P energy model were used to create the PMM. It was shown that the P_{fold} values computed with a PMM converge quickly toward the values computed by performing many MC simulation runs, when the number of nodes in the PMM increases. But computation with the PMM is several orders of magnitude faster than computation with MC simulation. The method was later extended to predict experimental measures of folding kinetics, such as folding rates, transition state ensembles, and Φ-values of residues (103).

In (104) an improved sampling method is proposed to generate the nodes of a PMM. The nodes are obtained by sub-sampling conformations of a protein along short trajectories obtained with MD simulation and merging conformations that are close (RMSD-wise) to each other. This approach makes it possible to assign transition durations to the arcs of the model (in addition to transition probabilities). So, it not only provides a more energy-pertinent coverage of the conformation space, but also adds temporal information that potentially allows more accurate computation of dynamic properties. The method was tested the 12-residue tryptophan zipper beta hairpin, which had previously been simulated on Folding@Home (105). The PMM was built by sub-sampling 22,400 conformations along 1,750 independent trajectories. The mean first passage time from the unfolded to the folded state and the folding rate were computed with the resulting model using first-step analysis. Their values agreed well with experimental results from fluorescence and IR.

A method is proposed in (106) to estimate the uncertainty in the set of transition probabilities in a PMM derived from MD simulation runs and to identify the nodes whose arcs have the largest uncertainty. Then one may reduce uncertainty by performing more simulations from these nodes.

4.4. Cell-Based and Hidden Markov Models

All Markov models to represent protein motion depend on a key assumption: the future state of a protein depends on its current state *s* only and not on past history prior to reaching *s*. This assumption enables a Markov model to be compact and yet capture the main features of the underlying dynamics. But single conformations rarely contain enough information to guarantee this assumption. So, a PMM may not have the ability to represent well protein motion over time. One way to alleviate this problem is to construct large PMMs by sampling many nodes, but this makes them more difficult to analyze and understand.

This drawback led to cell-based Markov models (CMMs) (5), in which each node is a collection of sampled conformations that roughly matches an attraction basin (cell) in the protein's energy landscape. The protein interconverts rapidly among different conformations within a basin *s* before it overcomes the energy barrier and transitions to another basin *s′*. The assumption is that after many inter-conversions within *s*, the protein "forgets" the history of how it entered *s* and transitions into *s′* with probability depending on *s* only. MD simulation is used to generate the data for building a CMM (5). Conformations sub-sampled along MD trajectories are first grouped into clusters so that self-transition probabilities for the states in the CMM are maximized, i.e., intra-state transitions are frequent (hence, fast) while inter-state transitions are rare (slow). Recent work builds CMMs at multiple resolutions through hierarchical clustering (107).

Related models, called transition networks, are described in (108, 109). A preliminary form of CMM was proposed earlier to analyze a simplified lattice protein model (110). The data for model construction was obtained by solving the master equation instead of performing MD simulation.

CMMs achieve the dual objectives of better satisfying the Markovian assumption and reducing the number of states. However, they still violate the Markovian assumption in a subtle way. Consider a protein at a conformation *c* near the boundary of an energy basin. The future state of the protein depends not only on *c*, but also on the protein's velocity, hence on past history. By requiring each conformation to belong to a single state, CMMs violate the Markovian assumption, especially near cell boundaries and in cells corresponding to shallow energy basins. To address this problem, in (111) a state is modeled by a probabilistic distribution over the collection of sampled conformations. Each conformation *c* now belongs to all states in the model, but with different probabilities (some very small). Conversely, for each state *s*, the model—a hidden Markov model (HMM)—gives a probability distribution over the conformation space. A major advantage of such an HMM over a CMM is that it can be scored by well-established tools computing its likelihood for a test dataset of MD trajectories. This scoring method makes it possible to determine automatically the optimal number of states. This approach was tested on two extensively studied peptides, alanine dipeptide and the villin headpiece subdomain (HP-35 NleNle), to estimate kinetic and dynamic folding quantities. The results were consistent with available experimental measurements. It was also shown that, although a widely accepted thermodynamic model of alanine dipeptide contains 6 states, a simpler model with only 3 states is almost equally good for predicting long-timescale motions.

Markov models derived from MD data are currently limited by the cost of MD simulation. So far they have been only applied to small proteins. However, faster computers and algorithms should eventually alleviate this limit. It will be possible to generate more data at faster rates, but the resulting datasets will remain difficult to understand, because of the sheer size of the data in high-dimensional spaces. Increasingly, the future challenge will be to gain biological insights from simulation data by deriving simple and yet powerful models. In that respect, CMMs and HMMs are promising possibilities.

5. CONCLUSION

Physics-based simulation is a valuable tool for investigating protein dynamics at high resolution. A host of complementary methods that focus on lower-resolution aspects of a protein's global conformation space have nonetheless shown significant utility in answering many questions of biological importance—with considerable advantages in performance. Further, the two approaches, which focus on differing aspects of a conformational landscape, may be used together to focus on key areas of interest to researchers.

REFERENCES

1. Adcock SA and McCammon JA. 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical Reviews*. 106(5):1589–1615.

2. Day R and Daggett V. 2003. All-atom simulations of protein folding and unfolding. *Advances in Protein Chemistry*. 66:373–403.

3. Scheraga HA, Khalili M, and Liwo A. 2007. Protein-folding dynamics: Overview of molecular simulation techniques. *Ann. Rev. Phys. Chemistry*. 58:57–83.

4. Moll M, Schwarz D, and Kavraki L. 2008. Roadmap Methods for Protein Folding. *Methods in Molecular Biology*. 413:219-239.

5. Chodera JD, Singhal N, Pande VS, Dill KA, and Swope WC. 2007. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chemical Physics*. 126(15):155101-155118.

6. Henzler-Wildman K and Kern D. 2007. Dynamic personalities of proteins. *Nature*. 450(7172):964-72.

7. Anfinsen CB and others. 1973. Principles that govern the folding of protein chains. *Science*. 181(96):223–230.

8. Ahmed A, Kazemi S, and Gohlke H. 2007. Protein flexibility and mobility in structure-based drug design. *Frontiers in Drug Design & Discovery: Structure-Based Drug Design in the 21st Century*. 3(1):455–476.

9. Carlson HA. 2002. Protein flexibility is an important component of structure-based drug discovery. *Current Pharmaceutical Design*. 8(17):1571–1578.

10. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, and Baker D. 2003. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins: Struct., Funct., and Bioinf.* 53(1):76-87.

11. Jacobs DJ. 2010. Ensemble-based methods for describing protein dynamics. *Current Opinion in Pharmacology*. 10:760-769.

12. Vorobjev YN and Hermans J. 2001. Free energies of protein decoys provide insight into determinants of protein stability. *Protein Sci.* 10(12):2498–2506.

13. Van Den Bedem H, Dhanik A, Latombe JC, and Deacon AM. 2009. Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. *Acta Crystallographica Section D: Biological Crystallography*. 65(10):1107–1117.

14. Levin EJ, Kondrashov D a, Wesenberg GE, and Phillips GN. 2007. Ensemble refinement of protein crystal structures: validation and application. *Structure*. 15(9):1040-52.

15. Dantas G, Kuhlman B, Callender D, Wong M, and Baker D. 2003. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Molecular Biology*. 332(2):449–460.

16. Chiti F and Dobson C. 2006. Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry*. 75:333-366.

17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, and Bourne PE. 2000. The protein data bank. *Nucleic Acids Research*. 28(1):235.

18. Burling F and Brunger A. 1994. Thermal motion and conformational disorder in protein crystal structures: Comparison of multi-conformer and time-averaging models. *Israel J. Chemistry*. 34(2):165.

19. Kuriyan J, Osapay K, Burley SK, Brünger AT, Hendrickson WA, and Karplus M. 1991. Probing disorder in high resolution protein structures by simulated annealing. *Proteins: Struct., Funct., and Genetics*. 10:340-358.

20. Baker ML, Zhang J, Ludtke SJ, and Chiu W. 2010. Cryo-EM of macromolecular assemblies at near-atomic resolution. *Nature Protocols*. 5(10):1697-708.

21. Schlick T. 1999. Algorithmic Challenges in Computational Molecular Biophysics. *Journal of Computational Physics*. 151(1):9-48.

22. Kumar S, Huang C, Zheng G, Bohm E, Bhatele A, Phillips JC, Yu H, and Kalé LV. 2008. Scalable Molecular Dynamics with NAMD on Blue Gene/L System. *IBM J. Res. and Dev.* 52:177–188.

23. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, et al. 2007. Anton, a special-purpose machine for molecular dynamics simulation. In *ACM SIGARCH Computer Architecture News*, 35: 1–12.

24. Stone JE, Hardy DJ, Ufimtsev IS, and Schulten K. 2010. GPU-accelerated molecular modeling coming of age. *J. Molecular Graphics and Modelling*. 29(2):116–125.

25. Tozzini V. 2005. Coarse-grained models for proteins. *Current opinion in structural biology*. 15(2):144-50.

26. Sherwood P, Brooks BR, and Sansom MSP. 2008. Multiscale methods for macromolecular simulations. *Current Opinion in Struct. Biology*. 18(5):630–640.

27. Lei H and Duan Y. 2007. Improved sampling methods for molecular simulation. *Current Opinion in Struct. Biology*. 17(2):187–191.

28. Sugita Y and Okamoto Y. 1999. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*. 314(1-2):141–151.

29. Skjaerven L, Hollup SM, and Reuter N. 2009. Normal mode analysis for proteins. *J. Molecular Structure: THEOCHEM*. 898(1-3):42-48.

30. Case DA. 1994. Normal mode analysis of protein dynamics. *Current Opinion in Struct. Biology*. 4(2):285–290.

31. Levitt M, Sander C, and Stern PS. 1985. Protein normal-mode dynamics: Trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Molecular Biology*. 181(3):423–447.

32. Haliloglu T, Bahar I, and Erman B. 1997. Gaussian dynamics of folded proteins. *Physical Review Letters*. 79(16):3090-3093.

33. Schröder GF, Brunger AT, and Levitt M. 2007. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure*. 15(12):1630-1641.

34. Thorpe MF. 2007. Comment on elastic network models and proteins. *Physical Biology*. 4(1):60-63; discussion 64-65.

35. Binder K and Heermann DW. 2010. *Monte Carlo Simulation in Statistical Physics: An Introduction*. 2nd. Springer.

36. Hansmann UHE and Okamoto Y. 1999. New Monte Carlo algorithms for protein folding. *Current Opinion in Struct. Biology*. 9(2):177–183.

37. Csermely P, Palotai R, and Nussinov R. 2010. Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends in Biochemical Sciences*. 35(10):539–546.

38. G.G. Hammes Y.C. Chang and Oas TG. 2009. Conformational selection or induced fit: A flux description of reaction mechanism. *Proc. Nat. Acad. of Sc.* 106(33):13737.

39. Zhou H-X. 2010. From induced fit to conformational selection: A continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophysical J.* 98(6):L15.

40. Kavraki LE, Svestka P, Latombe JC, and Overmars MH. 1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Tr. on Robotics and Automation*. 12(4):566–580.

41. Krogh A, Brown M, Mian IS, Sjölander K, and Haussler D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Molecular Biology*. 235(5):1501-31.

42. Callender RH, Dyer RB, Gilmanshin R, and Woodruff WH. 1998. Fast events in protein folding: The time evolution of primary processes. *Ann. Rev. Phys. Chemistry*. 49:173-202.

43. Brown ID. 2009. Recent developments in the methods and applications of the bond valence model. *Chemical Reviews*. 109(12):6858-6919.

44. Zhang M and Kavraki LE. 2002. A new method for fast and accurate derivation of molecular conformations. *J. Chemical Information and Computer Sciences*. 42(1):64–70.

45. Hartenberg RS and Denavit J. 1964. *Kinematic Synthesis of Linkages*. 1964. McGraw-Hill New York.

46. Chen J, Im W, and Brooks CL. 2005. Application of torsion angle molecular dynamics for efficient sampling of protein conformations. *J. Comp. Chem.* 26(15):1565-1578.

47. Gibson KD and Scheraga HA. 1990. Variable step molecular dynamics: An exploratory technique for peptides with fixed geometry. *J. Comp. Chem.* 11(4):468–486.

48. Coutsias EA, Seok C, Jacobson MP, and Dill KA. 2004. A kinematic view of loop closure. *J. Comp. Chem.* 25(4):510–528.

49. Craig JJ. 1989. *Introduction to Robotics: Mechanics and Control*. 2nd. Addison-Wesley New York.

50. Duffy J. 1980. *Analysis of Mechanisms and Robot Manipulators*. 1980. Edward Arnold.

51. Manocha D, Zhu Y, and Wright W. 1995. Conformational analysis of molecular chains using nano-kinematics. *Bioinformatics (Oxford, England)*. 11(1):71-86.

52. Mavroidis C and Roth B. 1994. Structural Parameters which reduce the number of manipulator configurations. *J. Mechanical Design*. 116(1):3-11.

53. Raghavan M and Roth B. 1993. Inverse kinematics of the general 6R manipulator and related linkages. *J. Mechanical Design*. 115(3):502-508.

54. Go N and Scheraga HA. 1970. Ring closure and local conformational deformations of chain molecules. *Macromolecules*. 3:178-187.

55. Zhang M, White RA, Wang L, Goldman R, Kavraki L, and Hassett B. 2005. Improving conformational searches by geometric screening. *Bioinformatics (Oxford, England)*. 21(5):624-630.

56. Milgram RJ, Liu G, and Latombe JC. 2008. On the structure of the inverse kinematics map of a fragment of protein backbone. *J. Comp. Chem.* 29(1):50-68.

57. Cortés J, Siméon T, Remaud-Siméon M, and Tran V. 2004. Geometric algorithms for the conformational analysis of long protein loops. *J. Comp. Chem.* 25(7):956-967.

58. Wang L-CT and Chen CC. 1991. A combined optimization method for solving the inverse kinematics problems of mechanical manipulators. *IEEE Tr. on Robotics and Automation*. 7(4):489-499.

59. Canutescu AA and Dunbrack RL. 2003. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* 12(5):963-972.

60. Fersht A and Serrano L. 1993. Principles of protein stability derived from protein engineering experiments. *Current Opinion in Struct. Biology*. 3:75-83.

61. Pace C. 2001. Polar group burial contributes more to protein stability than nonpolar group burial. *Biochemistry*. 40:310-313.

62. Schell D, Tsai J, Scholtz J, and C. Nick Pace. 2006. Hydrogen bonding increases packing density in the protein interior. *Proteins: Struct., Funct., and Bioinf.* 63(2):278-282.

63. Farrell D, Speranskiy K, and Thorpe MF. 2010. Generating stereochemically acceptable protein pathways. *Proteins: Struct., Funct., and Bioinf.* 78:2908-2921.

64. Jacobs DJ, Kuhn LA, and Thorpe MF. 2002. Flexible and rigid regions in proteins. *Rigidity Theory and Applications*:357–384.

65. Wells S, Menor S, Hespenheide B, and Thorpe MF. 2005. Constrained geometric simulation of diffusive motion in proteins. *Physical Biology*. 2(4):S127-S136.

66. Zavodszky M, Lei M, Thorpe M, Day A, and Kuhn LA. 2004. Modeling correlated main-chain motions in proteins for flexible molecular recognition. *Proteins: Struct., Funct., and Bioinf.* 57(2):243-261.

67. Jacobs DJ. 1998. Generic rigidity in three-dimensional bond-bending networks. *J. Physics A: Mathematical and General*. 31:6653.

68. Lee A, Streinu I, and Theran L. 2008. *Analyzing rigidity with pebble games*. 226. ACM Press.

69. Laman G. 1970. On graphs and rigidity of plane skeletal structures. *J. Engineering Mathematics*. 4(4):331–340.

70. Liwo A, Czaplewski C, Oldziej S, and Scheraga HA. 2008. Computational techniques for efficient conformational sampling of proteins. *Current Opinion in Struct. Biology*. 18(2):134–139.

71. Halperin D and Overmars MH. 1994. Spheres, molecules, and hidden surface removal. In *Proc. of the Tenth Ann. ACM Symp. Computational Geometry*, pp. 113–122.

72. Wu S, Liang MP, and Altman RB. 2008. The SeqFEATURE library of 3D functional site models: Comparison to existing methods and applications to protein function annotation. *Genome Biology*. 9(1):R8.

73. Sousa SF, Fernandes PA, and Ramos MJ. 2006. Protein–ligand docking: Current status and future challenges. *Proteins: Struct., Funct., and Bioinf.* 65(1):15–26.

74. van den Bedem H, Lotan I, Latombe JC, and Deacon A. 2005. Real-space protein-model completion: An inverse-kinematic approach. *Acta Crystallographica Section D*. D61:2-13.

75. Enosh A, Fleishman SJ, Ben-Tal N, and Halperin D. 2004. Assigning transmembrane segments to helices in intermediate-resolution structures. *Bioinformatics (Oxford, England)*. 20(Suppl 1):i122-i129.

76. Xiang Z. 2006. Advances in homology protein structure modeling. *Current Protein & Peptide Science*. 7(3):217.

77. Cahill S, Cahill M, and Cahill K. 2003. On the kinematics of protein folding. *J. Comp. Chem.* 24(11):1364–1370.

78. Singh R and Berger B. 2005. ChainTweak: Sampling from the neighbourhood of a protein conformation. In *Proc. Pacific. Symp. on Biocomputing*, pp. 54–65.

79. DePristo MA, de Bakker PIW, Lovell SC, and Blundell TL. 2003. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins: Struct., Funct., and Bioinf.* 51(1):41–55.

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

81. Yao P, Dhanik A, Marz N, Propper R, Kou C, Liu G, van den Bedem H, Latombe JC, Halperin-Landsberg I, and Altman RB. 2008. Efficient algorithms to explore conformation spaces of flexible protein loops. *IEEE/ACM Tr. on Comp. Biology and Bioinf.* 5(4):534-45.

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

83. Siciliano B and Khatib O. 2008. *Handbook of Robotics*. Springer.

84. Kolodny R, Guibas L, Levitt M, and Koehl P. 2005. Inverse kinematics in biology: The protein loop closure problem. *Int. J. Robotics Research*. 24(2-3):151.

85. Tosatto SCE, Bindewald E, Hesser J, and Manner R. 2002. A divide and conquer approach to fast loop modeling. *Protein Engineering Design and Selection*. 15(4):279-286.

86. van Vlijmen HWT and Karplus M. 1997. PDB-based protein loop prediction: parameters for selection and methods for optimization1. *J. Molecular Biology*. 267(4):975–1001.

87. P. Yao, Zhang L, and Latombe JC. 2011. Sampling-based exploration of folded state of a protein under kinematic and geometric constraints. *Proteins: Struct., Funct., and Bioinf.* DOI:10.1002/prot.23134.

88. Hsu D, Latombe JC, and Motwani R. 1999. Path Planning in Expansive Configuration Spaces. *International Journal of Computational Geometry and Applications (IJCGA)*. 9(4-5):495-512.

89. Silva D-A, Bowman GR, Sosa-Peinado A, and Huang X. 2011. A role for both conformational selection and induced fit in ligand binding by the LAO protein. *PLoS Computational Biology*. 7(5):e1002054.

90. Shehu A, Clementi C, and Kavraki LE. 2006. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins: Struct., Funct., and Bioinf.* 65(1):164–79.

91. Haspel N, Moll M, Baker ML, Chiu W, and Kavraki LE. 2010. Tracing conformational changes in proteins. *BMC Structural Biology*. 10 Suppl 1(May):S1.

92. Rohl CA, Strauss CEM, Misura KMS, and Baker D. 2004. Protein structure prediction using Rosetta. *Methods in Enzymology*. 383:66-93.

93. Tyka MD, Keedy DA, André I, Dimaio F, Song Y, Richardson DC, Richardson JS, and Baker D. 2011. Alternate states of proteins revealed by detailed energy landscape mapping. *J. Molecular Biology*. 405(2):607-18.

94. Altis A, Nguyen PH, Hegger R, and Stock G. 2007. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chemical Physics*. 126(24):244111.

95. Das P, Moll M, Stamati H, Kavraki LE, and Clementi C. 2006. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proc. Nat. Acad. of Sc.* 103(26):9885-9890.

96. Du R, Pande VS, Grosberg A, Tanaka T, and Shakhnovich ES. 1998. On the transition coordinate for protein folding. *J. Chemical Physics*. 108(1):334-351.

97. Hsu D, Latombe JC, and Kurniawati H. 2006. On the probabilistic foundations of probabilistic roadmap planning. *Int. J. Robotics Research*. 25(7):627-643.

98. Singh AP, Latombe JC, and Brutlag DL. 1999. A motion planning approach to flexible ligand binding. *Proc. Int. Conf. Intelligent Syst. for Molecular Biology (ISMB)*:252-261.

99. Amato NM, Dill KA, and Song G. 2003. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comp. Biology*. 10(3-4):239–255.

100. Thomas S, Song G, and Amato NM. 2005. Protein folding by motion planning. *Physical Biology*. 2:S148.

101. Thomas S, Tang X, Tapia L, and Amato NM. 2007. Simulating protein motions with rigidity analysis. *J. Comp. Biology*. 14(6):839-855.

102. Apaydin MS, Brutlag DL, Guestrin C, Hsu D, Latombe JC, and Varma C. 2003. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. *J. Comp. Biology*. 10:257-281.

103. Chiang T-H, Apaydin MS, Brutlag DL, Hsu D, and Latombe JC. 2007. Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics: Folding rates and phi-values. *J. Comp. Biology*. 14(5):578-593.

104. Singhal N, Snow C, and Pande VS. 2004. Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chemical Physics*:121:415-425.

105. Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, Larson SM, Rhee YM, et al. 2003. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*. 68(1):91-109.

106. Singhal N and Pande VS. 2005. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chemical Physics*. 123(20):204909-204912.

107. Huang X, Yao Y, Bowman GR, Sun J, Guibas LJ, Carlsson G, and Pande VS. 2010. Constructing multi-resolution markov state models (MSMs) to elucidate RNA hairpin folding mechanisms. *Proc. Pacific. Symp. on Biocomputing*:228-239.

108. Noé F and Fischer S. 2008. Transition networks for modeling the kinetics of conformational change in macromolecules. *Current Opinion in Struct. Biology*. 18(2):154-162.

109. Noé F, Krachtus D, Smith JC, and Fischer S. 2006. Transition networks for the comprehensive characterization of complex conformational change in proteins. *J. Chemical Theory and Computation*. 2(3):840-857.

110. Ozkan S, Dill K, and Bahar I. 2002. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.* 11(8):1958-1970.

111. Chiang TH, Hsu D, and Latombe JC. 2010. Markov dynamic models for long-timescale protein motion. *Bioinformatics*. 26(12):i269-i277.

FIGURES

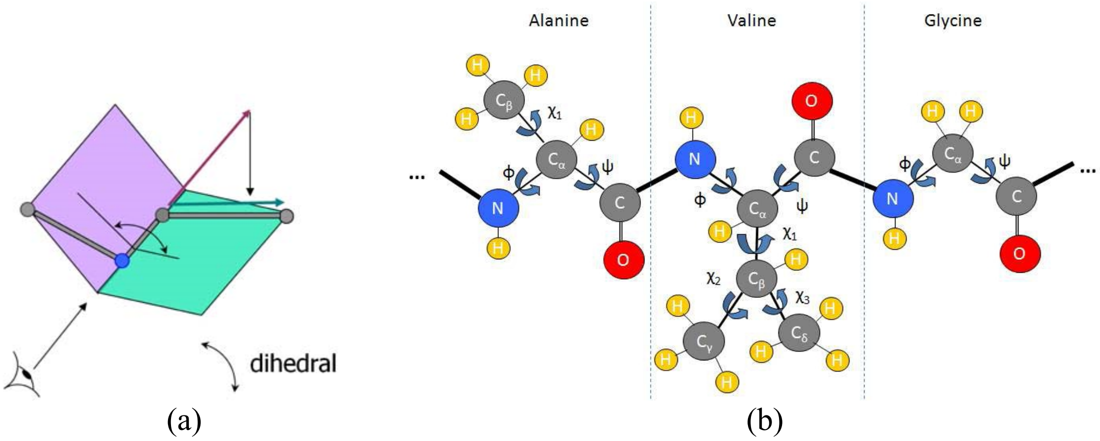
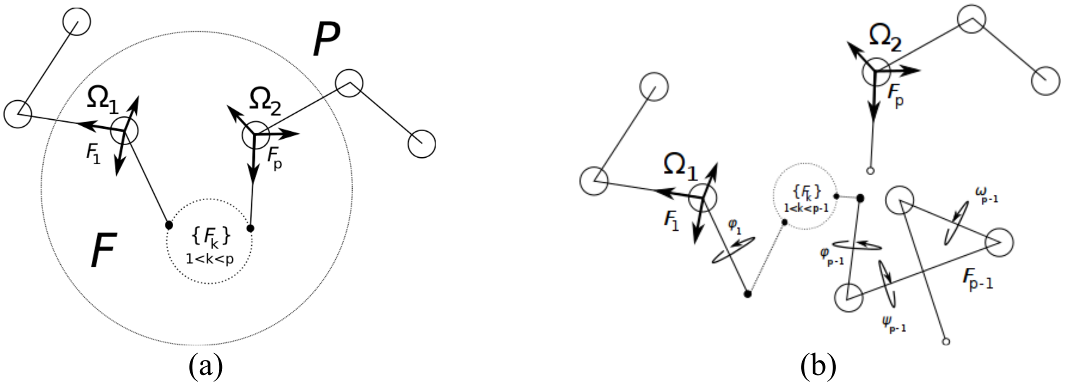


Figure 1: Linkage kinematic model: (a) Dihedral angle around a covalent bond. (b) Model of a protein fragment



(a) (b)

Figure 2: (a) Here coordinate frames Ω_1 and Ω_2 are placed with origins relative to the centers of the appropriate terminus atoms of a protein fragment F, with orientations defined relative to the bonds connecting the atom to its two neighboring atoms in the main-chain. (b) When Ω_2 and Ω_1 are consistent with the coordinate frames of their attachment points to the protein body *P*, *F* is geometrically consistent with *P*. Arbitrary choice of ϕ and ψ angles produce inconsistent (open) conformations; notice the last atom of F_{p-1} does not connect to the next sequential atom of F_p.