Journal of Artificial Intelligence Research 57 (2016) 509-572 Submitted November, 2015; published November, 2016

A Survey of Computational Treatments of Biomolecules by Robotics-Inspired Methods Modeling Equilibrium Structure and Dynamics

Amarda Shehu AMARDA@GMU.EDU
Department of Computer Science, Department of Bioengineering,
School of Systems Biology
George Mason University, Fairfax, VA, USA

Erion Plaku PLAKU@CUA.EDU
Department of Electrical Engineering and Computer Science
Catholic University of America, Washington, DC, USA

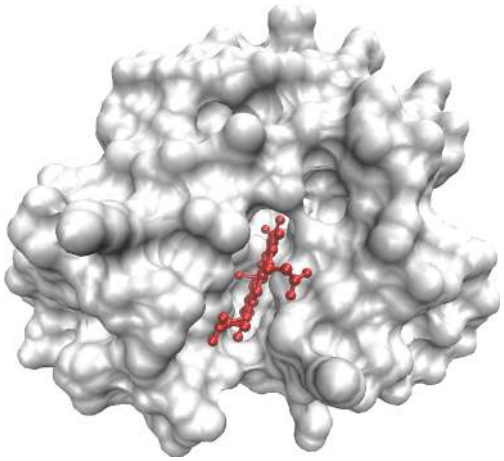
Abstract

More than fifty years of research in molecular biology have demonstrated that the ability of small and large molecules to interact with one another and propagate the cellular processes in the living cell lies in the ability of these molecules to assume and switch between specific structures under physiological conditions. Elucidating biomolecular structure and dynamics at equilibrium is therefore fundamental to furthering our understanding of biological function, molecular mechanisms in the cell, our own biology, disease, and disease treatments. By now, there is a wealth of methods designed to elucidate biomolecular structure and dynamics contributed from diverse scientific communities. In this survey, we focus on recent methods contributed from the Robotics community that promise to address outstanding challenges regarding the disparate length and time scales that characterize dynamic molecular processes in the cell. In particular, we survey robotics-inspired methods designed to obtain efficient representations of structure spaces of molecules in isolation or in assemblies for the purpose of characterizing equilibrium structure and dynamics. While an exhaustive review is an impossible endeavor, this survey balances the description of important algorithmic contributions with a critical discussion of outstanding computational challenges. The objective is to spur further research to address outstanding challenges in modeling equilibrium biomolecular structure and dynamics.

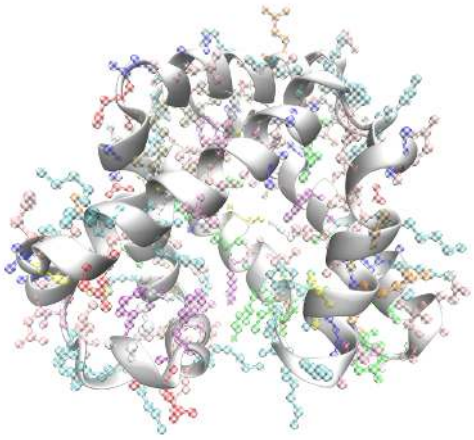
1. Introduction

"The way in which the chain of amino acid units in a protein molecule is coiled and folded in space has been worked out for the first time. The protein is myoglobin, the molecule of which contains 2,600 atoms." This is how John Kendrew began his feature article in *Scientific American* in 1961, reporting what was the first atomistic model of a protein structure[1] obtained via X-ray crystallography (Kendrew, Dickerson, Strandberg, Hart, Davies, Phillips, & Shore, 1960). This model is drawn in various graphical representations in Figure 1. For the pioneering work on resolving structures of globular proteins, Kendrew and Perutz were awarded the Nobel Prize in chemistry in 1962. This was the very same year Watson, Crick, and Wilkins shared the Nobel Prize in physiology or medicine for using X-ray crystallography data to determine the helical structure of DNA.

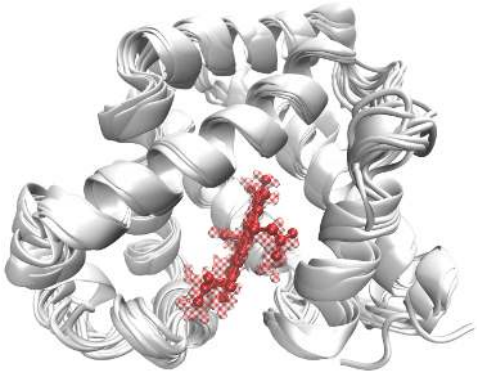
1. For the purpose of this survey, we will distinguish between structure and conformation. Structure will refer to a specific placement of the atoms that comprise a biomolecule in $\mathcal{R}^3$. The concept of conformation is defined in Section 2.



(a) X-ray model of myoglobin and heme group



(b) Model in atomistic detail (no heme group)



(c) Various models of myoglobin and heme group

Figure 1: (a) The X-ray model of myoglobin and the heme group bound to it determined by Kendrew are drawn here with the Visual Molecular Dynamics (VMD) software (Humphrey et al., 1996). The model can be found in the Protein Data Bank (PDB) (Berman et al., 2003), which is a repository of known protein structures, under PDB entry 1MBN. Drawing the surface of this protein facilitates visually locating the cavity where the heme group, which helps myoglobin to carry off oxygen to tissue, binds. The heme group is drawn in a ball-and-stick representation in red. (b) All heavy atoms that comprise the 153-amino acid long myoglobin chain are drawn in a ball-and-stick representation, color-coded by the amino acid to which they belong. The backbone that connects atoms of consecutive amino acids in the chain is drawn in white in the NewCartoon representation in VMD. (c) The X-ray model of myoglobin under PDB entry 1MBN is superimposed over the 12 models obtained for the same protein and the bound heme group from Nuclear Magnetic Resonance (NMR), deposited in the PDB under PDB entry 1MYF.

The ability to visualize structures of biomolecules in atomistic detail was a shot in the arm to molecular biology and marked the beginning of a revolution in molecular structural biology; a race soon ensued across wet laboratories to determine three-dimensional (3d) structures assumed by proteins and other biomolecules under physiological conditions. Since those early days, the set of protein structures resolved in the wet-laboratory, beginning with myoglobin and lysozyme (Kendrew, Bodo, Dintzis, Parrish, Wyckoff, & Phillips, 1958; Kendrew et al., 1960; Phillips, 1967), has grown

to over a hundred thousand, now freely available for anyone to download from the PDB (Berman et al., 2003).

Further pioneering work by Anfinsen, which earned him the Nobel Prize in Chemistry in 1973, demonstrated that the ability of a protein to carry out its biological function is dependent on its ability to fold onto a specific 3d structure reversibly (Anfinsen, 1973). The Anfinsen experiments led to the view that a folded structure corresponds to the global minimum of an underlying energy surface. They also showed that the information needed for a protein to assume its 3d, biologically-active structure is largely encoded in its amino-acid sequence. Since then, any study of biomolecular function has to consider the role of both sequence and structure (Fersht, 1999).

Figure 1(a), which shows the surface of the biologically-active structure of the myoglobin protein, exposes a central cavity that allows binding of the heme group to myoglobin. Figure 1(b) traces the amino-acid chain that makes up myoglobin and additionally draws the heavy atoms constituting each amino acid in this protein. These simple images illustrate two important points: first, that structure plays a central role in function (specifically, complementary geometric and physicochemical features of 3d structures of molecules are key to stable molecular interactions) (Boehr & Wright, 2008); second, that the ability of the amino-acid chain to fold onto itself makes protein structures complex. Understanding how and what biologically-active structure a biomolecule assumes in the cell is key not only to elucidating molecular mechanisms in the healthy and diseased cell, but also determining how to address the abnormal role of a biomolecule in such mechanisms in order to treat disease. In particular, research has shown that many abnormalities involve proteins with aberrant biological function (Soto, 2008; Uversky, 2009; Fernández-Medarde & Santos, 2011; Neudecker, Robustelli, Cavalli, Walsh, Lundstrm, Zarrine-Afsar, Sharpe, Vendruscolo, & Kay, 2012) due to external and internal perturbations (e.g., DNA mutations, copying errors) affecting the ability of these molecules to assume specific structures (Onuchic, Luthey-Schulten, & Wolynes, 1997; Ozenne, Schneider, Yao, Huang, Salmon, Zweckstetter, Jensen, & Blackledge, 2012; Levy, Jortner, & Becker, 2001; Miao, Sinko, Pierce, Bucher, Walker, & McCammon, 2014; Gorfe, Grant, & McCammon, 2008; Grant, Gorfe, & McCammon, 2009).

Any treatment of the relationship between structure and function would be incomplete if the "dynamic personality" of biomolecules is not taken into account (Jenzler-Wildman & Kern, 2007). While X-ray models of biomolecular structures seem to suggest rigid molecules with atoms frozen in space, an increasing number of wet-laboratory, theoretical, and computational studies have shown that biomolecules are systems of particles in perpetual motion. Indeed, Feynman taught early about the jiggling and wiggling of atoms (Feynman, Leighton, & Sands, 1963). Cooper and others later posited that the inherent dynamics of biomolecules could be explained under a general, theoretical treatment of molecules as thermodynamic systems striving towards their equilibrium, lowest free-energy state (Cooper, 1984). Thus, the inherent dynamics of biomolecules could be explained using fundamental physics principles; a statistical mechanics formulation also revealed the inherent uncertainty at any given time about the particular state of a molecule (Cooper, 1984).

The dynamics of molecular systems was investigated around the same time the first experimental models of protein structures were emerging. In 1967, Verlet simulated the dynamics of argon and demonstrated that such simulations were able to reproduce equilibrium properties (Verlet, 1967). Application of the Verlet algorithm for simulating protein dynamics would have to wait for one more decade. In 1977, McCammon and Karplus reported on a 9.2 picosecond-long trajectory showing in-vacuum, atomistic fluctuations of the bovine pancreatic trypsin inhibitor around its folded, active structure (the latter had already been obtained via wet-laboratory techniques) (McCammon,

Gelin, & Karplus, 1977). Advancements in wet-laboratory techniques, which had spewed about a dozen models of biologically-active protein structures by the late 70s, facilitated a revolution in computational structural biology. The pioneering algorithmic work of Verlet, Karplus, McCammon, Levitt, Warshel, and Lifson (for which Karplus, Levitt, and Warshel shared the 2013 Nobel Prize in chemistry) provided the earliest frameworks for computational treatments of biomolecules as a means to investigate equilibrium structure and dynamics (Fersht, 2013).

Since those early days, advances in wet-laboratory techniques have proceeded hand in hand with advancements in computational techniques, often feeding off each-other. The advent of NMR for structure determination provided evidence of the ability of biomolecules to fluctuate between different structures even at equilibrium (Kay, 1998, 2005). Figure 1(c) shows, in addition to the X-ray structure of myoglobin and its bound heme group, twelve models obtained via NMR, showcasing the intrinsic flexibility of this important biomolecule and its molecular partner in the cell. Nowadays, wet-laboratory techniques, such as NMR and cryo-Electron Microscopy (cryo-EM) can resolve equilibrium structures and quantify equilibrium dynamics. For example, NMR has been used to identify well-populated intermediate structures along a transition (Ådén & Wolf-Watz, 2007). Hybrid techniques that combine NMR relaxation measurements with X-ray models derived from room-temperature crystallographic, single-molecule spectroscopy techniques that tune optical radiation to observe one molecule, and others can now elucidate fast and slow dynamic processes lasting from a few picoseconds to a few milliseconds (Torella, Holden, Santoso, Hohlbein, & Kapanidis, 2011; Fenwick, van den Bedem, Fraser, & Wright, 2014; Karam, Powdrill, Liu, Vasquez, Mah, Bernatchez, Götte, & Cosa, 2014; Moerner & Fromm, 2003; Greenleaf, Woodside, & Block, 2007; Michalet, Weiss, & Jäger, 2006; Diekmann & Hoischen, 2014; Hohlbein, Craggs, & Cordes, 2014; Schlau-Cohen, Wang, Southall, Cogdell, & Moerner, 2013; Moffat, 2003; Schotte, Lim, Jackson, Smirnov, Soman, Olson, Phillips, Wulff, & Anfinrud, 2003; Roy, Hohng, & Ha, 2008; Fenwick et al., 2014; Hohlbein et al., 2014; Lee, M., Kim, & Suh, 2013; Socher & Imperiali, 2013; Gall, Ilioaia, Krüger, Novoderezhkin, Robert, & van Grondelle, 2015).

In particular, wet-laboratory techniques that employ fluorescence-based sensors, can provide information on dynamic, biological events by effectively monitoring changes in the signals of strategically-placed fluorophores (Socher & Imperiali, 2013). Depending on the placement of the fluorophores, the binding of two molecular partners or the switch/transition of one molecule between different structures can be monitored in real time. While such techniques are very promising and rapidly being adopted to study specific biological systems of interest, the reliance on fluorophores limits the generality of these techniques, as well as the structural detail that can be obtained. At the moment, wet-laboratory techniques obtain an incomplete view of equilibrium dynamics, as they are generally unable to span all the disparate length and time scales involved in a structural transition of a molecule (Maximova, Moffatt, Ma, Nussinov, & Shehu, 2016); While atomic motions occur on the picosecond scale, side-chain motions can take a few nanoseconds, and concerted motions among groups of atoms facilitating structural rearrangements for molecular recognition events can take anywhere from a few microseconds to a few milliseconds (Shaw, Maragakis, Lindorff-Larsen, Piana, Dror, Eastwood, Bank, Jumper, Salmon, Shan, & Wriggers, 2010; Lindorff-Larsen, Piana, Dror, & Shaw, 2011; Zagrovic, Snow, Shirts, & Pande, 2002; Piana, Lindorff-Larsen, & Shaw, 2012b); in extreme cases, binding of natural and drug molecules to proteins occurs on the hours scale (Hoelder, Clarke, & Workman, 2012).

Computational treatments of biomolecules are driven by the promise of complementing wet-laboratory treatments in obtaining a comprehensive and detailed characterization of equilibrium

dynamics. The current most well-known frameworks employed *in silico* are Molecular Dynamics (MD) (Verlet, 1967; McCammon et al., 1977) and Monte Carlo (MC) (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). In principle, the entire equilibrium dynamics of a molecule can be simulated by simply following the motions of the constitutive atoms along the physical forces that atoms impose on one another. This is the foundation of the MD framework. In contrast, in the MC framework, structural perturbation moves applied to atoms or bonds connecting atoms are not the result of physical forces but instead design decisions. Different local search strategies can be formulated to make use of such moves and iteratively explore neighborhoods in the structure space of a biomolecule.

The scope and capabilities of MD- and MC-based treatments of biomolecules have been significantly increased due to improvements in hardware and parallel computation strategies. Specialized architectures, such as Anton, a supercomputer designed for MD simulations, (Piana, Lindorff-Larsen, Dirks, Salmon, Dror, & Shaw, 2012a; Piana et al., 2012b; Lindorff-Larsen et al., 2011), GPUs (Stone, Phillips, Freddolino, Hardy, Trabuco, & Schulten, 2007; Harvey, Giupponi, & de Fabritiis, 2009; Tanner, Phillips, & Schulten, 2012; Götz, Williamson, Xu, Poole, Le Grand, & Walker, 2012), and petascale national supercomputers, such as BlueWaters, Titan, Mira, and Stampede (Dubrow, 2015; Zhao, Perilla, Yufenyuy, Meng, Chen, Ning, Ahn, Gronenborn, Schulten, Aiken, & Zhang, 2013) have allowed characterizing biomolecular structure and dynamics up to the microsecond time scale. Algorithmic improvements in dynamic load balancing (Fattebert, Richards, & Glosli, 2012), neighbor searches (Proctor, Lipscomb, Zou, Anderson, & Cho, 2012), and optimal force splitting (Batcho, Case, & Schlick, 2001) allow effectively distributing the simulation of the dynamics of molecular systems comprised of billions of particles (Perilla, Goh, Cassidy, Liu, Bernardi, Rudack, Yu, Wu, & Schulten, 2015).

In principle, a full account of the equilibrium dynamics of a biomolecule requires a comprehensive characterization of both the structure space available to the biomolecule at equilibrium as well as the underlying energy surface that governs accessibility of structures and transitions between structures at equilibrium. This remains challenging to do via MD and MC-based frameworks, and algorithmic enhancements of the classic MD and MC frameworks essentially aim to enhance their sampling of the structure space of a biomolecule. A review of state-of-the-art enhancements can be found in the work of Maximova et al. (2016).

In this survey paper, we focus instead on emerging contributions from the Robotics community on how to enhance sampling with complementary algorithmic strategies. Specifically, we review robotics-inspired methods designed to model structural excursions of a biomolecule at equilibrium by building conceptually over techniques designed originally for robot motion planning. These methods have now reached a crucial stage. They have been shown applicable to characterization of diverse molecular mechanisms in computational structural biology, such as protein-ligand binding, folding and unfolding of peptides, proteins, and RNA molecules, and transitions of small peptides and large proteins between thermodynamically-stable and semi-stable structural states. As this survey shows, these methods are capable of addressing challenging computational issues posed in each of these application settings, but they have yet to be widely adopted by the computational biology community at large. For various reasons, some of which are discussed in this survey, these methods are seen as providing an efficient but less detailed and less accurate characterization of biomolecular equilibrium structure and dynamics. This survey provides a critical review of robotics-inspired methods and lays out outstanding issues that need to be addressed for these methods to be considered reliable tools and be widely adopted for modeling biomolecular structure and dynamics.

This survey is organized as follows. A background of models of biomolecular energetics and geometry is provided in Section 2. Section 3 then introduces the main classes of problems in biomolecular modeling addressed with robotics-inspired methods, summarizes the robot motion planning frameworks over which such methods build, and concludes with a brief description of challenges faced by robotics-inspired methods in the context of biomolecular modeling. Section 4 provides examples of design decisions that address such challenges through a comprehensive and detailed review of robotics-inspired methods for modeling biomolecular structure and dynamics. Section 5 concludes this survey with a critical summary of remaining challenges and a discussion of several prospects for future research.

2. Background

The structural excursions that regulate the recognition events in which a biomolecule participates in the cell can be understood via a theoretical treatment of biomolecules as thermodynamic systems hopping between energetic states. These hops are fundamentally the result of concerted motions of the atoms that make up a biomolecule; in any physical system, constitutive particles are in a state of perpetual motion, fuelled by thermal excitation, all the while subjecting one another to physical forces (Cooper, 1984). These forces cumulatively drive a molecular system toward lower-energy states, while thermal excitations kick them off locally-optimal states, providing sufficient randomness to allow the entire system undergo a biased exploration of its structure space.

In the following we first summarize current knowledge on atomic forces and biomolecular energy functions that are employed in computational treatments of biomolecular structure and dynamics. The rest of this Section provides details on biomolecular geometry, showing how biomolecules can be treated mechanistically as modular systems composed of numerous, heterogeneous components for the purpose of characterizing their equilibrium structure and dynamics *in silico*.

2.1 Biomolecular Energetics: Molecular Mechanics

The physical interactions among the particles that make up a molecular system can in principle be measured via quantum mechanics (QM) methods. QM methods can carry out detailed and accurate electronic structure calculations but are currently limited in their applicability to molecular systems composed of no more than a few hundred atoms (Khaliullin, VandeVondele, & Hutter, 2013). Instead, molecular mechanics (MM) methods are now the methods of choice to evaluate structures of macromolecules, such as proteins, RNA, DNA, and other large molecular systems comprised of several molecules.

Though it was long known that atoms in a molecule subject one another to physical forces, such as Coulomb forces and others, it was the work of Levitt and Warshel in the Lifson laboratory at the Weizmann Institute of Science that propelled the design of consistent (now known as MM) energy functions for molecules. Lifson argued that it should be possible to come up with a small number of consistent, transferable parameters that do not depend on the local environment of an atom and allow analyzing the energetics of small crystalline molecules (Lifson & Warshel, 1968). Kendrew realized that such consistent energy functions could be used to conduct energetic evaluations on different placements (that is, structures) of the atoms comprising proteins and nucleic acids. Levitt and Lifson operationalized this realization to conduct energy refinements of protein structures (Levitt & Lifson, 1969).

MM energy functions have now become more detailed and accurate, but they are still estimates of the true potential energy of a molecule, summing up the possible, physical interactions between atoms in a molecule. Research on designing MM energy functions is active, and there are now many different functions offered from different computational chemistry labs across the world. These functions largely follow a common functional form, and categorize pairwise atomic interactions into local and non-local interactions; it should be noted that more accurate energy functions are available that consider n-particle interactions, but these are more computationally expensive and not widely adopted (Clementi, 2008). Local interactions concern modeling forces due to bonds, bond angles, and the periodicity of dihedral/torsion angles. Non-local interactions are divided into electrostatic (measured through the Coulomb potential) and van der Waals (measured through the Lennard-Jones – LJ – potential) interactions. These different types of interactions are typically linearly combined together, each with its own weight, to associate a potential energy value with a particular placement of atoms in a given molecular structure.

The following equation provides an example of the popular CHARMM energy function (Brooks, Bruccoleri, Olafson, States, Swaminathan, & Karplus, 1983) that is integrated in the NAMD software package for simulation of biomolecular dynamics (Phillips, Braun, Wang, Gumbart, Tajkhorshid, Villa, Chipot, Skeel, Kalé, & Schulten, 2005).

$$
\begin{aligned}
E_{\text{CHARMM}} = & \sum_{\text{bonds}} k_b \cdot (b - b_0)^2 + \\
& \sum_{\text{UB}} k_{\text{UB}} \cdot (S - S_0)^2 + \\
& \sum_{\text{valence angles}} k_\alpha (\alpha - \alpha_0)^2 + \\
& \sum_{\text{dihedral angles}} k_\theta \cdot (1 + \cos(n\theta - \delta)) + \\
& \sum_{\text{improper dihedral angles}} k_{\text{imp}} (\theta - \theta_0)^2 + \\
& \sum_{\text{non-bonded atoms i, j}} \epsilon_{ij} \Big[\big(\frac{Rmin_{ij}}{r_{ij}} \big)^{12} - 2 \big(\frac{Rmin_{ij}}{r_{ij}} \big)^{6} \Big] + \\
& \sum_{\text{non-bonded atomsi, j}} \frac{q_i \cdot q_j}{\epsilon \cdot r_{ij}}
\end{aligned}
$$

The k weights are constants, and the 0 subscript indicates equilibrium, ideal values of distances and angles. The first term effectively penalizes deviations of bond lengths from equilibrium values with a quadratic potential. The second term, also referred to as the Urey Bradley (UB) or the 1,3 term, introduces a similar penalty for pairs of atoms separated by two covalent bonds, with the distance between two atoms involved in a 1,3 interaction denoted by S. The third term is a quadratic potential for valence angles (between two consecutive bonds), denoted by α. The fourth term is a potential calculated over dihedral/torsion angles θ and models the presence of steric barriers between atoms separated by three covalent bonds. In this term, the n and δ variables are the multiplicity and the phase angles, respectively. In CHARMM, improper dihedral angles are specially penalized, as in the fifth term. The sixth term shows the LJ potential in CHARMM. In the LJ term summing up

van der Waals interactions between non-bonded atoms, r_{ij} measures the Euclidean distance between two non-bonded atoms (that are not covered by the UB term), and $Rmin_{ij} = (Rmin_i + Rmin_j)/2$ is the minimum interaction radius between the atoms, measured as half the sum of the known van der Waals radii $Rmin_i$ and $Rmin_j$ (ϵ_{ij} is a weight specific to the types of atoms i and j). The LJ term sums up a weak attraction at long distances and strong repulsion at short distances. The LJ term in CHARMM has a $12-6$ functional form, with an exponent of 12 for the repulsive sub-term and an exponent of 6 for the attractive sub-term. The last term in the CHARMM function measures electrostatic interactions via the Coulomb potential: q_i and q_j are the known partial charges of atoms i and j, r_{ij} measures the Euclidean distance between atoms i and j, and ϵ is the dielectric constant encoding the type of environment in which a biomolecule is (vacuum or different types of solvent environments).

Differences between available potential energy functions are due to different weights, different exponents used to measure the repulsion versus attraction terms in the van der Waals interaction, explicit estimation of hydrogen-bonding interactions outside the umbrella of van der Waals interactions, and more (Hornak, Abel, Okur, Strockbine, Roitberg, & Simmerling, 2006). The Amber suite of energy functions, integrated in the Amber MD simulation package (Case, Darden, Cheatham, Simmerling, Wang, Duke, Luo, Merz, Pearlman, Crowley, Walker, Zhang, Wang, Hayik, Roitberg, Seabra, Wong, Paesani, Wu, Brozell, Tsui, Gohlke, Yang, Tan, Mongan, et al., 2014), OPLS (Jorgensen, Maxwell, & Tirado-Reves, 1988), and CHARMM follow a similar functional form. Other similar functions are CEDAR (Hermans, Berendsen, van Gunsteren, & Postma, 1984) and GROMOS (van Gunsteren, Billeter, Eising, Hünenberger, Krüger, Mark, Scott, & Tironi, 1996), now incorporated in the GROMACS simulation package (Van Der Spoel, Lindahl, Hess, Groenhof, Mark, & Berendsen, 2005), and others. A review of these functions, known as physics-based function, can be found in the work of Ponder and Case (2003). Other functions, known as knowledge-based function, include additional terms derived from conducting statistics over known active structures of proteins in the PDB. Such functions are best suited for specific applications, such as rapid modeling of equilibrium structures. Rosetta (Leaver-Fay, Tyka, Lewis, Lange, Thompson, Jacak, Kaufman, Renfrew, Smith, Sheffler, Davis, Cooper, Treuille, Mandell, Richter, Ban, Fleishman, Corn, Kim, Lyskov, Berrondo, Mentzer, Popovi, & et. al., 2011) and Quark (Xu & Zhang, 2012) are recent examples of knowledge-based functions.

Whether physics-based or knowledge-based (or hybrid), all current molecular energy functions are models and, as such, they contain inherent errors that need to be taken into account when modeling biomolecular structure and dynamics (Hornak et al., 2006). In addition, in all such functions, despite the specific functional form, the most computationally expensive terms are the LJ and Coulomb terms due to the summation over pairs of atoms. These terms are also the ones that are most sensitive to small atomic motions. In particular, the 12-6 functional form of the LJ term provides great complexity and non-linearity to the energy surface that one can associate with the structure space of a biomolecule. It is quite common to reduce the total energy of a structure by a few hundred calories solely due to improvements in the LJ term from imperceptible changes in atomic positions. Moreover, small atomic displacements can lower the value of one term while increasing that of another in the energy function. From an optimization point of view, the terms that are linearly combined in an energy function are essentially conflicting optimization objectives. In computational biology, this issue is known as "frustration" and results in rugged or rough energy surfaces (that is, rich in local minima). In the broader AI community, such surfaces would be referred to as multi-modal. While true biomolecular energy surfaces are not overly rugged (known

as the principle of minimal frustration) (Clementi, 2008; Nevo, Brumfeld, Kapon, Hinterdorfer, & Reich, 2005), the *modeled* energy surfaces that one can probe *in silico* with the current energy functions have been shown exceptionally rugged (Olson & Shehu, 2012; Molloy, Saleh, & Shehu, 2013; Lois, Blawzdziewicz, & O'Hern, 2010).

2.2 The Biomolecular Energy Surface

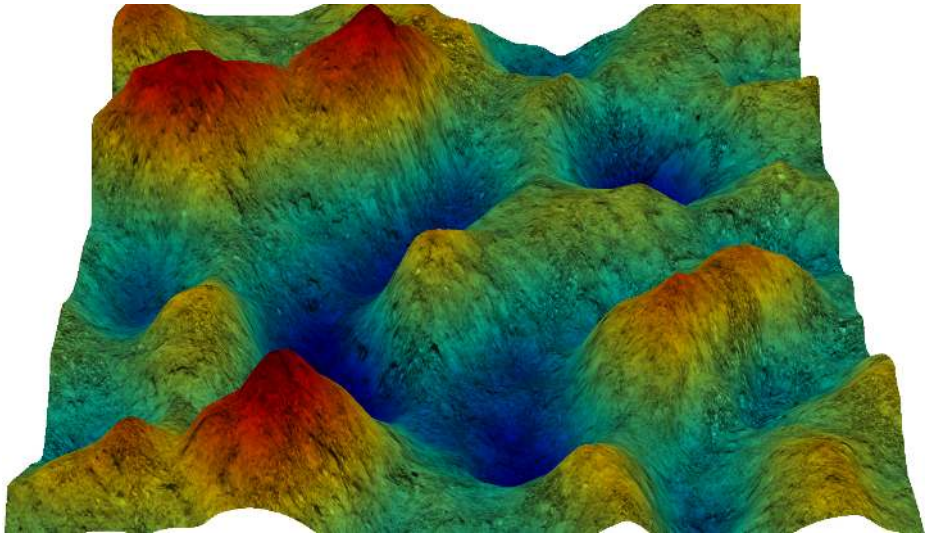
Equilibrium biomolecular dynamics can be visualized as structural excursions on the energy surface. The picture that emerges for proteins is that of a funnel-like, multidimensional energy surface (Onuchic et al., 1997; Dill & Chan, 1997). Projecting the surface onto few coordinates that capture relevant features of the different structures would allow summarizing and thus visualizing the energy surface in terms of a landscape (Onuchic & Wolynes, 2004).

The energy landscape shown in two different camera views in Figure 2 illustrates protein energy landscapes expected to be reconstructed *in silico*. Horizontal cross-sections of the landscape every dE units apart correspond to the different energetic states. The cross-sections go down in width as energy decreases; there are fewer options to place atoms in a molecule as potential energy gets lower without incurring energetic costs greater than dE. The width of a cross-section, or the structural diversity of an energetic state, is captured in the notion of entropy. Thermodynamically-stable states are those with low free energy F, measured as $F = \langle E \rangle - T \cdot S$, where $\langle E \rangle$ is the average potential energy over structures grouped together in the state, T is temperature, and S is entropy.

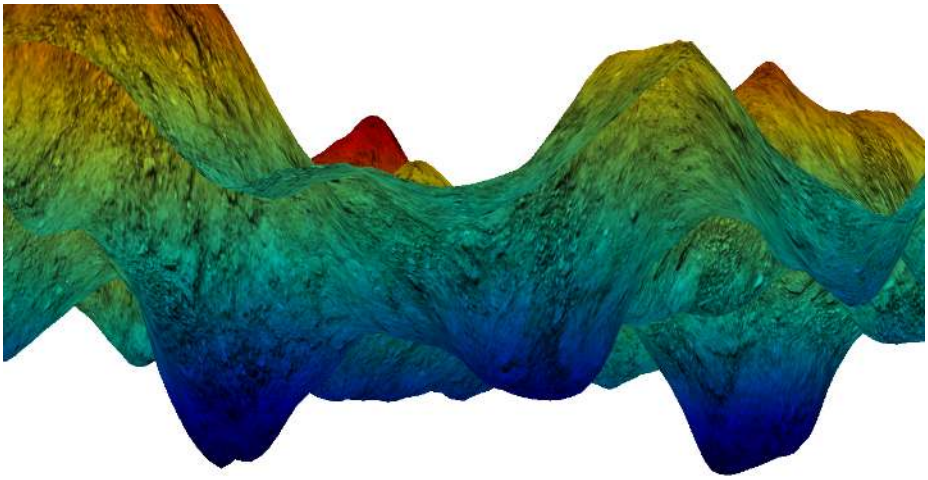
The first visual illustration, proposed by Dill and Chan (1997), highlighted the main features expected of true protein energy landscapes, a single, deep and wide basin corresponding to the thermodynamically-stable state and few other shallower, narrower basins corresponding to meta-stable states serving as possible kinetic traps. The landscape shown in Figure 2(a) is a synthetic one that is closer to the landscapes corresponding to existing MM energy functions; the landscape is not smooth but rather rich in local minima; in other words, the landscape is highly rugged or rough. A different camera view in Figure 2(b) emphasizes the presence of multiple, similarly-deep and wide basins among which current energy functions cannot further distinguish for the purpose of predicting the most stable state via energetic-based arguments; given the inherent errors, structure-function arguments cannot depend on small energetic differences.

The energy landscape view was instrumental in linking molecular structure, dynamics, and function. Viewing proteins and other biomolecules in terms of their energy landscapes gave rise to better understanding folding and binding as diffusion-like processes and not as a series of sequential, deterministic events. Under the new, landscape view (Baldwin, 1995), biomolecules can reach their most stable state at equilibrium by tumbling down the energy landscape along multiple routes (Bryngelson & Wolynes, 1987; Bryngelson, Onuchic, Socci, & Wolynes, 1995; Onuchic & Wolynes, 2004). In light of the new view, the intermediate, meta-stable states in which proteins would sometimes be found in the wet laboratory before transitioning to their most stable state correspond to other wide basins in the landscape. An illustration of this is provided in Figure 2(b).

The new view inspired a new understanding of dynamic molecular processes, known as conformational selection or population shift (Ma, Kumar, Tsai, & Nussinov, 1999; Tsai, Ma, & Nussinov, 1999b; Tsai, Kumar, Ma, & Nussinov, 1999a). Conformational selection refers to the idea that all states of an unbound molecular unit are present and accessible by the bound unit. For many unbound/uncomplexed biomolecules, there may be many semi-stable states at equilibrium. The proximity of a ligand or another molecular partner shifts the equilibrium (and thus the probability distri-



(a) Illustration of a complex energy landscape



(b) Tilted camera view highlights the presence of multiple energy basins

Figure 2: (a) The shown landscape illustrates what is often reconstructed *in silico*, rough landscapes rich in local minima. (b) The tilted camera view emphasizes the presence of multiple energetic basins. A basin is defined as the neighborhood of a local minimum in a *fitness* landscape. The interested reader is encouraged to learn more about features of landscapes that arise in optimization problems in Stadler's seminal review (Stadler, 2002). Given the high dimensionality of the structure space, many methods that probe energy landscapes cannot guarantee that a particular, sought stable or meta-stable state will be captured among the probed basins, or that none of the probed basins are artifacts of the energy function employed.

bution over possible states in which a system is found at equilibrium) towards one of the states that are close to optimal at equilibrium in the unbound/uncomplexed molecule. In other words, the presence of a binding partner can be considered an external perturbation to the unbound/uncomplexed energy landscape. Internal perturbations refer to changes in a biomolecule's composition itself due to changes to DNA, copy read errors, and other post-translation modifications that can occur. In many aberrant versions of a biomolecule, energy barriers between stable and semi-stable states can

drastically change and modify the underlying detailed structural mechanism regulating function, resulting in dysfunction or even loss of function (Clausen, Ma, Nussinov, & Shehu, 2015).

The principle of conformational selection allows employing analysis of energy landscapes of unbound molecules to identify structural states of interest for complexation events. The latter can be found among the meta-stable and stable states of the uncomplexed molecule and can thus be identified via modeling and simulation of the uncomplexed molecule.

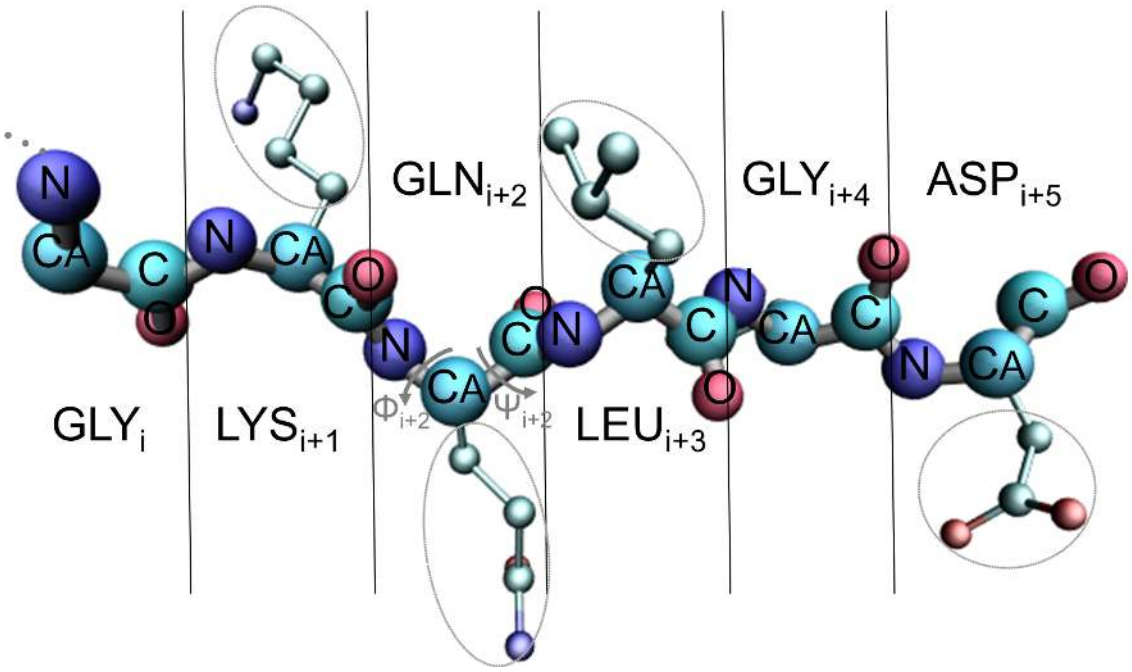
2.3 Computing Structures and Structural Transitions

The two main problems that can be addressed *in silico* to elucidate biomolecular equilibrium structure and dynamics concern (i) computing the ensemble of structures constituting the stable and meta-stable states relevant for biological function and (ii) computing the detailed structural transitions between such structures. The first problem is amenable to stochastic optimization, as it fundamentally involves locating deep and wide basins/minima in a nonlinear and multimodal energy surface. The second problem entails elucidating the different routes employed by a biomolecule as it switches between two structures. This survey focuses primarily on the second problem, that of computing structural transitions in bound and unbound biomolecules. Specifically, the survey reviews methods that employ robotics analogies to address this problem. While, in principle, robotics-inspired methods can also provide information on the structure space available to a biomolecule at equilibrium, other, more powerful stochastic optimization algorithms now exist for this purpose. We refer interested readers to the review by Shehu (2013), which surveys state-of-the-art evolutionary algorithms (EAs) capable of extracting efficient, discrete representations of protein energy surfaces.

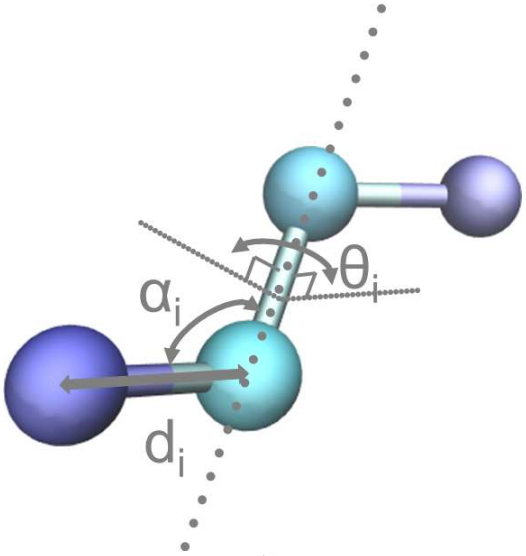
At a minimum, all algorithms aiming to model biomolecular structures are comprised of three functional units: (i) a way to represent/model a biomolecular structure; (ii) a way to modify such models in order to obtain new structures; (iii) and a way to evaluate the energetics of such structures. The existing MM energy functions summarized above in the context of biomolecular energetics provide a way to evaluate models of biomolecular structures explored *in silico*. The functional units (i) and (iii) are related, as the model chosen for a protein structure determines to a great extent what moves or perturbation operators can be designed to efficiently and effectively explore the structure space. Below, we summarize models that are now popular among the different algorithms employed to model equilibrium structures and dynamics of biomolecules. Details regarding the moves or perturbation operators designed to interface with such models are provided later in this survey when reviewing robotics-inspired methods.

2.4 Molecular Models: Selecting Variables of Interest

Covalent bonds link atoms together in a molecule. In protein molecules, atoms are organized into amino acids, which come in twenty different types in nature. All amino acids contain a common core of heavy atoms that make up its backbone and a unique set of heavy atoms that make up its side chain (hydrogen/light atoms are found both in the backbone and side-chain groups). The twenty naturally-occurring amino acids only differ in their side chain. Figure 3(a) shows the N, CA, C, and O heavy-atoms that comprise the backbone of an amino acid and further illustrates how amino acids are connected via covalent, peptide bonds in a serial fashion to form a (polypeptide) chain. What is referred to as a protein is often just one polypeptide chain; in protein-protein binding polypeptide chains stay together via non-covalent, weak interactions.



(a)



(b)

Figure 3: (a) In this chain of six amino acids, backbone atoms are N (gray), CA (black), C (gray), and O (silver). A peptide bond N_i-C_{i+1} links two amino acids (i proceeds from N- to C-terminus, which refer to backbone N and C atoms not in peptide bonds). Circled atoms comprise the side chain of each shown amino acid. (b) The three types of internal coordinates are shown here, the bond length d_i, the valence angle α_i between two consecutive angles, and the torsion or dihedral angle θ_i defined by three consecutive bonds. The dihedral angle is the angle between the two normals corresponding to each of the planes that can be defined by consecutive bonds j and $j+1$ and consecutive bonds $j+1$ and $j+2$. Depending on which backbone bonds they are defined, the θ_i dihedral angles are referred to as either ϕ or ψ, and annotated with the position of the amino acid on which they are defined (in the direction of the N- to the C-terminus). For instance, ϕ_i refers to the dihedral angle on the bond connecting the backbone N to the backbone CA atom of amino acid i, and ψ_i refers to the dihedral angle defined on the bond connecting the backbone CA to the backbone C atom of amino acid i. Characteristic values are observed for the ϕ, psi angles among equilibrium protein structures (Ramachandran et al., 1963).

Figure 3(a) illustrates that side chains dangle off the backbone of a polypeptide chain. Treating a protein molecule as a model where atoms are represented as balls and bonds between them as sticks exposes interesting questions regarding how to perform deformations of the model without breaking covalent bonds. This question is in essence what structure modeling researchers have to answer when defining variables to represent a molecule so as to be able to capture its intrinsic flexibility at equilibrium.

2.4.1 CARTESIAN COORDINATE-BASED MODELS

In the most intuitive model of a molecular structure, each Cartesian coordinate of each atom is selected as a variable. The Cartesian coordinate-based model is the preferred one by MD algorithms, which move individual atoms in a molecule according to the cumulative force that sums up interactions of an atom with all others in a molecule. However, this model is not ideal. First, it is redundant, demanding $3N$ variables for a molecule of N atoms. In small peptide and drug molecules, the number of atoms may be in the dozens, but even in small proteins, the number of atoms can easily surpass a hundred; as a result, the variable space has hundreds of dimensions.

Many strategies have been offered to reduce the number of variables by essentially removing certain atoms from modeling. For example, side-chain atoms are the first to be sacrificed in protein structure modeling, since it has been demonstrated that the main features of equilibrium protein structures are captured by the backbone (Rose, Fleming, Banavar, & Maritan, 2006). Once such reduced structures are modeled, side chains can be modeled via side-chain packing algorithms. In other studies focusing on modeling molecular interactions, only atoms that comprise the interaction site are explicitly modeled. These decisions effectively result in reduced models (and thus fewer dimensions of the selected variable space), ranging from CA traces, where only the central CA atom is modeled in each amino acid, to backbone models, where only the backbone chain is tracked in 3d space (Papoian, Ulander, Eastwood, Luthey-Schulten, & Wolynes, 2004; Matysiak & Clementi, 2004; Das, Matysiak, & Clementi, 2005; Matysiak & Clementi, 2006; Hoang, Trovato, Seno, Banavar, & Maritan, 2007; Rose et al., 2006). There is now a rich literature on reduced models and the energy functions designed to interface with these models (Clementi, 2008).

Representing a molecular structure in terms of the Cartesian coordinates of all (or a subset of) the constitutive atoms is appealing, as any new instantiation in the variable space can be readily evaluated in terms of its energetics. We recall that the central LJ and electrostatic/Coulomb terms in the energy functions that are now widely adopted in biomolecular structure and dynamics modeling operate over 3d coordinates of atoms. However, the Cartesian coordinate-based model is both redundant and ineffective. First, the model results in an excessive number of variables, which poses great challenges for any sampling-based method aimed at probing the structure space one sample at a time. Second, it is ineffective, as it does not encode in it any of the explicit and implicit geometric constraints present in molecular structures.

Many efforts in the computational biophysics community target the reduction of coordinates. The resulting models are referred to as coarse-grained models or representations. The first such model, employed in an MC simulation of the folding of the bovine pancreatic trypsin inhibitor, represented each amino-acid residue with one pseudo-atom (Levitt & Warshel, 1975). Work on coarse-grained and multiscale models (in the latter, different parts of the structure are represented at different levels of detail/resolution at different time and length scales) has been key to extend the spatio-temporal reach of MD and MC simulations of biomolecular dynamics. Such work has

the additional onerous task of designing accompanying energy functions that reproduce known thermodynamic properties even at lower or mixed structural resolution. Indeed, the 2013 Nobel Prize to Warshel recognized seminal work by him in multiscale models built with the QM/MM method (Warshel & Levitt, 1976; Warshel, 2003; Kamerlin, Haranczyk, & Warshel, 2009; Mukherjee & Warshel, 2011, 2012; Dryga, Chakrabarty, Vicatos, & Warshel, 2011; Rychkova, Mukherjee, Bora, & Warshel, 2013; Mukherjee & Warshel, 2013). The interested reader is directed to the review by Clementi (2008) on coarse-grained models. The review by Zhou (2014) focuses on multiscale models.

2.4.2 ENCODING VARIABLE DEPENDENCIES IN CARTESIAN COORDINATE-BASED MODELS

Outside of the realm of modeling chemical reaction processes, such as bond formation and breaking, when modeling biomolecular equilibrium structures and dynamics, other application setups require preserving certain structural features that can be formulated as local and non-local constraints. Some constraints, such as keeping bonded atoms at a distance no more than the ideal/equilibrium length of their bond, are known as explicit, local constraints. They are trivially extracted from specification of the chemical composition of a molecule, as they involve neighboring atoms. Equilibrium conditions place additional, implicit constraints on biomolecular structures. The need to preserve favorable Lennard-Jones interactions, for instance, places (non-local/long-range) constraints over non-bonded atoms. Such long-range constraints cannot be effectively captured by a model where there is a variable for each Cartesian coordinate of each atom. Any perturbation operator that interfaces with such a model and modifies such variables has no information on invalid or energetically-unfavorable assignments to subsets of variables, as variable dependencies are not captured in the model. An external energy model is crucial here in the form of an energy function to evaluate the results of the perturbation operator and detect variable instantiations resulting in violations.

The Cartesian coordinate-based model can encode variable dependencies. The latter can be extracted via several techniques, including multivariate analysis techniques that analyze known equilibrium structures of a biomolecule to identify subsets of atoms that exhibit simultaneous displacements; that is, move in concert. The essential premise of such techniques is that such known structures are good examples of solutions or near-solutions of the energy function, and that analysis of such examples will expose variable dependencies. These dependencies can then be employed to design reduced Cartesian coordinate-based models that readily encode in them the energetic constraints satisfied by the provided examples (solution or near-solution structures) and effective perturbation operators that readily yield new near-solution instantiations in the reduced variable space (Clausen & Shehu, 2015).

Multivariate Analysis Techniques to Obtain Collective Variables The sub-field of statistical techniques for the identification of collective motions, which are also referred to as collective coordinates or collective variables is rich, and a review is not the subject of our survey. Instead, we point to recent work in the narrow context of biomolecular modeling, where variance-maximizing techniques, such as Principal Component Analysis (PCA) (Shlens, 2003), Isomap (Tenenbaum, de Silva, & Langford, 2000), Locally Linear Embedding (Roweis & Saul, 2000), Diffusion Maps (Coifman, Lafon, Lee, Maggioni, Nadler, Warner, & Zucker, 2005), and others (van der Maaten, Postma, & van den Herik, 2009) have been employed to analyze biomolecular structures and dynamics (Teodoro, Phillips, & Kavraki, 2003; Das, Moll, Stamati, Kavraki, & Clementi, 2006; Plaku, Stamati, Clementi, & Kavraki, 2007; Gorfe et al., 2008; Grant et al., 2009; Hori, Chikenji, & Takada, 2009; Maisuradze,

Liwo, & Scheraga, 2009; Rohrdanz, Zheng, Maggioni, & Clementi, 2011; Zheng, Rohrdanz, Maggioni, & Clementi, 2011) and, more importantly, identify variables that represent collective motions of atoms (Zheng, Rohrdanz, & Clementi, 2013; Clausen & Shehu, 2015; Clausen et al., 2015; Molloy, Clausen, & Shehu, 2016; Maximova, Plaku, & Shehu, 2015). In addition to these methods, other ones such as Normal Mode Analysis (NMA) (Ciu & Bahar, 2005) also have a rich history in computational structural biology (Atilgan, Durell, Jernigan, Demirel, Keskin, & Bahar, 2001; Delarue & Sanejouand, 2002; Kim, Chirikjian, & Jernigan, 2002b; Zheng & Doniach, 2003; Tama, Valle, Frank, & Brooks, 2003; Bahar & Rader, 2005; Maragakis & Karplus, 2005; Zheng & Brooks, 2005; Zheng, Brooks, & Hummer, 2007; Yang, Song, Carriquiry, & Jernigan, 2008; Yang, Mâjek, & Bahar, 2009; Das, Gur, Cheng, Jo, Bahar, & Roux, 2014). The normal modes extracted from NMA are also often employed as as effective perturbation operators in robotics-inspired methods (Tama & Sanejouand, 2001; Kim, Jernigan, & Chirikjian, 2002a; Kirillova, Cortés, Stefaniu, & Siméon, 2008; Schuyler, Jernigan, Wasba, Ramakrishnan, & Chirikjian, 2009; Teknipar & Zheng, 2010; Baron, 2013; Al-Bluwi, Vaisset, Siméon, & Cortés, 2013). Our summary and highlights of robotics-inspired methods later in this survey describe in greater detail collective variables and their employment in effective perturbation operators.

2.5 Internal Coordinate- and Angular-Based Models

The internal coordinate model has been offered as an effective alternative to the Cartesian coordinate-based model (Burgess & Scheraga, 1975). In the internal-coordinate model, the only variables selected are bond lengths, angles between two consecutive bonds, and torsion or dihedral angles between three consecutive bonds. Figure 3(b) provides an illustration. This model allows for fast forward kinematics, as changes to Cartesian coordinates as a result of changes to the values of these variables can be efficiently calculated via accumulation of rigid-body transformations (Craig, 1989; Zhang & Kavraki, 2002a).

Internal coordinate-based models are now the norm in non-MD based molecular structure modeling. An additional simplification is made for equilibrium protein structures. Analysis of deposited equilibrium structures of proteins reveals that bond lengths and bond angles are constrained to characteristic values (Engh & Huber, 1991). This is a consequence of the energetic constraints placed on structures at equilibrium and is exploited to idealize protein geometry in modeling by effectively removing bond lengths and bond angles from the list of variables in the model. This leaves only dihedral angles defined over three consecutive bonds as variables (ϕ, ψ backbone angles and at most four dihedral side-chain angles per amino acid, as shown in Figure 3(a)) and is computationally appealing, as the number of dihedral angles for a polypeptide chain of N atoms is on average $3N/7$ (Abayagan, Totrov, & Kuznetsov, 1994). It is worth noting that bond lengths and bond angles do change even at equilibrium, but at a faster pace than other motions. Employing idealized geometry allows devoting computation to obtaining the slower fluctuations first. Once structures representative of a molecule's equilibrium dynamics are obtained, deviations of bond lengths and bond angles can be introduced and studied via more detailed models.

2.5.1 BIOMOLECULES AS KINEMATIC CHAINS WITH REVOLUTE JOINTS

Idealizing protein geometry reveals mechanistic analogies with kinematic chains with revolute joints. Similarly to how a joint rotation changes positions of following links, so does rotation by a dihedral angle change positions of following atoms (Craig, 1989). These analogies have been

employed by robotics researchers to apply algorithms that plan motions for kinematic chains with revolute joints to the study of protein conformations (Manocha & Zhu, 1994; Singh, Latombe, & Brutlag, 1999; Apaydin, Singh, Brutlag, & Latombe, 2001; Amato, Dill, & Song, 2003; Apaydin, Brutlag, Guestrin, Hsu, & Latombe, 2003; Song & Amato, 2004; Cortés, Siméon, & Tran, 2004; Cortés, Siméon, Guieysse, Remaud-Siméon, & Tran, 2005; Lee, Streinu, & Brock, 2005; Kim et al., 2002a; Chiang, Apaydin, Brutlag, Hsu, & Latombe, 2007; Shehu & Olson, 2010; Molloy et al., 2013; Molloy & Shehu, 2013; Haspel, Moll, Baker, Chiu, & E., 2010; Shehu, Clementi, & Kavraki, 2006). Unlike typical articulated robotic mechanisms, protein chains pose hundreds rather than a dozen variables (a short backbone of 50 amino acids poses 100 dihedral angles as variables).

The analogies between protein chains and kinematic chains with revolute joints are popular among robotics researchers proposing robotics-inspired methods for modeling biomolecular structure and dynamics. For instance, torsional angles were employed early to model protein ligand binding (Singh et al., 1999) and remain popular in modeling the kinetics of folding in small protein and RNA molecules (Han & Amato, 2001; Amato et al., 2003; Song & Amato, 2004; Thomas, Song, & Amato, 2005; Thomas, Tang, Tapia, & Amato, 2007; Tang, Thomas, Tapia, Giedroc, & Amato, 2008; Tapia, Thomas, & Amato, 2010). Such angles have also proved popular in computing functionally-relevant structures of peptides and proteins (Haspel, Tsai, Wolfson, & Nussinov, 2003; Shehu et al., 2006; Shehu, Clementi, & Kavraki, 2007; Shehu, Kavraki, & Clementi, 2007, 2008; Cortés et al., 2004; Shehu, Kavraki, & Clementi, 2009; Shehu, 2009; Shehu & Olson, 2010; Molloy et al., 2013), as well as in modeling peptides and proteins switching between different functionally-relevant structures (Cortés et al., 2005; Jaillet, Cortés, & Siméon, 2008; Haspel et al., 2010; Jaillet, Corcho, Perez, & Cortés, 2011; Molloy et al., 2016; Molloy & Shehu, 2013, 2015; Devaurs, Molloy, Vaisset, Shehu, Cortés, & Siméon, 2015; Molloy & Shehu, 2016).

Techniques to Obtain Reduced Angular-based Models The number of variables in angular-based models can be reduced further via various techniques. For instance, consecutive dihedral angles can be bundled together into fragments to capture variable dependencies. This technique, known as molecular fragment replacement and introduced in the context of MC-based methods for *de novo* protein structure prediction (Bradley, Misura, & Baker, 2005), allows operationalizing on the observation that a limited number of configurations are observed for k-bundles of consecutive dihedral angles among stable protein structures at equilibrium (Han & Baker, 1996). This technique has been incorporated in robotics-inspired methods for modeling equilibrium protein structure and dynamics (Shehu & Olson, 2010; Molloy et al., 2013; Molloy & Shehu, 2013, 2016). Other application-specific techniques analyze structures to reduce or prioritize the number of dihedral angles for manipulation by a perturbation operator. Rigidity-based techniques, for instance, analyze a given structure to detect least-constrained regions and suggest an order for which dihedral angles to modify first or more often in order to focus computational resources to computing the large structural deformations first (Thorpe & Ming, 2004; Wells, Menor, Hespenheide, & Thorpe, 2005; Fox & Streinu, 2013). Rigidity-based analysis has been incorporated in robotics-inspired methods for modeling protein dynamics (Thomas et al., 2007). Other techniques are aimed specifically at modeling structural transitions in large proteins (Raveh, Enosh, Furman-Schueler, & Halperin, 2009; Haspel et al., 2010). In these, a comparison of the two structures for which a transition is sought identifies the differently-valued dihedral angles. In a similar fashion as in rigidity-based analysis, these angles are prioritized and modified more often by perturbation operators in order to capture possibly large structural deformations in a reasonable amount of time. All these techniques make

several assumptions about which variables participate in the process of interest, and we highlight such assumptions and their implications later in this survey.

Biomolecular Structure versus Biomolecular Conformation In light of the various models that can be employed to represent biomolecular structure, a distinction needs to be made between the terms structure and conformation. The term structure is meant to refer to the specification of the Cartesian coordinates of the atoms that comprise a molecule (even if not all atoms are explicitly modeled, as in a backbone or reduced structure). The term conformation is meant to be more general and refer to the specification of the values of variables selected in the model of a structure; that is, a conformation is a particular instantiation in the employed variable space. For instance, a conformation is the instantiation of angles if an angular-based model is employed, and forward kinematics allows obtaining the structure encoded by such a conformation. It is worth noting that the terms conformation and structure are often used interchangeably and in a slight abuse of terminology in biomolecular modeling literature. For instance, many algorithms that explicitly modify structures via Cartesian coordinate-based models are referred to as conformational search algorithms; of course, when such models are employed, a structure can be extracted trivially from a conformation. In this survey, the distinction made above will be observed when referring to structures and conformations.

In the specific domain of robotics-inspired methods, the term (molecular) conformation is equivalent to (robot) configuration. However, in keeping with the broader computational biology literature, we will employ the term conformation when referring to macromolecules, such as proteins, RNA, and DNA, reserving the term configuration for small molecules (also referred to as ligands) that bind to macromolecules. In addition, while the term variable is often referred to as parameter in the broader computational biology and biophysics literature and degree of freedom (dof) in robotics and AI literature, we will employ the more general, non-domain specific term of variable and variable space. That is, a (molecular) conformation is an instantiation in the space of selected variables. Depending on the number of variables selected to model the molecule under investigation, the variable space may be high-dimensional. Mapping a conformation to its corresponding structure allows associating a structure space to the employed variable space. Moreover, energy functions allow associating an energy surface to the structure space, and interesting observations regarding stable and semi-stable structural states and excursions among such states can be made by analyzing the structure space and low-dimensional projections/embeddings of its underlying energy surface, hence the energy landscape.

Educational Resources The purpose of the material related above is to provide enough detail on biomolecular geometry to allow seeing how biomolecules can be treated mechanistically as modular systems composed of numerous, heterogeneous components for the purpose of characterizing them *in silico*. Further information for readers of varying levels of background or interest can be found in online, educational learning modules designed to introduce computer scientists to computational structural biology. One such set of modules, publicly accessible under the cnx project and highly popular with students and researchers, can be found at `http://cnx.org/contents/9cMfjngH@6.3:ppj-3H2A@14/Structural-Computational-Biolo`. In these modules, for instance, interested readers can learn about protein architecture in greater detail. Other modules mirror the material summarized here on representations and energy functions, and yet others introduce readers to forward and inverse kinematics for modular mechanical systems, making these modules a good supplement to the survey of robotics-inspired methods presented here.

3. Summary of Biomolecular Modeling Problems and Robot Motion Planning Frameworks

We introduce the main classes of problems in biomolecular modeling that are addressed with robotics-inspired methods. The robot motion planning frameworks over which these methods build are summarized next. The section concludes with a summary of challenges faced by algorithmic realizations of such frameworks for modeling biomolecular structure and dynamics. These challenges are a preview of important design decisions that are detailed in Section 4 in the context of reviewing representative methods.

3.1 Representative Problems in Biomolecular Modeling

Two main classes of molecular mechanisms are studied by robotics-inspired methods, those that involve more than one molecule associating/complexating with or disassociating from one another, and those that involve a dynamic, uncomplexated molecule. As application setups, both concern informing our understanding of dynamic events involving dynamic molecules.

In the first application setup, robotics-inspired methods aim to model and understand protein-ligand binding events. Provided unbound structures of the protein receptor and a small ligand molecule, the objective is to elucidate how the ligand approaches and then binds the protein receptor. In a related problem, the reverse process is addressed. The ligand is bound to the receptor, and the goal is to determine motions of the ligand and the protein receptor that allow disassociation. Modeling protein-ligand binding is important not only for a general understanding of our biology but also for computation-aided drug discovery. While other molecular recognition events such as protein-protein, protein-DNA, protein-RNA, and protein-membrane binding conceptually fall in the same category as protein-ligand binding, they are more challenging due to the higher number of variables needed to model the association or complexation event and are currently beyond the domain of applicability of robotics-inspired methods.

The second application setup concerns modeling and understanding the dynamics of molecules. Almost exclusively, the focus of robotics-inspired techniques in this category are uncomplexed protein and RNA molecules. The goal is to elucidate the structural deformations or motions that allow a protein or an RNA molecule to transition between two structural states of interest. These states can be the unfolded and folded state, in which case the goal is to highlight folding and unfolding paths, or they can both be stable or semi-stable structures employed by the molecule to recognize and lock onto different molecules, in which case the goal is to formulate hypothesis regarding the impact of structure in molecular recognitions in the healthy and diseased cell.

Figures 4-5 illustrates these problems. In Figure 4, a robotics-inspired method can highlight how the ligand approaches the protein receptor, as well as where it binds onto the receptor and into what configuration. The variables of interest need at a minimum to include the translational and rotational variables of the ligand, as well as internal coordinates to capture the potential structural flexibility of the ligand. The receptor can be considered frozen in 3d space. A more accurate setup would also consider internal coordinates of the receptor in order to model its possible structural flexibility upon ligand binding. However, the resulting number of variables would be too large. A robotics-inspired method can not only elucidate the bound ligand configuration and its bound placement relative to the receptor, but also the possible routes of successive configurations and placements that the ligand may follow to approach the binding site(s). In addition, as Figure 5(a) illustrates, a robotics-inspired method can show the possible routes of successive structures employed by a

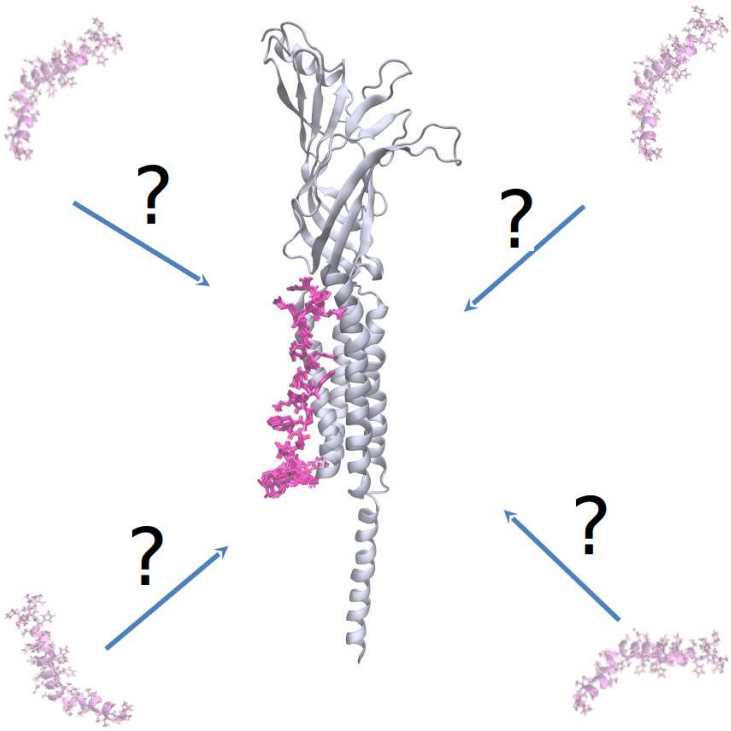
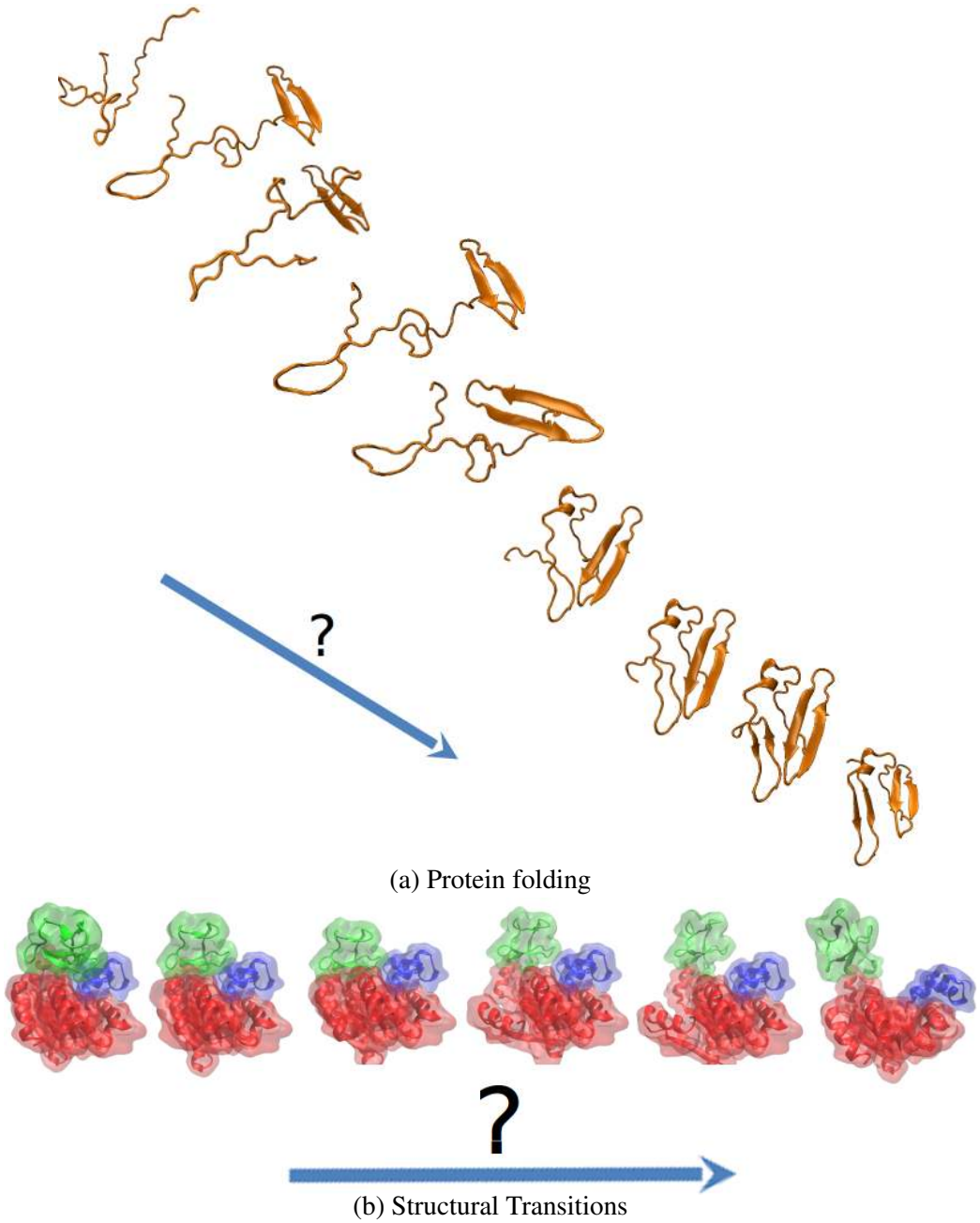


Figure 4: (a) Where and how does the ligand bind to the protein receptor? Many methods are designed to elucidate the step-by-step process of how a ligand approaches a protein molecule, where it binds, and with what configuration.

protein to fold, thus shedding light into the process of protein folding and unfolding. Similar setups consider RNA molecules. Figure 5(b) illustrates that often robotics-inspired methods are employed to reveal not only folding or unfolding routes of a protein, but also structural transitions between any two structures of interest; knowledge of the most probable routes that carry out the transition allows understanding at a structural level the mechanism by which a biomolecule regulates its biological activity in the cell.

3.2 Foundations of Robotics-inspired Treatments of Biomolecules

The fundamental assumption of robotics-inspired treatments of biomolecules is that mechanistic analogies between molecular chains and robot chains allow putting together efficient algorithms for rapid exploration of molecular structure spaces and modeling of excursions of molecules on such spaces (Manocha & Zhu, 1994; Singh et al., 1999; Apaydin et al., 2001; Amato et al., 2003; Apaydin et al., 2003; Song & Amato, 2004; Cortés et al., 2005; Kim et al., 2002a; Chiang et al., 2007; Kirillova et al., 2008). That is, instead of simulating how a molecule navigates its energy surface via gradient-based and other local search techniques, more powerful techniques can be put together by building on algorithms demonstrated to have high exploration capability on robot configuration spaces. Robotics-inspired treatments of biomolecules draw from techniques for fast forward and



(a) Protein folding

(b) Structural Transitions

Figure 5: (a) How do proteins fold? Shedding light into the process of protein folding is an important goal, and many robotics-inspired methods are devoted to elucidating this process step by step by computing the most probable succession of structures assumed by a protein navigating from the unfolded to the folded state. (b) How do proteins transition between the diverse structures they use for interacting with different partners in the cell? Robotics-inspired methods seek to elucidate structural transitions between meta-stable and stable structural states of a protein.

inverse kinematics and, more importantly, sampling-based algorithms developed in the algorithmic robotics community to address the robot motion-planning problem (Choset & et al., 2005).

The objective in robot motion planning is to obtain paths that take a robot from a given, start configuration to a given, goal configuration. The robot motion planning problem bears mechanistic analogies to the problem of computing conformations along a transition trajectory of a biomolecule; in both problems, the driving objective is to uncover what of the underlying (molecular) conformation or (robot) configuration space is employed in motions of an articulated system from a start to a goal conformation or configuration. Analogies between molecular bonds and robot links and molecular atoms and robot joints help to draw from techniques that perform fast kinematics for kinematic linkages (Manocha, Zhu, & Wright, 1995; Zhang & Kavraki, 2002b); that is, specifying values for the variables selected to represent a molecular conformation, rapidly update Cartesian coordinates of the corresponding structure (Zhang & Kavraki, 2002b). In the inverse kinematics setting, these techniques allow rapidly obtaining values to underlying variables consistent with Cartesian-based constraints (Chirikjian, 1993; Manocha & Canny, 1994; Zhang & Kavraki, 2002a; Kolodny, Guibas, Levitt, & Koehl, 2005).

At a higher level, robotics-inspired methods operationalize on two key observations. The first observation, originally made for robot configuration spaces, is that solution-containing regions of a high-dimensional and non-linear variable space can only be found by heuristic rather than exact approaches; stochastic, or sampling-based techniques can construct the distribution of constraint-satisfying instantiations in a two-stage manner; an initial distribution is first constructed via sampling of the unconstrained variable space. Instantiations are then evaluated through external (energetic) models capable of capturing inter-variable dependencies and penalizing violations of constraints. Violating samples are either removed or down-weighted so the initial, uninformed distribution gradually converges to that containing solutions. The second observation is that transitions of a system between two given solutions can be modeled via discrete, kinetic models that essentially embed solutions in graph-like structures amenable to rapid, shortest, or lowest-cost path queries, provided that lengths or other cost metrics can be associated with a given series of configurations in a path. Methods that embed solutions in a tree are referred to as tree-based methods, and those that embed solutions in a graph are referred to as roadmap-based methods.

3.3 The Motion Planning Framework: Tree- and Roadmap-Based Methods

In the context of molecular modeling, tree-based methods grow a tree in conformation space from a given, start to a given, goal conformation representing the structures bridged by the sought transition. The tree is incrementally extended, with every iteration adding a new conformation node and a new branch to the tree. Depending on whether the tree is pulled towards configurations sampled at random over the configuration space or pushed from leaves towards new regions of the configuration space, the method is known as Rapidly Random Exploring Tree (RRT) (LaValle & Kuffner, 2001), or Expansive Spaces Tree (EST) (Hsu, Kindel, Latombe, & Rock, 2002), accordingly. It is important to note here that the sampling and connectivity go hand in hand, as every sampled conformation is added to the growing tree. The growth of the tree is biased so the goal conformation can be reached in a reasonable computational time. As a result, tree-based methods are efficient but limited in their sampling. They are known as single-query methods, as they can only answer one start-to-goal query at a time; that is, only one path of consecutive conformations that connect the start to the goal can be extracted. Running them multiple times to sample an ensemble of conformation paths for the same query results in an ensemble with high inter-path correlations due to the biasing of the conformation tree.

Roadmap-based methods adapt the Probabilistic Road Map (PRM) framework (Kavraki, Švestka, Latombe, & Overmars, 1996). Rather than grow a tree in conformation space, these methods detach the sampling of conformations from the connectivity model that encodes neighborhood relationships among conformations in the conformation space. Typically, a sampling stage first provides a discrete representation of the conformation space of interest, with conformations satisfying explicit or implicit geometric and energetic constraints, and then a roadmap building stage embeds the sampled conformations in a graph/roadmap by connecting each one to its nearest neighbors. Roadmap-based methods can provide richer information regarding a dynamic event, as multiple paths may exist in the roadmap connecting two structures of interest. Moreover, these methods support multiple queries, as in principle the same graph can be used to extract paths connecting different given structures. These structures can be specified as conformations and connected to nearest neighbors in the graph, and then the graph can be queried for optimal paths. In practice, it is difficult to obtain broad and dense sampling of sufficient regions in the conformation space of a molecule so as to elucidate diverse excursions between structures of interest.

Below we provide further detail on tree-based and roadmap-based methods in robotics to familiarize the reader with the diverse design decisions employed and even adapted by robotics-inspired methods for biomolecular modeling.

3.3.1 NON-HIERARCHICAL TREE-BASED ROBOT MOTION PLANNING METHODS

Tree-based methods can be categorized broadly based on the strategies employed to select the vertex from which to expand the tree. Non-hierarchical strategies consider all the vertices as possible candidates. Hierarchical strategies place the tree vertices at a bottom layer and introduce additional high-level layers that group similar vertices together, proceeding from the top to the bottom layer during the selection process.

RRT is one of the most widely used non-hierarchical tree-based motion planning methods. In RRT (LaValle & Kuffner, 2001), at each iteration, the tree is expanded towards a randomly-sampled configuration $q_{\rm rand}$. The nearest vertex, $q_{\rm near}$, in the tree to $q_{\rm rand}$ is determined according to a distance metric. A local planner attempts to connect $q_{\rm near}$ to $q_{\rm rand}$. Often the local planner interpolates over the underlying variables to generate intermediate configurations. In the basic version of RRT, the iteration stops after one interpolation step. In the connect version, the expansion continues until $q_{\rm rand}$ is reached or the interpolation results in an invalid configuration, e.g., collision with obstacles. This process of sampling a configuration and expanding from the nearest neighbor in the tree is repeated until the goal is reached. Figure 6 shows such a tree. By using random sampling and nearest neighbors, RRT exhibits a Voronoi bias which enables the expansion of the tree toward unexplored regions. To also bias the search towards the goal, $q_{\rm rand}$ is often selected with probability b (often set to 0.05) as the goal configuration and with probability $1 - b$ uniformly at random.

Over the years, different RRT variants have been developed in order to improve the exploration. The adaptive dynamic domain RRT (ADDRRT) (Jaillet, Yershova, LaValle, & Siméon, 2005) associates a sampling radius with each tree vertex and dynamically adjusts the radius based on the success of the local planner. The reachability-guided RRT (RGRRT) (Shkolnik, Walter, & Tedrake, 2009) relies on the notion of reachable sets to increase the likelihood of successful tree expansions. The obstacle-based RRT (OBRRT) (Rodriguez, Tang, Lien, & Amato, 2006b) increases sampling near obstacles, PCARRT (Dalibard & Laumond, 2009) relies on PCA, and The selective retraction-based RRT (SRRRT) (Lee, Kwon, Zhang, & Yoon, 2014) uses bridge sampling and se-

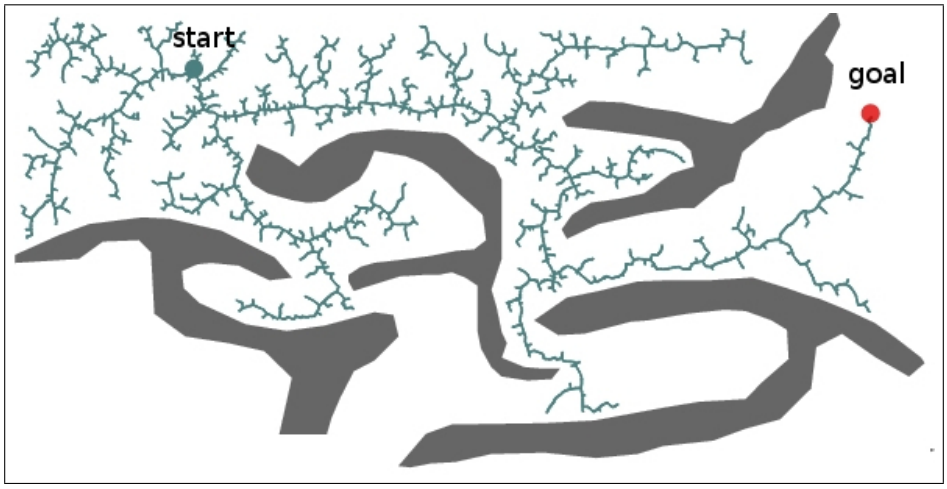


Figure 6: The tree built by RRT is shown here on a simplistic environment.

lective retraction in order to facilitate expansions inside narrow passages. The utility-guided RRT (UGRRT) (Burns & Brock, 2007) associates a utility measure with each vertex and uses it to promote expansions that increase the utility. RRT-Blossom (Kalisiak & van de Panne, 2006) creates a flood-fill behavior to locally explore the area surrounding each vertex. Machine learning has also been used to derive a distance metric that captures the cost-to-go in order to improve the exploration in RRT (Palmieri & Arras, 2015). The reachable volume RRT (RVRRT) (McMahon, Thomas, & Amato, 2015) relies on the notion of reachable volumes in order to restrict sampling to feasible regions and improve the performance of RRT on highly-constrained problems. The abstraction-guided RRT (fRRT) (Kiesel, Burns, & Ruml, 2012) uses A* search on a grid-based decomposition to bias RRT sampling towards low-cost regions. RRT* (Karaman & Frazzoli, 2011) rewires the branches in RRT to find optimal solutions with respect to a given cost function.

EST (Hsu et al., 2002) takes a different approach from RRT by pushing the frontier of the tree towards unexplored areas. Instead of relying on expansions from the nearest neighbor, EST maintains a probability distribution over the tree vertices. At each iteration, a vertex v is selected with probability inversely proportional to the density of a small neighborhood around v. This allows EST to push the tree towards less-explored regions of the configuration space.

3.3.2 HIERARCHICAL TREE-BASED MOTION PLANNING METHODS

Hierarchical tree-based methods rely on a scheme which first selects a region and then a vertex from which to expand the tree. Regions are often defined based on a decomposition of a low-dimensional projection of the configuration space. The rationale is that, by grouping similar vertices, better selections can be made at the region level to effectively guide the tree exploration. For instance, the single-query, bidirectional, lazy collision-checking (SBL) method (Sánchez & Latombe, 2002) pushes the tree toward sparse regions by using a grid-based decomposition and uniform probability distributions to select non-empty grid cells. The kinodynamic planning by interior-exterior cell exploration) (KPIECE) method (Sucan & Kavraki, 2012) relies on a multi-level grid decomposition constructed over user-defined or random linear projections. The synergistic combination of layers of planning (SyCLoP) method (Plaku, Kavraki, & Vardi, 2010) uses discrete search over a

low-dimensional triangular or grid decomposition to guide the tree exploration along short region paths to the goal. The guided sampling tree (GUST) method (Plaku, 2015) partitions the motion tree into equivalence classes and relies on multi-objective criteria based on shortest-path distances, selection penalties, and progress made to determine equivalence classes which could result in rapid expansions toward the goal.

The path-directed subdivision tree (PDST) method (Ladd & Kavraki, 2004, 2005) relies on a grid subdivision of the configuration space. Each cell in the decomposition keeps track of the tree branches that fall into it. Initially, the tree has only the root vertex and there is only one cell corresponding to the minimum and maximum values for each variable. At each iteration, a tree branch is selected for expansion. The cell c that contains the selected tree branch is divided into two cells, c_1 and c_2, along the largest dimension. The tree branches in c are also split according to the boundaries between c_1 and c_2. This ensures the invariant that a tree branch is contained entirely in one cell. The selected tree branch b is expanded by picking a configuration q along b and using propagation to add a new branch starting at q. Propagation is problem-dependent and could correspond to moving in a random direction.

The propagation continues until a collision is found or a maximum number of steps is reached. The branch is split when it exits the cell boundaries. A key component of the PDST is the weighting scheme associated with each cell based on its volume, number of branches, and number of previous selections. When selecting a tree branch, in order to increase the coverage, priority is given to cells that have large volumes but have not been well-explored. After each iteration, the selected cell c is penalized in order to ensure that other cells will eventually be selected for expansion. This is necessary to avoid oversampling and guarantee probabilistic completeness (Ladd & Kavraki, 2005). PDST has also been combined with artificial potential fields in order to expand the tree from the region with the lowest potential (Bekris & Kavraki, 2007).

3.4 Roadmap-Based Robot Motion Planning Methods

The introduction of PRM (Kavraki et al., 1996) shifted the focus from complete to probabilistically-complete motion-planning algorithms, which guarantee to find a solution, when it exists, with probability approaching one as time tends to infinity. While complete algorithms were limited to simple problems with $2-3$ variables, PRM made it possible to efficiently solve high-dimensional problems. The underlying idea in PRM is to construct a roadmap which captures the connectivity of the (obstacle-)free configuration space. The roadmap is populated by generating a number of collision-free configurations. Each configuration is generated by sampling values for the variables uniformly at random. The configuration is discarded, if it results in collision. Otherwise, it is added to the roadmap. To capture the connectivity, neighboring roadmap configurations are connected with collision-free paths. Figure 7 provides an illustration of a roadmap created by PRM. A common approach is to compute for each roadmap configuration its k-nearest neighbors according to a distance metric. A path between two configurations is often obtained by linear interpolation. If the path is collision free, it is added as an edge to the roadmap graph. A path from a given start to a goal configuration is found by first connecting the start and the goal configurations to the roadmap and then searching the roadmap graph. A* is often used to efficiently compute the shortest roadmap path. Additional sampling may be required when the initial roadmap does not contain a path from the start to the goal.

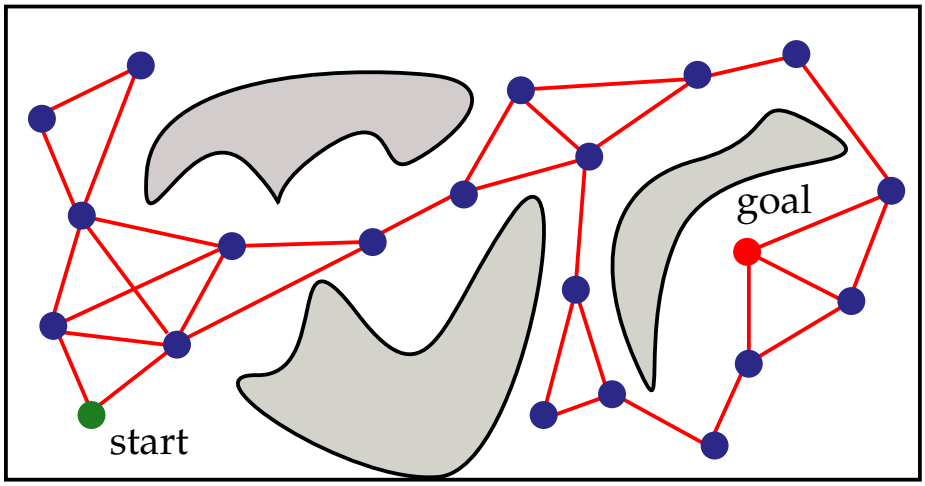


Figure 7: An illustration of a roadmap that can be used to answer start-to-goal queries.

Over the years, numerous strategies have been proposed to enhance the sampling in PRM. Obstacle-based PRM (OBPRM) (Amato, Bayazit, Dale, Jones, & Vallejo, 1998) seeks to increase sampling near obstacles in order to improve the connectivity inside narrow passages. BridgePRM (Sun, Hsu, Jiang, Kurniawati, & Reif, 2005) has a similar objective but uses a bridge test to generate samples halfway between two obstacles. Machine learning has also been used in conjunction with a portfolio of samplers to enhance sampling in narrow passages (Hsu, Sánchez-Ante, & Sun, 2005). The region-sensitive adaptive PRM (RESAMPL) (Rodriguez, Thomas, Pearce, & Amato, 2006a) uses the notion of entropy to identify regions that can enhance sampling. TogglePRM (Denny & Amato, 2013) switches between the free configuration space and the obstacle space in order to facilitate connections in narrow passages. ANC-Spatial (Ekenna, Thomas, & Amato, 2016) uses a spatial-learning approach to enhance the roadmap connectivity by determining appropriate connection methods for each roadmap vertex. PRM* (Karaman & Frazzoli, 2011) is a variant of PRM that leads to optimal solutions with respect to a given cost function. The modification is surprisingly simple, as it requires only using a variable number of nearest neighbors instead of a fixed k.

3.5 Biomolecular Modeling Challenges for Tree- and Roadmap-Based Methods

Both tree- and roadmap-based methods experience the curse of dimensionality in several ways. A central issue concerns the breadth of sampling in possibly high-dimensional and complex variable spaces. In particular, in the context of biomolecular modeling, the decision of which variables to represent is key, as it determines the dimensionality and complexity of the variable/conformation space. This decision is tightly tied with the application setting or the class of biomolecular systems considered. As reviewed in Section 2, while in many adaptations the selected variables are all or a subset of the backbone dihedral angles (Han & Amato, 2001; Amato et al., 2003; Song & Amato, 2004; Thomas et al., 2005, 2007; Jaillet et al., 2008; Tang et al., 2008; Tapia, Tang, Thomas, & Amato, 2007; Tapia et al., 2010; Jaillet et al., 2011; Shehu & Olson, 2010; Molloy et al., 2013; Molloy & Shehu, 2013), in others the selected variables capture collective motions of atoms in the 3d Cartesian space (Kim et al., 2002b, 2002a; Kirillova et al., 2008; Schuyler et al., 2009; Al-Bluwi et al., 2013; Maximova et al., 2015). Whether values for variables are sampled individually or in

tandem from *a-priori* compiled databases of "good moves" (Shehu & Olson, 2010; Molloy et al., 2013; Molloy & Shehu, 2013), or whether diverse perturbation/sampling operators are employed that make use of different sets of variables (Gipson, Moll, & Kavraki, 2013; Molloy & Shehu, 2016), the dimensionality and size of the variable space remains a key challenge for tree- and roadmap-based treatments of biomolecules.

The choice of variables is key to the design of effective sampling or perturbation operators for generating conformations that satisfy a set of desired geometric and/or energetic constraints for the biomolecular modeling problem at hand. Samples obtained uniformly at random have very low probability of being low-energy or in the region of interest for a sought structural transition. In particular, for long protein chains with hundreds or more backbone dihedral angles, a conformation sampled at random is highly unlikely to be physically-realistic.

Biased sampling techniques can be used to remedy this issue (Amato et al., 2003; Song & Amato, 2004), but it is hard to know *a priori* which perturbation operators will be effective. Recent work recognizes this issue and addresses it by offering diverse sampling operators on possibly diverse sets of variables (Raveh et al., 2009; Gipson et al., 2013; Molloy & Shehu, 2016). In particular, the work of Molloy and Shehu (2016) implements a probabilistic scheme that selects among a rich menu of operators making use of angular or Cartesian variables.

It is worth noting that sampling operators may generate samples not in a vacuum but by incremental modifications of existing samples. While earlier generations of tree- and roadmap-based methods typically obtained new conformations by sampling values over the selected variables (Singh et al., 1999), recent methods generate samples in neighborhoods of existing "parent" samples (Shehu & Olson, 2010; Molloy & Shehu, 2013; Maximova et al., 2016) (hence, the often used term perturbation operator). This later strategy has a higher chance of yielding physically-realistic conformations, as perturbation operators that perturb a selected sample to obtain a new one tend to preserve some good structural features in the new sample while introducing enough change to explore new regions of the variable space (Olson, Hashmi, Molloy, & Shehu, 2012). In the context of perturbation operators, selection schemes are critical to control sampling. A recent phenomenon in robotics-inspired methods has been the recognition that selection schemes, which are central to hierarchical tree-based robot motion planning (as reviewed above), can also be employed in both tree- and roadmap-based methods to steer biomolecular sampling to regions of interest (Shehu & Olson, 2010; Molloy & Shehu, 2013; Maximova et al., 2016).

An additional challenge with ensuring high sampling capability is that biomolecules have underlying complex energy surfaces that encode energetic constraints. Therefore, the criterion for accepting a sampled conformation and adding it to the vertex list of the tree or roadmap needs to either rely on an *a-priori* set energy threshold or be probabilistic in nature. The latter setting provides a balance between obtaining low-energy conformations while allowing a particular algorithm to go over high-energy barriers as needed to sample more of the conformation space (Jaillet et al., 2008, 2011; Molloy & Shehu, 2013; Devaurs et al., 2015; Molloy & Shehu, 2016).

Both roadmap- and tree-based methods rely on local planners or local deformation techniques to connect neighboring conformations. In tree-based methods that push the tree out in the variable space by generating child conformations from selected parent conformations via perturbation operators, the child becomes a neighbor of the selected parent; the neighborhood function does not rely on the notion of a distance in the variable space but is instead tied to the parent-child relationship. In other methods, a sampled conformation needs to be connected to one or more nearest neighbors. Nearest-neighbor calculations need specification of a distance function between conforma-

tions, which is non-trivial in high-dimensional spaces. Most adaptations employ least Root-Mean-Squared-Deviation (lRMSD), which is a modification of Euclidean distance after differences due to rigid-body motions have been removed through optimal superimposition of the protein conformations under comparison (McLachlan, 1972). lRMSD is carried out over Cartesian coordinate-based instantiations. Other distance functions use L1 or related variants defined over dihedral angles.

It is generally challenging to find computationally-efficient and dynamics-integrating local planners for biomolecular conformations. If conformations q_1 and q_2 are nearby in variable or structure space, the local path that is encoded with an edge in the tree or roadmap should encode the process of diffusion. The local path should provide evidence of the diffusion of the biomolecule from q_1 to q_2 in the presence of thermal vibrations. It is not readily obvious how to use dynamics to steer a biomolecule from q_1 to q_2. While in principle MD simulations can be employed in search of such paths, there is no guarantee that these simulations will reach q_2. While biased MD simulations can be employed to reach q_2, such simulations modify the energy surface and do not model the actual dynamics (Ma & Karplus, 1997). Moreover, any corrections to biased MD simulations to model the actual dynamics prove computationally expensive (Ovchinnikov & Karplus, 2012) in the context of robotics-inspired methods that evaluate thousands of edges.

As a result, the majority of robotics-inspired methods for biomolecular modeling employ local planners that carry out linear interpolations over the variables of the conformations that need to be connected via a tree branch or roadmap edge. Such planners produce unrealistic conformations, and significant time can be spent correcting geometry and other ensuing energetic violations via energy minimization. Recent work proposes complex, local planners that are not based on interpolation but are instead re-formulations of the motion computation problem; that is, the local planners themselves are tree- or roadmap-based methods (Molloy & Shehu, 2016). The latter idea is borrowed from the similarly challenging setting of motion planning for manipulators, where linear interpolation is also not effective (Nielsen & Kavraki, 2000). When making use of complex local planners, a prioritized path sampling scheme is needed to prioritize the application of these computationally-demanding planners on the most promising paths in order to control computational cost (Nielsen & Kavraki, 2000). The work of Molloy and Shehu (2016) provides an implementation of prioritized path sampling for biomolecular modeling.

4. Robotics-Inspired Methods for Equilibrium Biomolecular Structure and Dynamics

Table 1 categorizes different robotics-inspired methods by the robot motion planning frameworks they adapt and the application setups they address. The table is not comprehensive by any means, but it may be useful to readers selecting to focus on specific applications. In the rest of this Section we describe these methods in greater detail, paying particular attention to recent, state-of-the-art methods that showcase the current capabilities of robotics-inspired treatments of biomolecules.

4.1 Tree-Based Methods for Modeling Equilibrium Biomolecular Structure and Dynamics

Tree-based methods have been employed to model biomolecular flexibility and compute conformation paths connecting given structures (Cortés et al., 2005; Shehu, 2009; Shehu & Olson, 2010; Jaillet et al., 2011; Haspel et al., 2010). Some tree-based methods address decoy sampling for the *de novo* protein structure prediction problem (Shehu & Olson, 2010; Olson, Molloy, Hendi, & Shehu, 2012; Molloy et al., 2013) and map the entire energy landscape and pathways connecting stable

states in molecular loops, peptides, and proteins (Porta, Thomas, Corcho, Canto, & Perez, 2007; Jaillet et al., 2011; Porta & Jaillet, 2013; Devaurs et al., 2015; Molloy et al., 2016). Others have focused on specific flexible sub-chains, such as loops, rather than entire protein chains (Cortés et al., 2004, 2005; Yao, Dhanik, Marz, Propper, Kou, Liu, van den Bedem, Latombe, Halperin-Landsberg, & Altman, 2008; Barbe, Cortés, Siméon, Monsan, Remaud-Siméon, & Andre, 2011).

Table 1: Categorization of of Tree- and Roadmap-based Methods by Application Setting

Application	Tree-based Methods	Roadmap-based Methods
Protein Loop Motions	RLG-RRT (Cortés et al., 2005; Cortés, Jaillet, & Siméon, 2007), ML-RRT (Barbe et al., 2011)	LoopTK (Yao et al., 2008)
Protein-Ligand Binding	ML-RRT (Cortés et al., 2007)	PCR (Singh et al., 1999; Apaydin et al., 2001), SRS (Apaydin et al., 2003)
Protein Structure Prediction	FeLTr (Shehu & Olson, 2010; Molloy et al., 2013)	
Protein and RNA (Un)Folding		SRS (Apaydin et al., 2003), PRM-FP (Amato et al., 2003; Song & Amato, 2004; Tang, Kirkpatrick, Thomas, Song, & Amato, 2005; Thomas et al., 2005, 2007; Tapia et al., 2007; Tang et al., 2008; Tapia et al., 2010), MaxFlux-PRM (Yang, Wu, Li, Han, & Huo, 2007; Li, Yang, Han, & Huo, 2008)
Peptide and Protein Structural Transitions	NMA-RRT (Kirillova et al., 2008; Al-Bluwi et al., 2013), PathRover (Enosh, Raveh, Furman-Schueler, Halperin, & Ben-Tal, 2008; Raveh et al., 2009), T-RRT (Jaillet et al., 2011), PDST (Haspel et al., 2010), Sprint (Molloy & Shehu, 2013), Multi-T-RRT (Devaurs et al., 2015)	SRS (Molloy et al., 2016), Spiral (Molloy & Shehu, 2016), SoPRIM (Maximova et al., 2015)
Peptide and Protein Energy landscape Mapping	T-RRT (Jaillet et al., 2011), Multi-T-RRT (Devaurs et al., 2015)	SoPRIM (Maximova et al., 2015)

4.1.1 MODELING PROTEIN LOOP MOTIONS AND PROTEIN-LIGAND DISASSOCIATION

Early adaptations of the RRT algorithm for biomolecules focused on modeling the equilibrium dynamics of protein loops (Cortés et al., 2005) and protein-ligand interactions (Cortés et al., 2004). For instance, the method presented in the work of Cortés et al. (2005) proceeds in two stages to model large-amplitude structural changes in a protein loop at equilibrium. The first stage obtains an ensemble of collision-free conformations of a loop in a protein structure. This is achieved through the Random Loop Generator RRT (RLG-RRT) algorithm, which effectively samples the loop conformation space that satisfies kinematic closure constraints. The loop is divided into an active and a passive part. The passive part is selected to contain 6 variables, whereas the active part contains the rest of the angular variables selected to represent a loop conformation. The RLG-RRT algorithm

directly samples values for the variables in the active part via a scheme that increases the probability of obtaining a conformation that satisfies loop closure. Once the active part of a loop conformation has been obtained, an exact 6R inverse kinematics technique is applied to solve for the passive variables under the loop closure constraints. Closed loop conformations are added to the tree if they are collision-free. The tree is run for a fixed amount of time, and a goal region is defined around the given goal conformation in order to extract many paths from one execution of the RLG-RRT algorithm. Conformations in extracted paths are then subjected to short energetic minimizations in order to elucidate energetically-feasible, large-amplitude motions of the loop under investigation.

Analysis of extracted loop motions in the work of Cortés et al. (2004) reveals that results are comparable with classic molecular modeling methods while obtained with a performance gain of several orders of magnitude. The work of Cortés et al. (2004) demonstrates the efficacy of the two-stage approach for studying the activity-regulating mobility of the 17-residue long loop 7 in the amylosucrase enzyme from Neisseria polysaccharea.

Ideas similar to RLG-RRT are employed in the LoopTK (toolkit) algorithm (Yao et al., 2008) to explore the closed, collision-free conformations of flexible loops ranging in length for 5 to 25 amino acids. The algorithm relies on an interplay of sampling and deformation to obtain loops that satisfy the kinematic closure constraints and are collision-free. The sampling procedure focuses on obtaining geometrically-diverse, closed loops. The deformation procedure is based on earlier related work on loop modeling (Lotan, van den Bedem, Deacon, & Latombe, 2004; van den Bedem, Lotan, Latombe, & Deacon, 2005). The procedure makes use of the null space technique to explore the self-motion manifold (the constrained, closure space) around a closed loop to resolve steric clashes while not violating the closure constraints. LoopTK is shown to efficiently handle long loops up to 25 amino acids and even generate biologically-interesting, calcium-binding conformations. The toolkit is available at https://simtk.org/home/looptk.

The time demands of RRT-RLG on problems with hundreds of variables are addressed by Cortés et al. (2007) by proposing the Manhattan-like RRT (ML-RRT) algorithm to efficiently compute paths for a small protein-bound ligand to exit the protein active site. ML-RRT borrows ideas from mechanical disassembly and divides the variables into two groups, active and passive. In particular, the variables that model the internal and rigid-body motions of the ligand are designated as active, and the subspace of these active variables is sampled as in the RLG-RRT algorithm. The variables that model the internal motions of amino acids on the protein receptor's active site are designated as passive, and they are slightly perturbed if they hinder motions of the ligand. This decoupling proves effective, as it allows for possible ensuing collisions between the ligand and the protein to be addressed in a domino-like scheme, as illustrated in Figure 8.

ML-RRT has been shown to efficiently model motions of small ligands, side chains, loops, and backbone (Cortés, Le, Lehl, & Siméon, 2010; Barbe et al., 2011). The work of Cortés et al. (2010) subjects paths extracted from executions of the ML-RRT algorithm to a randomized path smoothing post-processing technique. The technique is carried out in the composite space of all the parameters resulting in simultaneous motions of the ligand and the protein in the final path. The work of Barbe et al. (2011) subjects loop conformations in paths extracted from ML-RRT to minimization of an MM energy function so as to reveal critical, physically-realistic intermediate conformations and bottlenecks along the open-to-closed loop motion of the Burkholderia cepacia lipase lid domain.

Adaptations of robot motion planning frameworks to model loop structures and motions represent only a fraction of diverse methods designed for loop modeling. Interested readers are referred to a survey in the work of Shehu and Kavraki (2012).

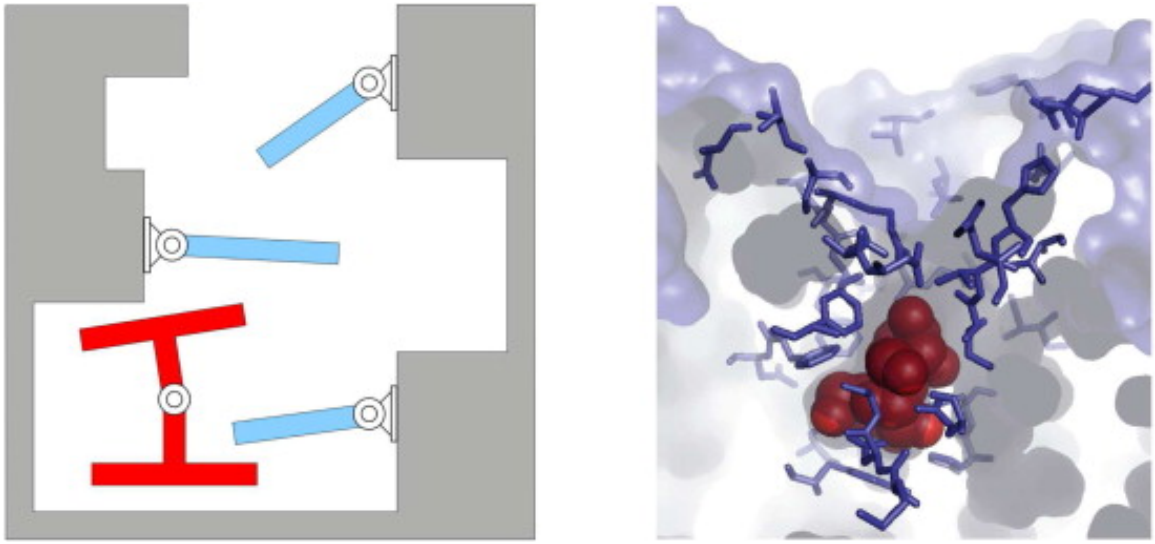


Figure 8: This figure is reproduced from the work of Al-Bluwi et al. (2012). The left panel illustrates the disassembly planning problem for two articulated objects. The ML-RRT algorithm proposed in the work of Cortés et al. (2007) problem models the escape of a ligand from a protein's binding site as disassembly problem. The red H-shaped object on the left image can be considered as the ligand in the right image, and the blue sticks in the left image can be considered as the flexible side chains on the binding site of the receptor protein in the right image. The figure is reproduced with permission of the Computer Science Review Journal.

4.1.2 MODELING PEPTIDE AND PROTEIN STRUCTURAL TRANSITIONS AND MAPPING PEPTIDE ENERGY LANDSCAPES

The RLG-RRT algorithm is modified in the work of Enosh et al. (2008) to model structural transitions in proteins and, in particular, model open and close motions in potassium channels. The main modification to the RLG-RRT algorithm by Enosh et al. concerns the addition of an energetic test to the collision-free test performed before deciding whether a generated conformation should be added to the tree. Several novel analysis techniques are introduced. Clustering is conducted over many paths obtained from several executions of the algorithm in order to identify common intermediate conformations in the paths connecting given start and goal conformations. Path alignment is employed to obtain the most energetically-favored path among all those computed. A schematic of the method proposed by Enosh et al. and a visualization of the most energetically-favored path obtained on the KscA protein are shown in Figure 9.

Another extension of the RRT-based algorithm in the work of Enosh et al. (2008) is presented by Raveh et al. (2009); the more efficient PathRover algorithm is proposed for modeling structural transitions on many proteins. PathRover achieves its computational efficiency in two main ways. First, its application on many proteins is made possible by restricting the number of dihedral angles used as variables. Three strategies are used to identify a subset of dihedral angles to define the variable space: careful inspection of structures, relevant literature, computational tools for detecting hinge regions like NMA, and comparison of structural changes in alternative (native or homologue) structures. In particular, the FlexProt alignment algorithm (Shatsky, Nussinov, & Wolfson, 2002) is used to compare start and goal structures and reveal structurally-different regions. Variables are manually restricted to dihedral angles in such regions. Second, PathRover limits the exploration to

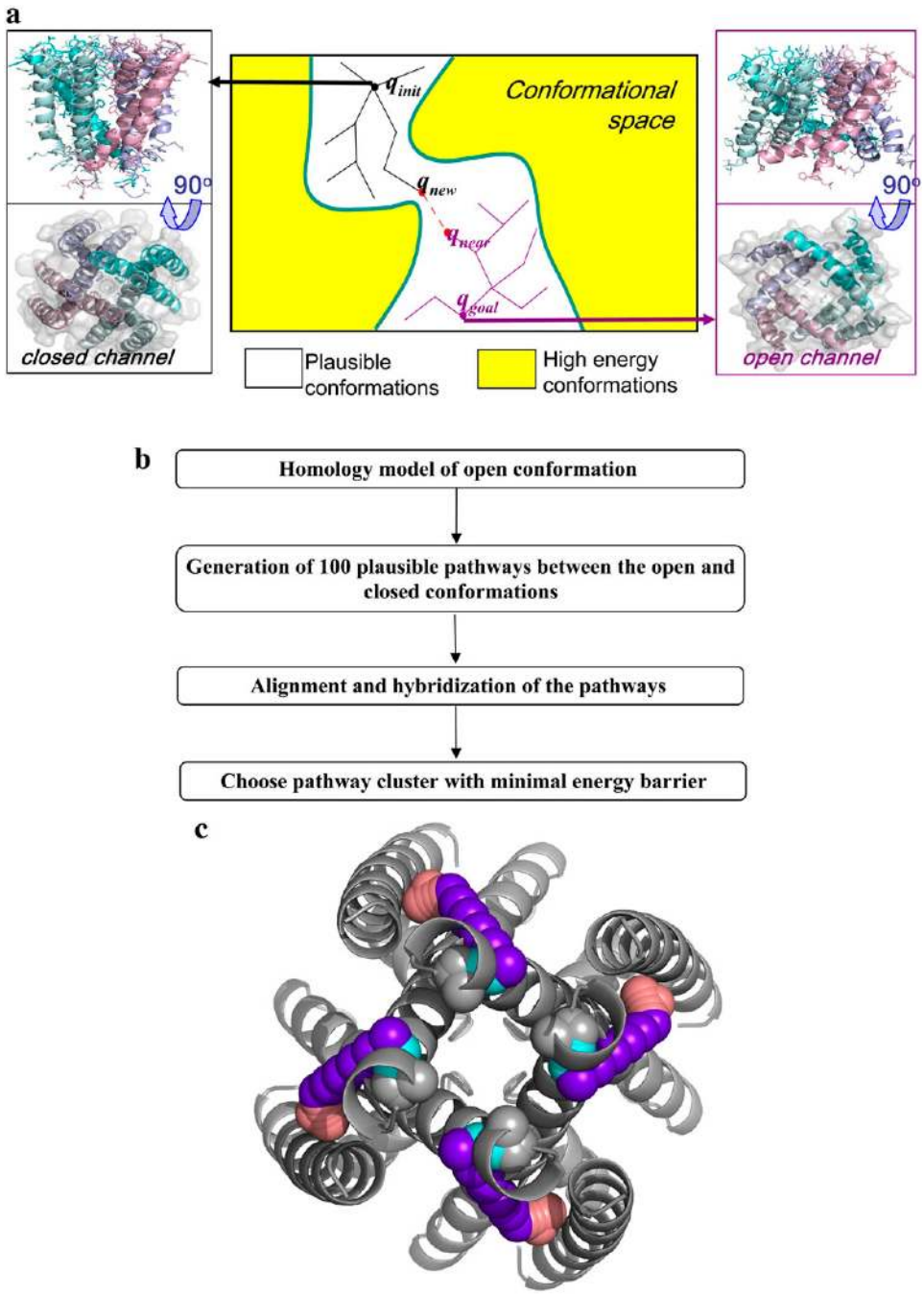


Figure 9: This figure is reproduced from the work of Enosh et al. (2008). The schematic of the method is shown in (a). The PathRover algorithm is executed multiple times to extract 100 plausible paths connecting the given open and closed conformations. The paths are clustered and aligned to reveal a path cluster with minimal energy barrier. This cluster is visualized for the KscA protein in (b), which shows a putative three-phase motion between the close and open conformations. This figure is reproduced with permission of the Biophysical Journal.

regions of the space consistent with available wet-laboratory data. Branch termination criteria are employed to stop the tree from being pulled towards regions that do not improve agreement with wet-laboratory data. The integration of wet-laboratory data aims to circumvent known inaccuracies of modern biomolecular energy functions.

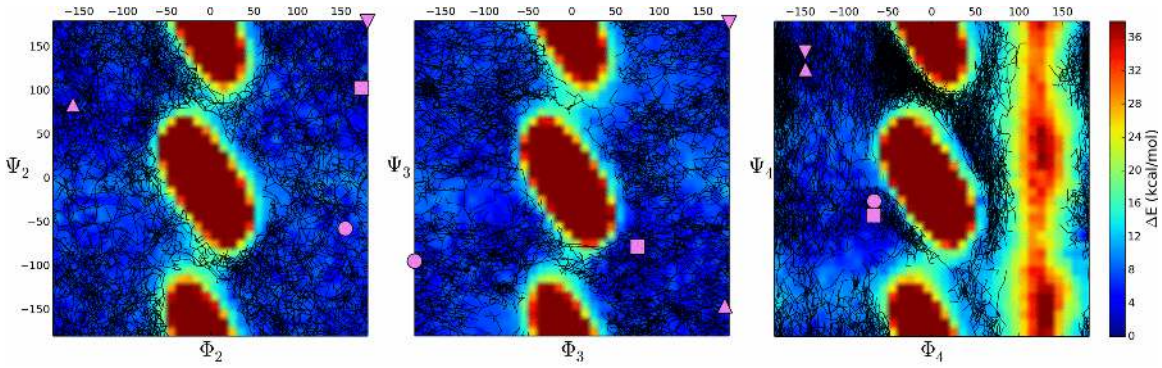
The RRT-based algorithms summarized above demonstrate the utility of RRT for modeling structural transitions in proteins. In particular, the PathRover algorithm utilized several schemes to reduce the number of variables in order to make the problem of modeling structural transitions tractable (Raveh et al., 2009). The work of Haspel et al. (2010) continued in this spirit via clever, reduced representations of large proteins. In contrast, work in the Simeon and Cortes labs credited with introducing RRT-based algorithms to biomolecular modeling focused instead on techniques to enhance sampling. The Transition-RRT (T-RRT) algorithm proposed by Jaillet et al. (2008) was shown particularly effective in this regard.

T-RRT and its bi- and multi-tree variants have been recently proposed to explore and obtain comprehensive maps of energy landscapes of small peptides, such as dialanine and Met-Enkephalin (Jaillet et al., 2011; Devaurs et al., 2015). The main modification to the baseline RRT algorithm in T-RRT concerns the introduction in the acceptance criterion of a state transition test based on the Metropolis criterion. New conformations are added to the tree if they pass the transition test (hence the name, T-RRT). The goal in T-RRT is to steer the tree towards exploration of low-energy regions in order to map energy minima in the potential energy surface while relaxing the transition test as needed to cross energy barriers that may trap the exploration to a particular local minimum.

The dynamic modification of the state transition test makes use of a reactive temperature scheme. In the Metropolis criterion, an effective temperature effectively controls the height of energy barriers that can be crossed by two consecutive conformations. In T-RRT, this temperature is increased when the number of attempts to pull the tree towards low-energy regions reaches a user-specified threshold; that is, the number of failures to grow the tree is taken as an indication of the presence of an energy barrier, and the effective temperature is increased in order to relax the state transition test. As soon as a successful edge is added to the tree, the temperature is then lowered by a pre-specified factor in order to resume the overall bias of pulling the tree towards local minima. The effect of this reactive temperature scheme is that the search is balanced between unexplored regions and low-energy regions of the variable space. Application of T-RRT in the work of Jaillet et al. (2011) shows that the algorithm can map the entire known energy landscape of the dialanine peptide when run in an exploration mode. Another setting, where T-RRT is used to obtain paths that connect discovered minima, also shows that recovered transitions between known stable states of dialanine are in strong agreement with transitions known from experiment and affirmed in other simulation studies.

Further work addresses the issue of limited sampling when the goal is to obtain accurate representations of energy landscapes of longer peptides, such as Met-Enkephalin (Devaurs et al., 2015). The T-RRT algorithm is used by Devaurs et al. only to reveal conformation paths connecting already-identified meta-stable states. These states are identified by an EA known as Basin Hopping (BH), which has been shown to effectively sample local minima of the energy surfaces of biomolecules (Olson et al., 2012). In the work of Devaurs et al., BH operates over dihedral angles and provides a sample-based, discrete representation of the energy surface of a peptide. The local minima are clustered to reveal wide basins corresponding to meta-stable states.

A variant of the T-RRT algorithm, referred to as Multi-T-RRT, is also proposed by Devaurs et al. to connect all identified states. The algorithm builds n single trees, each rooted at a conformation representative of a unique meta-stable state. The algorithm proceeds in iterations, at each iteration



Symbol	Conformation
■	
●	
▲	
▼	

Pathway	Cost
■ → ▼	48.2
■ → ▲ → ▼	178.1
■ → ● → ▼	219.2
■ → ● → ▲ → ▼	253.4
● → ▼	93.7
● → ▲ → ▼	127.9
● → ■ → ▼	172.0
● → ■ → ▲ → ▼	301.9

Figure 10: In this figure, reproduced from the work of Devaurs et al. (2015), the graph obtained from a single run of multi-T-RRT with cycles is projected over key dihedral angles in the top panel. The algorithm is seeded with four meta-stable states of the Met-Enkephalin peptide, drawn as pink triangles, squares, or circles. The states capture the unfolded state, the folded state, and two intermediates (shown in the bottom panel). Costs of various paths are shown in the table on the bottom panel. This figure is reproduced with permission of IEEE Trans Nano BioScience 2015.

randomly selecting one of the n trees for expansion with a conformation q. The conformation nearest to q in any of the other $n-1$ trees is identified, and if it is within an extension step-size, q merges two trees. Iterations continue until all trees are merged in a graph. The graph is queried for minimum-cost paths connecting any states of interest. The identified meta-stable states and the minimum-cost paths connecting these states are found to be comparable to those reported by other studies that employ more computationally-demanding exploration strategies (Devaurs et al., 2015). A representative result of the information that can be extracted from this combination of BH and Multi-T-RRT is shown in Figure 10.

One limitation of readily applying T-RRT and its variants to obtain similar, detailed characterization for proteins rather than short peptides is the dimensionality of the conformation space. In the work of Jaillet et al. (2011), this space has few dimensions due to the limited number of dihedral angles in small peptides, such as dialanine and Met-Enkephalin. Detail, while desired and possible on characterizations of short peptides, needs to be sacrificed in order to model large-scale structural transitions in proteins. NMA-RRT (Kirillova et al., 2008), PDST (Haspel et al., 2010), and Sprint (Molloy & Shehu, 2013) present three different algorithms that make use of representations of reduced detail to model large-scale structural transitions in proteins.

Normal Mode Analysis (NMA) is used to obtain larger-scale moves (low-frequency modes revealed by the NMA on a conformation) (Kirillova et al., 2008). These moves are then employed to generate new conformations in the RRT framework. The NMA-RRT algorithm proposed by Kirillova et al. (2008) essentially conducts the RRT search over a low-dimensional variable space of

the low-frequency modes. Since the normal modes provided by the NMA on a conformation only allow to get out of the local minimum represented by a conformation, NMA needs to be repeated regularly during the RRT search in order to explore the breadth of the conformation space. This can be computationally demanding, and application of NMA-RRT is limited to extraction of minimum-cost paths connecting two conformations of interest rather than a comprehensive map of the energy landscape and its connectivity in proteins. The work of Kirillova et al. shows that precious information can be extracted regarding structural transitions in proteins, such as adenylate kinase, even when focusing on motions largely driven by normal modes. A complementary study of minimal energy paths in adenylate kinase via NMA (Maragakis & Karplus, 2005) shows that the modes are sufficient to capture the structural transition between the open and closed structures in this protein; all known wet-lab structures are found to be within 3.0Åof these mode-based minimal energy pathways (Maragakis & Karplus, 2005).

A recent extension of NMA-RRT aims to reduce the computational demands of the algorithm. The extension employs a further reduced representation of a protein chain based on tripeptides and employs NMA on conformations of such reduced representations. The reactive temperature scheme in T-RRT is employed to broaden sampling and capture large-scale motions connecting significantly different structural states in large proteins of several hundred amino acids (Al-Bluwi et al., 2013).

Employing reduced representations has expanded the applicability of tree-based algorithms for treating large biomolecules. The PDST algorithm is adapted by Haspel et al. (2010) to model a transition between two structures of interest in large proteins of more than 200 amino acids. The assumption is made that secondary structures do not unfold in the sought transition, which is largely valid when modeling domain motions in proteins. Under this assumption, only backbone dihedral angles on loops connecting secondary structures are selected as variables.

In the work of Haspel et al. (2010), a bias scheme is used on 10% of the iterations to steer the tree towards the goal conformation. The bias scheme employs a Euclidean distance between feature vector representations of conformations in the tree. Given a conformation, its corresponding feature vector contains Euclidean distances between centers of mass of its secondary structure units. Conformations are evaluated through a detailed coarse-grained energy function that combines terms from the energy function used in the work of Shehu et al. (2009) and the Amber ff03 energy function. If a sampled conformation is evaluated to have energy above a set threshold of 100 kcal/mol more of the energy of the start conformation, the conformation is subjected to 20 steps of steepest descent and retained only if its energy decreases below this threshold. This is a rather coarse energetic constraint, but paths collected from 100 runs of the algorithm not only reach the goal conformation in less time than methods based on Simulated Annealing, but also reveal credible motions consistent with experimental data on large, well-characterized proteins such as GroEL (Haspel et al., 2010).

The Sprint algorithm proposed by Molloy and Shehu (2013) uses a complementary approach to simplifying the search space explored for paths connecting given structures of medium-size proteins. Sprint addresses the issue of sampling in high-dimensional variable spaces by employing a popular idea from *de novo* structure prediction. The fragment replacement technique is used to divide a protein chain into bundles of consecutive dihedral angles, and values for a bundle or fragment are sampled from an *a-priori* constructed database of fragment configurations on known, native protein structures. The fragment replacement technique is used to expand the tree at every iteration.

While the work of Molloy and Shehu (2013) adapts the EST framework via this expansion procedure to model structural transitions in small- and medium-size proteins, the Fragment Monte Carlo Tree Exploration (FeLTr) algorithm proposed by Shehu and Olson (2010) uses related con-

cepts to sample the space of near-native protein conformations for the purpose of *de novo* structure prediction in small proteins.

In both Sprint and FeLTr, a state-transition test is used to steer the tree towards low-energy conformations over time; that is, a probabilistic, Metropolis-like criterion is used to determine whether a child conformation should be added to the tree. While in FeLTr (Shehu & Olson, 2010) a fixed scaling parameter (analogous to a fixed effective temperature) is used in the Metropolis-like criterion, Sprint (Molloy & Shehu, 2013) integrates a reactive temperature scheme in order to allow the tree to go over high-energy regions when no low-energy routes can be found and thus expand exploration capability. The reactive temperature scheme in Sprint is slightly different from that in T-RRT (Jaillet et al., 2008). While in T-RRT the effective temperature is increased or decreased by fixed amounts, Sprint moves temperature along a proportional cooling scheme often employed in Simulated Annealing Monte Carlo methods (Shehu et al., 2009). Upon failures to expand the tree, temperature moves up to the next value in the cooling scheme; upon successes, temperature goes down to the next value in this scheme.

Both FeLTr and Sprint operationalize on the idea that it is easier to push rather than pull the tree in conformation space when good moves are available or compiled *a priori* to generate energetically-feasible child conformations from selected parent conformations in the tree. While in the work of Molloy and Shehu (2013) the tree is rooted at a given start conformation and the goal is to get within a tolerance region of the given goal conformation, in the work of Shehu and Olson (2010) the tree is rooted at an extended conformation, and the termination criterion is a compromise between a desired number of low-energy conformations and running time.

A central idea in both Sprint and FeLTr is that the growth of the tree can be controlled via a selection mechanism; at each iteration, a conformation in the tree is selected for expansion. The selection penalizes the tree from growing towards regions of the conformation space that have been oversampled, thus resulting in enhanced sampling of the conformation space. Two discretization layers are employed. In FeLTr, the first layer maps conformations in the tree on a 1d grid whose cells are energy levels of width 2 kcal/mol. In Sprint, the first layer maps conformations on a 1d grid based on their lRMSD from the goal conformation. The second layer in both algorithms maps conformations over a geometric projection. The second layer is a 3d grid, where conformations are associated with 3 shape-based global coordinates (Shehu & Olson, 2010).

The selection mechanism uses both discretization layers. First, it selects an energy level according to a probability distribution function. The latter is defined over weights associated with energy levels according to some weighting function. Different weighting functions are analyzed for how strong a global energetic bias needs to be in order to reproduce the native structure (Molloy et al., 2013). Once an energy level is selected, cells of the geometric projection grid that belong to conformations in the selected energy level are analyzed. A second weighting function over cells of the grid biases against selecting a cell that has been selected many times before and/or already has many conformations in it. Once a cell is selected, any conformation in it is selected for expansion uniformly at random, since conformations in a cell are energetically- and geometrically-indistinguishable.

Extensions of FeLTr have explored both the effect of different weighting functions over the discretization layers and the employment of different projection coordinates (Molloy et al., 2013; Olson et al., 2012). Different coarse-grained energy functions that are considered state-of-the-art in *de novo* structure prediction, including the Rosetta suite of energy functions, are employed in the framework and directly compared for how they steer the search towards near-native conformations (Molloy et al., 2013). Molloy and Shehu (2013) also investigate the impact of different

projection schemes and selection mechanisms on both the diversity and energetic profiles of Sprint-extracted paths in the context of computing structural transitions.

Applications of Sprint on different start and goal structure pairs of the calmodulin and adenylate kinase proteins show that the algorithm is able to find paths that reach the goal conformation (Molloy & Shehu, 2013). Soft global biasing schemes are found to provide the right compromise between tree depth (that is, lower energies) and diversity of paths (that is, geometrically-diverse conformations). Detailed energetic and structural analysis on computed paths for two hallmark proteins, such as calmodulin and adenylate kinase, reveals that Sprint yields accurate characterizations of structural transitions in these proteins. Energetic profiles of extracted paths indicate the presence of high-energy regions that need to be crossed on specific transitions in calmodulin, in agreement with wet-laboratory characterizations. Analysis on adenylate kinase shows that known intermediate structures of this protein are present in the the conformation paths computed by Sprint (Molloy & Shehu, 2013).

4.2 Roadmap-Based Methods for Modeling Equilibrium Biomolecular Structure and Dynamics

Roadmap-based methods have been employed to model protein-ligand binding (Singh et al., 1999), protein and RNA folding and unfolding (Song & Amato, 2004; Chiang et al., 2007; Chiang, Hsu, & C., 2010), and protein structural transitions (Molloy & Shehu, 2016; Maximova et al., 2015).

4.2.1 MODELING PROTEIN-LIGAND BINDING

The adaptation of the roadmap-based motion planning framework for protein-ligand binding by Singh et al. (1999) is the first occurrence of robotics-inspired treatments of biomolecular structure and dynamics. The adaptation was simplistic but provided key design issues that were then replicated and extended by many robotics researchers. One of the key simplifications is that the protein receptor is kept rigid, and the only variables of interest are those allowing to model rigid-body motions of the ligand around the receptor and internal motions of the ligand. Small ligands are considered, so that the $6 + p$ variables to allow modeling of such motions do not go above a dozen. Sampling proceeds uniformly at random over the $6 + p$ variables, but ligand configurations added to the roadmap pass a geometric and energetic criterion. The geometric criterion ensures that ligand configurations are within some predefined distance of the center of mass of the receptor.

The energetic criterion is probabilistic: two dynamically-updated thresholds, $E_{\min}$ and $E_{\max}$ values, corresponding to minimum and maximum energy values over sampled configurations, are recorded. Ligand configurations with energy higher than $E_{\max}$ are rejected. Other configurations are retained with probability $(E_{\max} - E(q))/(E_{\max} - E_{\min})$. The energy function incorporates both terms to evaluate the internal energy of a ligand configuration as well as terms evaluating interactions between a ligand configuration and the rigid protein receptor.

Retained ligand configurations are embedded in a nearest-neighbor graph, using lRMSD to measure the distance between two ligand configurations and a user-set parameter, k, for the number of nearest neighbors. A simple local planner interpolating over all $p + 6$ variables of two neighboring configurations is used to estimate the feasibility of $q \leftarrow q'$ and $q' \leftarrow q$ edges by generating consecutive configurations. Consecutive configurations q_i are generated by the linear interpolation planner to connect q and q' until the distance between two consecutive configurations in the generated series is no higher than 1Å. The $q \rightarrow q'$ and $q' \rightarrow q$ edges are added to the roadmap only if all q_i

configurations have energies below $E_{\max}$. Weights are added to retained edges as follows:

$$w(q \to q^{'}) = -\sum_{i=0}^{s-1} \log[P(q_i \to q_{i+1})]$$

where

$$P(q_i \to q_{i+1}) = \frac{e^{(E_{i+1}-E_i)/(K_B T)}}{\left(e^{(E_{i+1}-E_i)/(K_B T)} + e^{(E_{i-1}-E_i)/(K_B T)}\right)}$$

In the above equations, q_{i-1}, q_i, q_{i+1} are three consecutive configurations with corresponding energies E_{i-1}, E_i, E_{i+1}, K_B is the Boltzmann constant, and T is the effective temperature. The weight of a path $q_{\text{start}} \rightsquigarrow q_{\text{goal}}$, which connects a start configuration to a goal configuration, is then the sum of the weights of the edges in it. The weight of a path initiated at an unbound configuration and terminating in a bound configuration estimates the association rate (the cost of the ligand approaching and binding to the protein receptor). The weight of the reverse path estimates the disassociation rate (the cost of the ligand leaving the binding site and diffusing in space).

The resulting roadmap represents a distribution of energetically-credible paths of the ligand approaching and then binding the receptor. In the work of Singh et al. (1999), the bound configuration of the ligand in the $p+6$ variable space is presumed not to be known, and RMSD-based clustering of sampled lowest-energy ligand configurations is employed to reveal a few likely bound candidates. Analysis reveals that the true bound configuration is indeed present in the top-populated clusters; however, many false positives are reported, as well. Weights of paths terminating in and initiated from the lowest-energy ligand configurations are analyzed in order to determine what other characteristics can be used to discriminate between true and false positives. Paths terminating at the true, bound configurations are found to have high association rates; the reverse paths are found to have high disassociation rates. This important result elucidates that effective binders are not only those that allow the ligand to reach the lowest interaction energy but also trap it at the binding site via high-energy barriers.

4.2.2 MODELING PROTEIN AND RNA (UN)FOLDING

Singh et al. (1999) provided a much needed template and has served as the foundation for many robotics-inspired treatments of biomolecules. In particular, a suite of roadmap-based algorithms and extensions were designed in the Amato lab to model unfolding of small proteins. A review of roadmap-based methods to study molecular motions in the Amato lab is available in the work of Tapia et al. (2010), whereas a review of roadmap-based methods for the specific protein folding problem is presented in the work of Moll, Schwartz, and Kavraki (2008). A seminal contribution in this category is the Probabilistic Conformation Roadmap (PCR) algorithm (Apaydin et al., 2001), which builds upon the template presented by Singh et al. (1999) to study protein folding.

PCR addresses a complex application domain, as the number of variables needed to model the intrinsic flexibility of protein chains can easily reach 100 or more. In PCR and other extensions that followed, most notably in the Amato lab, the variables employed are all or a subset of the backbone dihedral angles of a protein chain (Amato et al., 2003; Song & Amato, 2004; Tang et al., 2005; Thomas et al., 2005, 2007; Tapia et al., 2007; Tang et al., 2008; Tapia et al., 2010). In such variable spaces, uniform random sampling is ineffective and likely to result in conformations with severe internal collisions. For this reason, work in the Amato lab on PCR-based algorithms has gradually

shifted to sampling strategies based on incremental perturbations of a given native/folded conformation until memory of the folded conformation has been lost. Specifically, backbone dihedral angles of the folded conformation are perturbed by small amounts by use of a Gaussian distribution until a minimum number of conformations is obtained for each category (0 to 100% in 10% increments) of the percentage of native contacts. The lower the number of native contacts is in a conformation, the more likely that conformation is to belong to the unfolded state. While the acceptance criterion for sampled conformations is as in the work of Singh et al. (1999), the energy function is different, as it measures the internal energy of a protein chain. The function contains terms favoring hydrogen bonds, disulfide bonds, and hydrophobic interactions.

The sampled conformations that pass the energetic/acceptance criterion are embedded in a nearest-neighbor graph, with the number of nearest neighbor conformation k specified by the user. In contrast to the original PCR algorithm, directed (u, v) edges in the graph are weighted based on the Boltzmann-related Metropolis criterionas in: $P_{(u,v)} = e^{\frac{E(u)-E(v)}{K_B \cdot T}}$, where $E(.)$ is the energy of a conformation, K_B is the Boltzmann constant, and T is an *a-priori* set temperature determining the height of energy barriers crossed by an edge. In this early formulation of edge weights, no reactive temperature schemes are employed as in the later tree- and roadmap-based algorithms for structural transitions. Instead, T is a user-controlled parameter that determines to a great extent the ability of the algorithm to navigate the underlying energy surface.

In the works of Song and Amato (2004) and Thomas et al. (2005), all N best paths that end at the folded conformation and start at conformations with 0 native contacts are extracted and analyzed. Analysis of such paths has shown that, despite several design decisions intended to simplify the protein folding problem, PCR-based algorithms can predict the order of secondary structure formation. Agreement with wet-laboratory data has validated the general usage of PCR-based algorithms to provide a coarse-grained treatment of folding and unfolding pathways for protein chains. Other works by Amato and collaborators also show the applicability of PCR-based algorithms to study RNA folding and unfolding (Tapia et al., 2007; Tang et al., 2008; Tapia et al., 2010) .

The sampling strategy of incremental perturbations is effective on protein chains of no more than 60 amino acids (Song & Amato, 2004) but scales poorly on longer chains (Thomas et al., 2005). Ensuing work improves sampling for protein chains up to 110 amino acids by further reducing the number of variables modeled to represent conformations (Thomas et al., 2007). Specifically, rigidity analysis is employed to detect least-constrained regions in a given structure. The dihedral angles belonging to such regions are selected more often for perturbation in the sampling stage. This modification is shown effective in revealing subtle folding differences between protein G and two of its sequence variants. In particular, the modification is also shown to be promising for capturing other dynamic events in proteins beyond folding to study large-scale conformational changes involved in structural transitions of the calmodulin protein. Related ideas have been employed by other researchers to compute temperature-dependent optimal folding paths in peptides and proteins (Yang et al., 2007; Li et al., 2008). The MaxFlux-PRM algorithm proposed by Yang et al. (2007) to study structural transitions in the dialanine peptide and folding of a β-hairpin is then shown by Li et al. (2008) to be capable of predicting folding pathways of the engrailed homeodomain protein.

Further work in the Amato lab has focused on exploiting the conformation roadmap to extract quantities summarizing folding kinetics in protein and RNA molecules. Tapia et al. (2007) introduce two new analysis techniques, the Map-based Master Equation (MME) and Map-based MC (MMC) technique. This work shows that treating the roadmap as a map of the folding landscape can be

exploited to estimate kinetic metrics that are typically only extracted from MD simulation studies. Metropolis MC simulations can be conducted over the roadmap, moving between roadmap vertices observing the edge probabilities in the roadmap. Different statistics can then be calculated over the MMC walks, including folding rates and population kinetics. Tang et al. (2008) show that statistics summarizing RNA folding predict well the same relative gene expression rate for wild-type MS2 phage RNA and three of its mutants, in good agreement with wet-laboratory data.

4.2.3 STOCHASTIC ROADMAP SIMULATION AND MARKOV STATE MODELS FOR MODELING PROTEIN AND RNA (UN)FOLDING AND PROTEIN STRUCTURAL TRANSITIONS

The idea that reliable statistics can be extracted from molecular conformation roadmaps was presented earlier by Latombe and colleagues (Apaydin et al., 2003). The stochastic roadmap simulation (SRS) framework is formalized in relation to a key analogy between a roadmap and a Markov state model (MSM); the concept of a stochastic roadmap with probabilistic edges was presented earlier (Song & Amato, 2000), but the analogy with an MSM went missing till the 2003 formalization by Latombe and colleagues (Apaydin et al., 2003). The latter laid bare the analogies between a stochastic roadmap and what would later be referred to as a point-based MSM. In such an MSM, states of the MSM are the single-conformation vertices of the stochastic roadmap, and the probabilistically-weighted edges connecting vertices in the roadmap are the state-to-state transitions in the MSM. The analogy brought to focus that a stochastic roadmap better encodes the stochastic nature of biomolecular motions, and the analogy with an MSM could even be used to extract interesting summary statistics regarding physics-driven stochastic processes.

In addition to recognizing that biased random walks can be carried out over the roadmap and employed to extract statistics of interest (Tapia et al., 2007), the SRS-MSM analogy highlights that effective, algebra-based techniques from (Markov chain) transition state theory can be employed to extract average statistics without launching a single simulation (or random walk over the roadmap). Folding rates, p_{fold} values, ϕ values, and other estimates of kinetics, such as transition rates, can be obtained without needing to perform many random walks but by in-order propagation of transition probabilities. The analogy between a stochastic roadmap and a point-based MSM is shown to result in correctly-predicted p_{fold} values on small proteins modeled at the secondary structure level with $6-12$ variables (Apaydin, Brutlag, Hsu, & Latombe, 2002; Apaydin et al., 2003). Further work demonstrated that the transition state ensemble (the set of conformations with $p_{\text{fold}}=0.5$), folding rates, and ϕ values could be predicted on 16 different proteins but at a fraction of the computational time that would be needed by a framework launching numerous MC simulations (Chiang, Apaydin, Brutlag, Hsu, & Latombe, 2006; Chiang et al., 2007) .

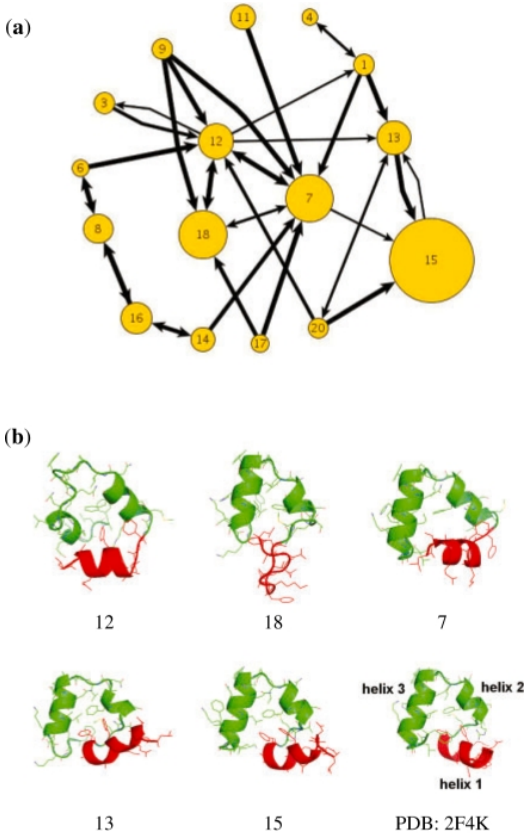
While the SRS-MSM analogy permits interesting mathematics, practical issues such as how to ensure that the transition matrix is not prohibitive in size to allow solving of linear algebra equations have to be addressed on a case-by-case basis. The formalization presented by Apaydin et al. (2003) did not discuss practical design decisions such as how to group conformations into states and how to estimate transition probabilities between two sub-ensembles, but rather on the mathematics that would be possible by the analogy between a stochastic roadmap and an MSM. Analogies with cell-based MSMs, where states are homogeneous sub-ensembles of conformations rather than single conformations need addressing practical issues regarding how to organize conformations into states and how to associate transition probabilities between states.

Since the seminal work of Apaydin et al. (2003), analogies between SRS and cell-based MSMs have been largely limited, partly due to the lack of clear objectives in design decisions that are general in their ability to transform a roadmap into an MSM of manageable size. For instance, the fundamental assumption was that if conformations were obtained via an MD simulation at some temperature T, then the probability of an edge representing a transition from a vertex u to a vertex v could be measured via the Boltzmann-related Metropolis criterion $e^{-(E(v)-E(u))/(K_B \cdot T)}$. This realization allowed Apaydin, Latombe, and colleagues to see the clear connection between the stochastic (probabilistic) roadmap of structures and a point-based MSM, with vertices seen as states of the MSM and edges between vertices in the roadmap as transitions between states in the MSM. However, practical considerations as to how to convert these single conformation vertex probabilities into state-state transition probabilities were not discussed.

The issue of how to associate probabilities in the first place to conformations sampled via other non-MD algorithms was also not discussed. Two groups of researchers have started operationalizing on the seminal ideas presented by Apaydin et al. (2003). Work by Latombe and colleagues has focused either on point-based MSMs or on summarizing and uncovering the MD-simulated dynamics of synthetic and small peptides via cell-based MSMs (Chiang et al., 2007, 2010). Complementary work in the Shehu lab has focused on non-MD approaches and extracting average statistics to model and compare transitions in healthy and aberrant forms of disease-participating, small- to medium-size proteins (Molloy et al., 2016).

Chiang et al. (2010) offer a novel representation of states not as individual conformations (Apaydin et al., 2003; Singhal, Snow, & Pande, 2004) or even disjoint regions of conformation space (Ozkan, Dill, & Bahar, 2002; Chodera, Singhal, Pande, Dill, & Swope, 2007) (as in cell-based MSMs) but instead as overlapping probabilistic distributions over the conformation space. This distribution relies on the key recognition that a single conformation does not contain enough information to be uniquely mapped to a state and leads to the presence of hidden states in what is referred to as a Markov Dynamics Model (MDM) rather than an MSM (Chiang et al., 2010). In the MDM, emission probabilities of hidden states measure the probability with which a conformation belongs to a state. Both transition and emission probabilities are estimated over trajectories of conformations obtained from many MD simulation trajectories. A principled criterion based on the ability of a model to predict long-timescale kinetics allows discriminating between possible MDMs and selecting an optimal one. The MDM embedded over conformations obtained from MD trajectories simulating folding of the fast-folding villin headpiece subdomain (HP-35 NleNle) is shown in Figure 11. The MDM presents a highly-interpretable discrete kinetic model of the folding of this small sub-domain, built over more than 400 MD trajectories, each 1μs long. Figure 11 shows that a few states, 7, 12, 13, 15 and 18, are the most frequently-visited states that significantly influence the long-term dynamics.

Molloy et al. (2016) present strategies to embed conformations sampled via a non-MD method in a cell-based MSM. The ability to formulate a cell-based MSM relies on dense sampling of the conformation space of interest. The latter provides significantly challenging to do in the MD setting or even in the robotics-inspired setting. Instead, complementary work in the Shehu lab on EAs is used to obtain a rich ensemble of local minima conformations in healthy and variant sequences of a given protein (Clausen & Shehu, 2015; Clausen et al., 2015). These conformations are organized into states via a simple lRMSD-based clustering scheme. A nearest-neighbor graph is then imposed over the states, but an additional lRMSD constraint is imposed so as to connect only nearby states via an edge. The assumption is made that transitions are possible between nearby states, and prob-

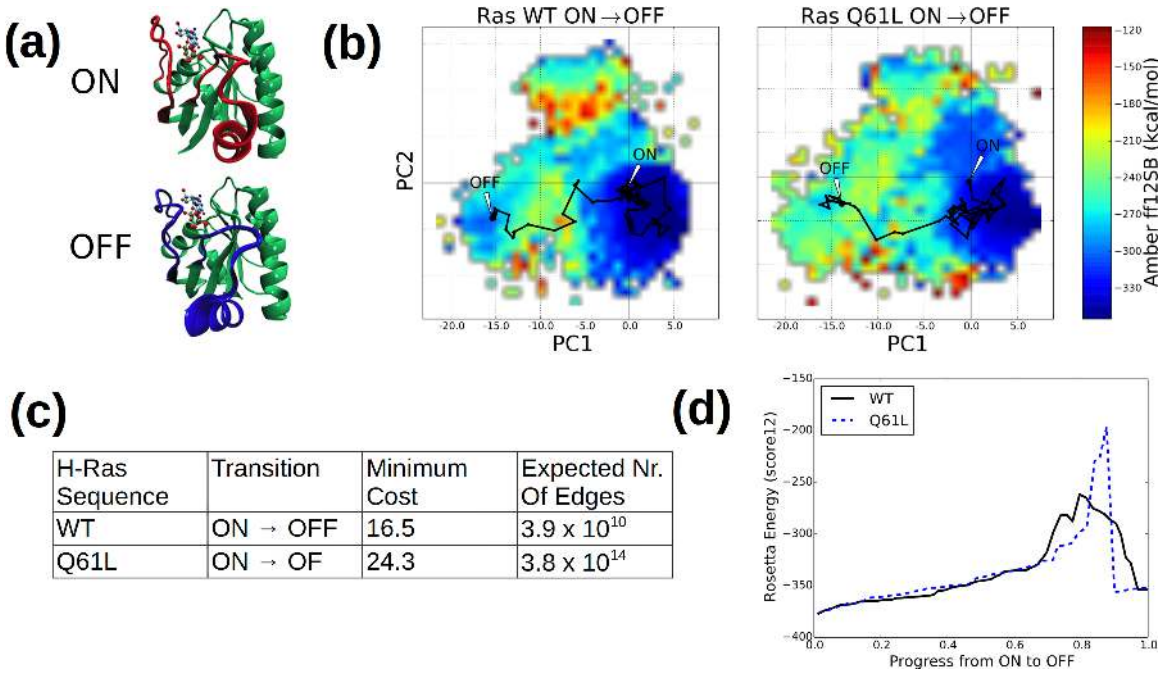


(c)

State Sequence	Prob.
$12 \rightarrow 7 \rightarrow 15$	0.0311
$12 \rightarrow 13 \rightarrow 15$	0.0294
$12 \rightarrow 1 \rightarrow 13 \rightarrow 15$	0.0076
$12 \rightarrow 18 \rightarrow 7 \rightarrow 15$	0.0052
$12 \rightarrow 18 \rightarrow 13 \rightarrow 15$	0.0047

Figure 11: In this figure, reproduced from the work of Chiang et al. (2010), (a) shows the MSM connecting twenty identified states of the villin headpiece peptide. The size of each node in the MSM is proportional to the probability of the corresponding state in the stationary distribution. The width of each edge is proportional to the transition probability between corresponding states. States with probability < 0.01 in the stationary distribution, self-transitions, and edges with transition probability < 0.002 are not drawn to avoid cluttering. The initial conformations are most likely to belong to state 12, and the native conformation is most likely to belong to state 15. (b) Representative conformations are shown from states 7, 12, 13, 15, and 18. The residues forming the important helix 1 in the villin headpiece peptide are drawn in red. (c) The most likely state transition sequences from states 12 to 15 are shown here. This figure is reproduced under the Bioinformatics Journal's terms of the Creative Commons Attribution Non-Commercial License.

abilities of such transitions can be estimated via a Boltzmann-like probability. The latter makes use of the concept of the energy of a state. Several schemes are employed to determine the energy of a state, ranging from the minimum to the average value over energies of conformations grouped in a state.



H-Ras Sequence	Transition	Minimum Cost	Expected Nr. Of Edges
WT	ON → OFF	16.5	3.9×10^{10}
Q61L	ON → OF	24.3	3.8×10^{14}

Figure 12: This figure is reproduced from the work of Molloy et al. (2016). Panel (a) shows two wet-laboratory structures representative of the ON and OFF structural states of the H-Ras catalytic domain. H-Ras switches between these two states to regulate its biological activity in the cell. The loop regions where the change is localized are shown in red and blue. The reactant (GTP) and product (GDP) are also drawn where they bind H-Ras. Panel (b) shows two-dimensional projections of the probed energy surface of the H-Ras wildtype (WT) and the oncogenic Q61L variant. Sampled conformations are projected on the top two principal components (PC) obtained via Principal Component Analysis of sampled conformations. The color-coding follows the Amber ff12SB internal energy values of the all-atom structure corresponding to each sampled conformation. The ON → OFF minimum-cost paths obtained by querying the stochastic roadmap constructed over sampled conformations are shown, as well. The costs of these paths are shown in the table in panel (c). The average number of edges over all possible ON → OFF paths are obtained by treating the roadmap as an MSM. The actual energy profiles of the minimum-cost paths obtained for the WT and Q61L variant are shown in panel (d). This figure is reproduced with permission of Robotica 2016.

The result of this process is a stochastic roadmap that can be used to answer lowest-cost path queries, as traditionally the case in roadmap-based methods, as well as yield average statistics, such as the average number of edges in a transition, via the analogy of the stochastic roadmap with an MSM. A path smoothing algorithm based on the conjugate peak refinement technique (Fischer & Karplus, 1992) provides more detail with state-state paths and improves their energetic profile. The average statistics, while not direct measurements of transition rates due to the lack of timescale information from non-MD methods, allow conducting comparisons between wildtype (WT) and variant (mutated) sequences of proteins of interest. In the work of Molloy et al. (2016), such statistics are employed to obtain a structural explanation for the role of specific mutations on the biological activity in two proteins implicated in human disorders. Figure 12 showcases some representative results from application of the SRS-based approach in the work of Molloy et al. to the WT and Q61L variant of the H-Ras protein. Ras sequence mutations have been implicated in various human cancers (Karnoub & Weinberg, 2008).

Comparison of the energy landscapes, costs of the minimum-cost paths, energy profiles of these paths, and the expected number in edges over all paths in each H-Ras sequence in Figure 12 provides a structural explanation for the impact of the Q61L mutation on the biological activity of this enzyme. The mutation introduces an energy barrier between the ON and OFF states, and this barrier increases both the cost of the minimum cost path and the number of expected edges in transition paths. Taken together, these results suggest that the Q61L mutation, while preserving the stability of the ON and OFF states, cannot form the transition state mimic, in agreement with wet-laboratory studies (Gremer, Gilsbach, Ahmadian, & Wittinghofer, 2008; Gibbs, Schaber, Allard, Sigal, & Scolnick, 1988).

4.2.4 ADDRESSING LIMITED SAMPLING IN ROADMAP-BASED METHODS FOR MODELING PROTEIN STRUCTURAL TRANSITIONS

Sampling remains a key issue in adaptations of roadmap-based methods for biomolecular modeling. While generally the focus of robotics-inspired methods has been on demonstrating their ability to reproduce experimental knowledge qualitatively or even quantitatively on specific systems under investigation, their general applicability has been largely sacrificed. For instance, the roadmap-based methods applied to model the discrete secondary structure formation events in protein folding and unfolding are largely not applicable to model folding or other transition events in proteins more than 150 amino acids where the states sought to the bridged by the transition may be farther than 10Å away from each-other. Strategies to reduce the number of variables so as to control the dimensionality of the variable space have important ramifications. For instance, rigidity-based techniques base their conclusions of where the most flexible regions are on analysis of a specific structure. NMA techniques suffer from a similar issue, and regular application of NMA on sampled conformations adds to the computational time demands of an algorithm. Other techniques that make assumptions on which regions do or do not participate in a particular transition event rule out the possibility of potentially complex, cooperative events. Others that bundle variables together and obtain values for them from pre-compiled databases make similar assumptions on what types of structural changes facilitate a transition.

Sampling will remain a challenge, but two complementary directions are being explored. The first direction values broad applicability over specific improvements. Molloy and Shehu (2016) propose that the community needs a benchmark testing dataset and a baseline approach against which specific improvements and extensions can be evaluated. In particular, this work ignores system-specific insights into which variable and which sampling schemes can be more effective over others but instead compiles a broad set of variables and sampling/perturbation operators that can be selected via a probabilistic scheme. Different schemes can be employed at different stages of a roadmap-based method based on the distance of the conformations that need to be connected and the size of the biomolecule under investigation. A general baseline implementation shows comparable performance to system-specific methods and promises that further improvements can guarantee a baseline performance over a broad set of biomolecules and problem instances. Related ideas building on the concept of a move selector are presented by Gipson et al. (2013).

The second direction sacrifices broad applicability in the interest of improving the predictive capability of roadmap-based methods to the point that reliable hypotheses can be formulated to further guide wet-laboratory experimentation. Maximova et al. (2015) recognize that roadmap-based methods do not have to operate in a *de novo* setting but instead exploit the rich set of wet-laboratory

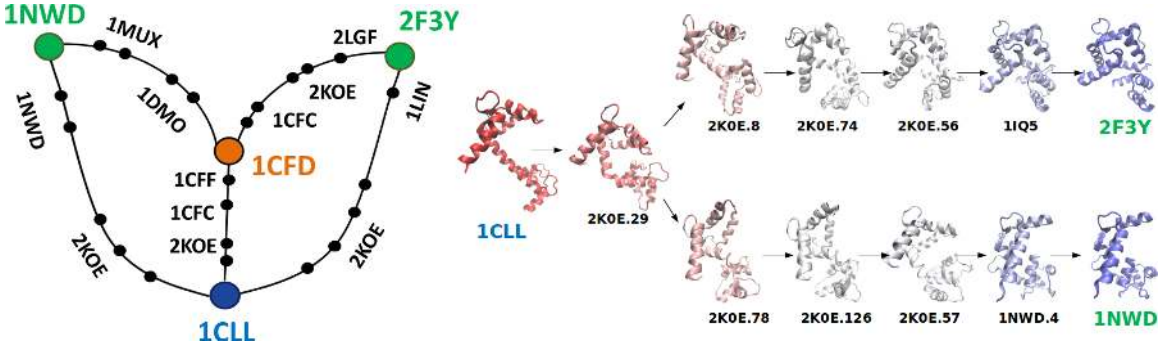


Figure 13: This figure is reproduced from the work of Maximova et al. (2015). The left panel shows a schematic that summarizes all paths within a small energetic threshold of the minimum-cost path connecting structure pairs of interest in calmodulin. Analysis of these paths reveals that known, wet-laboratory structures mediate transitions of interest. The PDB ids of these mediating structures are shown along each of the paths. The right panel shows successive structures in the minimum-cost paths found for transitions of calmodulin from structure with PDB id 1CLL to that with PDB id 2F3Y and then from structure with PDB id 1CLL to that with PDB id 1NWD. Numbers indicate model number within an NMR entry. This figure is reproduced with permission of IEEE Society 2015.

structures to determine the variable space of interest. In particular, the SoPRIM algorithm proposed by Maximova et al. subjects wet-laboratory structures of different sequences of a protein to a statistical multivariate analysis to determine variables that represent collective motions of atoms. Sampling focuses on this space of variables and a multiscaling technique converts samples to all-atom structures that are local minima of the Amber ff14SB energy function. The samples are embedded in a roadmap, and distance constraints ensure that edges are only placed between neighboring samples. Edges are weighted based on the concept of minimum cost, recording only energetic increases.

In an additional contrast to existing roadmap-based treatments, the work of Maximova et al. (2015) yields not only the minimum-cost path connecting a given start to a given goal structure, but allows extracting additional paths with similar costs. The concept of tours is employed, based on related work in robotics. The tours allow to investigate specific hypotheses regarding the participation of known meta-stable structures in a transition. A set of such structures can be specified, and all minimum-cost tours that consider all subsets and orders of such structures are reported. Analysis of tours with costs no higher than a specific threshold over the minimum-cost path reveals precious information regarding important function-regulation transitions in several proteins, including Ras and calmodulin. A summary result is shown in Figure 13.

Figure 13 extracts several energetically-credible paths representing the various, equiprobable routes of transitions of calmodulin from its open, unbound state (represented by structure with PDB id 1CLL) to two different, closed peptide and protein-bound states (represented by structures with PDB id 2F3Y and 1NWD). The schematic summary of these paths in Figure 13 highlights that these open-to-closed transitions in calmodulin may not make use of the calcium-bound structure (PDB id 1CFD). Indeed, paths that go through this structure have higher energetic cost. A different intermediate structure emerges from the analysis of paths. This structure (under PDB id 2K0E) also binds calcium but is slightly different for that under PDB id 1CFD. The succession of structures shown in 13 makes clear that the domain collapse, re-arrangement, and partial unfolding of the helix that links the N- and C-terminal domains in calmodulin are gradual, as captured in various structures in the NMR ensemble with PDB id 2K0E. This result is in good agreement with

the wet-laboratory study in the work of Gsponer, Christodoulou, Cavalli, Bui, Richter, Dobson, and Vendruscolo (2008), which, in addition to contributing the NMR entry under PDB id 2K0E to the Protein Data Bank, also concludes that correlated motions within the 2K0E Ca(2+)-CaM state direct the structural fluctuations toward complex-like substates (Gsponer et al., 2008). While the wet-laboratory study by Gsponer et al. (2008) was restricted to MLCK binding of CaM, the results obtained by the SoPRIM algorithm (Maximova et al., 2015) suggest that the same mechanism observed by Gsponer et al. (2008) prepares CaM for binding to other peptides (the C-terminal Domain of Petunia Glutamate Decarboxylase in 1NWD and the IQ domain in 2F3Y). The work of Maximova et al. (2015) points to a general mechanism for the apo-to-closed/complexed dynamics of calmodulin, where correlated motions within the calcium-bound state direct the fluctuations and population shift of this protein to its peptide-bound states.

5. Outstanding Challenges and Directions of Research

Robotics-inspired methods are becoming more powerful and diverse in their algorithmic strategies and the problems they address in biomolecular modeling. While this survey has focused on tree- and roadmap-based methods for modeling protein-ligand binding, protein *de novo* structure prediction, protein and RNA folding and unfolding, structural transitions in peptides and proteins, and energy landscape mapping, other methods are building on related ideas to efficiently map ligand migration channel networks in dynamic proteins (Lin & Song, 2011; Na & Song, 2015) or even model antibody aggregation processes (Hoard, Jacobson, Manavi, & Tapia, 2016). While we have attempted to provide a broad and deep survey of robotics-inspired methods for biomolecular modeling, an exhaustive survey is not possible. This particular sub-domain at the interface of Robotics and computational structural biology is rapidly progressing, as demonstrated by the increasing number of adaptations and applications showcased in this survey over earlier, related reviews of robotics-inspired methods (Al-Bluwi et al., 2012; Gipson, Hsu, Kavraki, & Latombe, 2012). As this survey showcases, several algorithmic challenges remain. Below we provide a partial list of these challenges and prospects for future research.

5.1 Problem-Specific versus General Treatments

There is a pressing need in the community for benchmarks. While work has been largely driven by specific biological systems and problems of interest, such data-driven research has often resulted in specific design decisions that are not easily transferable to other systems and other problems. For instance, key decisions on how to reduce dimensionality of the variable space and design compliant sampling strategies and perturbation operators on a specific problem instance may not be applicable to another problem. A realization of the need for baseline, general treatments and benchmarks is leading researchers towards non-specific treatments to establish benchmarks and baseline performance. Better sharing of problem instances, metrics, and algorithms with known baseline performance will also be key to allow researchers to build on existing work and expedite progress.

5.2 Sampling

There is a growing realization that sampling will remain a central issue, despite clever reduced representations and sampling strategies. While the community of researchers adapting robot motion planning treatments for biomolecular modeling has been successful at integrating important knowledge

about biomolecules in model selection, sampling strategies, and energetic evaluations, this community has largely remained isolated from complementary work in AI on stochastic optimization of continuous, non-linear variable spaces. In particular, there is a growing body of work in the evolutionary computation community on optimization of complex fitness landscapes. Some ideas from this community have successfully been employed in *de novo* structure prediction (Shehu, 2013) and mapping of protein energy landscapes (Clausen & Shehu, 2015; Clausen et al., 2015; Sapin, Carr, De Jong, & Shehu, 2016). These ideas are also beginning to be incorporated in robotics-inspired treatments of biomolecular dynamics (Molloy et al., 2016). Better awareness and integration of effective practices of other communities dealing with similarly challenging high-dimensional problems is likely to address issues in sampling and lead to more powerful robotics-inspired treatments. In this context, we see great opportunity for AI researchers to make contributions in sampling-based treatments of biomolecular dynamics.

5.3 Decorrelations of Paths

In particular, applications of tree- and roadmap-based methods for modeling structural transitions of biomolecules, path correlation is an issue. Path correlations can potentially skew any statistics of interest and even yield to incorrect conclusions about a structural transition. The culprit in tree-based methods is the bias that is applied to steer the conformation tree to the goal conformation. Even multiple executions of a tree-based method are likely to result in similar paths. To some extent, this source of path correlations can be addressed. For instance, Molloy and Shehu (2013) makes use of an additional projection layer to steer the tree towards under-sampled regions of the conformation space. This is shown to improve path diversity. Yet another culprit that is shared by tree- and roadmap-based methods is density of sampling. For instance, undersampling of specific regions may lead to the conclusion that the region is not energetically favorable for the biomolecule at hand. Further investigation is needed to quantify and reduce path correlations in robotics-inspired methods. This direction is also ripe for cross-fertilization of ideas from different sub-communities in AI.

5.4 Injection of Dynamics

A common criticism of robotics-inspired methods is that they are essentially geometric treatments of biomolecules. While to some extent geometric treatments are accepted in modeling biomolecular structure, they are seen as inadequate in modeling biomolecular dynamics. Modeling dynamics is largely seen as exclusive to MD simulation frameworks. In a somewhat colloquial and simplistic characterization of robotics-inspired methods, "biomolecular dynamics has nothing to do with robot motion planning." This characterization can be overcome by pointing out that the superficial analogies are only used to inspire robotics researchers, but the deeper analogies that are exploited and shown to have an impact are those on selection of models, variables, fast forward and inverse kinematics, and effective sampling strategies. It is worth noting that the latter are not exclusively the domain of robotics-inspired researchers. On the contrary, issues of effective variable selection or representation, variation operators, employment of such operators in sampling strategies, and others are of broad interest to AI researchers working on optimization problems as part of modeling abstract, mechanical, or biological systems.

At various places, this survey has highlighted that robotics-inspired methods are capable not only of reproducing wet-laboratory knowledge and data but also of providing novel findings to direct

further experimentation in wet laboratories. Still, a valid criticism of robotics-inspired methods is that edges in trees or roadmaps do not provide a detailed view of the diffusion between the two conformations they connect. While the survey points out in Section 4 several challenges with integrating MD trajectories in robotics-inspired framework, it is important that the community think of ways to do so effectively. Injection of ideas from the AI community at large may prove beneficial here. A growing body of work in computational biophysics is pointing to effective frameworks of biomolecular dynamics that integrate thousands or more short MD trajectories in MSMs to capture biomolecular dynamics. Cross-fertiziliation of ideas from the AI and biophysics communities is likely to prove fruitful in explicitly integrating MD in robotics-inspired methods.

5.5 Beyond Path Computations: Roadmaps and MSMs

As this survey highlights in Section 4, MSMs have become very popular in computational biophysics literature to organize and extract statistics from many, independent MD simulations of biomolecular folding or other structural transitions (Jayachandran, Vishal, & Pande, 2006; Chodera et al., 2007; Noé & Fischer, 2008; Prinz, Keller, & Noé, 2011a; Noé, Doose, Daidone, Löllmann, Sauer, Chodera, & Smith, 2011; Pérez-Hernández, Paul, Giorgino, De Fabritiis, & Noé, 2013; Weber, Jack, & Pande, 2013; Deng, Dai, & Levy, 2013; Chodera & Noé, 2014; Malmstrom, Lee, Van Wart, & Amaro, 2014; Song & Zhuang, 2014; Shukla, Hernández, Weber, & Pande, 2015). Several survey articles are dedicated to reviewing MSM-based treatments of biomolecular dynamics (Pande, Beachamp, & Bowman, 2010; Gipson et al., 2012; Maximova et al., 2016) review MSM-based treatments of biomolecular dynamics. Works by Chiang et al. (2006), Chiang et al. (2007), Chiang et al. (2010), and Molloy et al. (2016) provide an important first step in the integration of MSMs in the analysis of conformation spaces probed via robotics-inspired algorithms. While Chiang et al. (2010) and Molloy et al. (2016) address some of the issues on to convert roadmaps into MSMs, many others remain, including definition of structural states, possible undersampling of specific states, feedback mechanisms to address undersampling, and rigorous calculation of transition probabilities. Some of these issues are also contended with in the computational biophysics community, and initial treatments have emerged (Singhal et al., 2004; Singhal & Pande, 2005; Prinz, Wu, Sarich, Keller, Senne, Held, Chodera, Schütte, & Noé, 2011b; Malmstrom et al., 2014; Da, Sheong, Silva, & Huang, 2014). We see a great opportunity here for AI researchers, particularly those with expertise in machine learning, to coordinate efforts with computational biophysicists. Such efforts will undoubtedly lead to richer and more powerful computational treatments of biomolecular dynamics.

5.6 Cross-Fertilization of Ideas

As this survey shows, work in modeling biomolecular structure and dynamics is highly interdisciplinary, and great progress is achieved when ideas from different communities are combined and integrated in computational treatments. There is a rich set of scientific questions that can be formulated to understand the role of biomolecular structure and dynamics in human biology and health. These questions often result in exceptionally challenging computational problems that necessitate sophisticated algorithmic treatments. Treatments that add to the current knowledge of biomolecular systems in chemistry, physics, and biophysics are likely to advance not only our modeling capabilities but also make important, general contributions to AI research.

Acknowledgements

Funding for this work is provided in part by the National Science Foundation. The work A. Shehu is supported by NSF-CCF1421001, NSF-ACI1440581, and NSF-IIS1144106. The work of E. Plaku is supported by NSF-ACI1440581, NSF-IIS1449505, and NSF-IIS1548406.

References

Abayagan, R., Totrov, M., & Kuznetsov, D. (1994). ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comput Chem*, *15*(5), 488–506.

Ådén, J., & Wolf-Watz, M. (2007). NMR identification of transient complexes critical to adenylate kinase catalysis. *J Amer Chem Soc*, *129*(45), 14003 –14012.

Al-Bluwi, I., Siméon, T., & Cortés, J. (2012). Motion planning algorithms for molecular simulations: A survey. *Comput Sci Rev*, *6*(4), 125–143.

Al-Bluwi, I., Vaisset, M., Siméon, T., & Cortés, J. (2013). Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. *BMC Struct Biol*, *13*(S2), Suppl 1.

Amato, N. M., Bayazit, B., Dale, L., Jones, C., & Vallejo, D. (1998). OBPRM: An obstacle-based PRM for 3D workspaces. In *Workshop Algorithm Found Robot*, Vol. 86 of *Springer Tracts in Advanced Robotics*, pp. 156–168. Springer.

Amato, N. M., Dill, K. A., & Song, G. (2003). Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J Comput Biol*, *10*(3-4), 239–255.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230.

Apaydin, M. S., Brutlag, D. L., Guestrin, C., Hsu, D., & Latombe, J.-C. (2003). Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. *J Comput Biol*, *10*(3-4), 257–281.

Apaydin, M. S., Brutlag, D. L., Hsu, D., & Latombe, J.-C. (2002). Stochastic conformational roadmaps for computing ensemble properties of molecular motion. In *Workshop Algorithm Found Robot*, pp. 131–147, Nice, France. IEEE.

Apaydin, M. S., Singh, A. P., Brutlag, D. L., & Latombe, J.-C. (2001). Capturing molecular energy landscapes with probabilistic conformational roadmaps. In *Intl Conf Robot Autom (ICRA)*, Vol. 1, pp. 932–939, Seoul, Korea. IEEE.

Atilgan, A., Durell, S., Jernigan, R., Demirel, M., Keskin, O., & Bahar, I. (2001). Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*, *80*(1), 505–515.

Bahar, R., & Rader, A. J. (2005). Coarse-grained normal mode analysis in structural biology. *Curr Opinion Struct Biol*, *204*(5), 1–7.

Baldwin, R. L. (1995). The nature of protein folding pathways: the classical versus the new view. *J Biomol NMR*, *5*(2), 103–109.

Barbe, S., Cortés, J., Siméon, T., Monsan, P., Remaud-Siméon, M., & Andre, I. (2011). A mixed molecular modeling-robotics approach to investigate lipase large molecular motions. *Proteins: Struct Funct Bioinf*, *79*(8), 2517–2529.

Baron, R. (2013). Fast sampling of A-to-B protein global conformational transitions: From Galileo Galilei to Monte Carlo anisotropic network modeling. *Biophys J*, *105*(7), 1545–1546.

Batcho, P., Case, D. A., & Schlick, T. (2001). Optimized particle-mesh ewald/multiple-time step integration for molecular dynamics simulations. *J Chem Phys*, *115*(9), 4003–4018.

Bekris, K. E., & Kavraki, L. E. (2007). Greedy but safe replanning under kinodynamic constraints. In *Intl Conf Robot Autom (ICRA)*, pp. 704–710, Rome, Italy. IEEE.

Berman, H. M., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nat Struct Biol*, *10*(12), 980–980.

Boehr, D. D., & Wright, P. E. (2008). How do proteins interact?. *Science*, *320*(5882), 1429–1430.

Bradley, P., Misura, K. M., & Baker, D. (2005). Toward high-resolution *de novo* structure prediction for small proteins. *Science*, *309*(5742), 1868–1871.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*, *4*(2), 187–217.

Bryngelson, J. D., Onuchic, J. N., Socci, N. D., & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct Funct Genet*, *21*(3), 167–195.

Bryngelson, J. D., & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA*, *84*(21), 7524–7528.

Burgess, A. W., & Scheraga, H. A. (1975). Assessment of some problems associated with prediction of the three-dimensional structure of a protein from its amino-acid sequence. *Proc Natl Acad Sci USA*, *72*(4), 1221–1225.

Burns, B., & Brock, O. (2007). Single-query motion planning with utility-guided random trees. In *Intl Conf Robot Autom (ICRA)*, pp. 3307–3312, Rome, Italy. IEEE.

Case, D. A., Darden, T. A., Cheatham, T. E. I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., Walker, R. C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C., Mongan, J., et al. (2014). Amber 14. http://ambermd.org/.

Chiang, T. H., Apaydin, M., Brutlag, D., Hsu, D., & Latombe, J. (2006). Predicting experimental quantities in protein folding kinetics using stochastic roadmap simulation. In *Res Comput Mol Biol*, Vol. 3909 of *Lecture Notes in Computer Science*, pp. 410–424. Springers.

Chiang, T. H., Apaydin, M. S., Brutlag, D. L., Hsu, D., & Latombe, J.-C. (2007). Using stochastic roadmap simulation to predict experimental quantities in protein folding kinetics: folding rates and phi-values. *J Comput Biol*, *14*(5), 578–593.

Chiang, T. H., Hsu, D., & C., L. J. (2010). Markov dynamic models for long-timescale protein motion. *Bioinformatics*, *26*(12), 269–277.

Chirikjian, G. S. (1993). General methods for computing hyper-redundant manipulator inverse kinematics. In *Intl Conf Intell Robot Sys (IROS)*, Vol. 2, pp. 1067–1073, Yokohama, Japan. IEEE.

Chodera, J. D., & Noé, F. (2014). Markov state models of biomolecular conformational dynamics. *Curr Opinion Struct Biol*, *25*, 135–144.

Chodera, J. D., Singhal, N., Pande, V. S., Dill, K. A., & Swope, W. C. (2007). Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics. *J Chem Phys*, *126*(15), 155101.

Choset, H., & et al. (2005). *Principles of Robot Motion: Theory, Algorithms, and Implementations* (1st edition). MIT Press, Cambridge, MA.

Ciu, Q., & Bahar, I. (2005). *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems* (1st edition). CRC Press.

Clausen, R., Ma, B., Nussinov, R., & Shehu, A. (2015). Mapping the conformation space of wildtype and mutant H-Ras with a memetic, cellular, and multiscale evolutionary algorithm. *PLoS Comput Biol*, *11*(9), e1004470.

Clausen, R., & Shehu, A. (2015). A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes. *J Comp Biol*, *22*(9), 844–860.

Clementi, C. (2008). Coarse-grained models of protein folding: Toy-models or predictive tools?. *Curr Opinion Struct Biol*, *18*(1), 10–15.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc Natl Acad Sci USA*, *102*(21), 7426–7431.

Cooper, A. (1984). Protein fluctuations and the thermodynamic uncertainty principle. *Prog Biophys Mol Biol*, *44*(3), 181–214.

Cortés, J., Jaillet, L., & Siméon, T. (2007). Molecular disassembly with RRT-like algorithms. In *Intl Conf Robot Autom (ICRA)*, pp. 3301–3306, Roma, Italy.

Cortés, J., Le, D. T., Lehl, R., & Siméon, T. (2010). Simulating ligand-induced conformational changes in proteins using a mechanical disassembly method. *Phys Chem Chem Phys*, *12*(29), 8268–8276.

Cortés, J., Siméon, T.AND Remauld-Siméon, M., & Tran, V. (2004). Geometric algorithms for the conformational analysis of long protein loops. *J Comput Chem*, *25*(7), 956–967.

Cortés, J., Siméon, T.AND de Angulo, R., Guieysse, D., Remaud-Siméon, M., & Tran, V. (2005). A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, *21*(S1), 116–125.

Craig, J. (1989). *Introduction to robotics: mechanics and control* (2nd edition). Addison-Wesley, Boston, MA.

Da, L. T., Sheong, F. K., Silva, D. A., & Huang, X. (2014). Application of Markov state models to simulate long timescale dynamics of biological macromolecules. *Adv Exp Med Biol*, *805*, 29–66.

Dalibard, S., & Laumond, J.-P. (2009). Control of probabilistic diffusion in motion planning. In *Workshop Algorithm Found Robot*, Vol. 57 of *Springer Tracts in Advanced Robotics*, pp. 467–481. Springer.

Das, A., Gur, M., Cheng, M. H., Jo, S., Bahar, I., & Roux, B. (2014). Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model. *PLoS Comput Biol*, *10*(4), e1003521.

Das, P., Matysiak, S., & Clementi, C. (2005). Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc Natl Acad Sci USA*, *102*(29), 10141–10146.

Das, P., Moll, M., Stamati, H., Kavraki, L. E., & Clementi, C. (2006). Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proc Natl Acad Sci USA*, *103*(26), 9885–9890.

Delarue, M., & Sanejouand, Y. H. (2002). Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J Mol Biol*, *320*(5), 1011–1024.

Deng, N.-J., Dai, W., & Levy, R. M. (2013). How kinetics within the unfolded state affects protein folding: An analysis based on markov state models and an ultra-long md trajectory. *J Phys Chem B*, *117*(42), 12787–12799.

Denny, J., & Amato, N. M. (2013). Toggle PRM: A coordinated mapping of C-free and C-obstacle in arbitrary dimension. In *Workshop Algorithm Found Robot*, Vol. 86 of *Springer Tracts in Advanced Robotics*, pp. 297–312. Springer.

Devaurs, D., Molloy, K., Vaisset, M., Shehu, A., Cortés, J., & Siméon, T. (2015). Characterizing energy landscapes of peptides using a combination of stochastic algorithms. *IEEE Trans NanoBioScience*, *14*(5), 545–552.

Diekmann, S., & Hoischen, C. (2014). Biomolecular dynamics and binding studies in the living cell. *Physics of Life Reviews*, *11*(1), 1–30.

Dill, K. A., & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol*, *4*(1), 10–19.

Dryga, A., Chakrabarty, S., Vicatos, S., & Warshel, A. (2011). Realistic simulation of the activation of voltage-gated ion channels. *Proc Natl Acad Sci USA*, *109*(9), 3335–3340.

Dubrow, A. (2015). What got done in one year at NSF's Stampede supercomputer. *Comput Sci Eng*, *17*(2), 83–88.

Ekenna, C., Thomas, S., & Amato, N. (2016). Adaptive local learning in sampling based motion planning for protein folding. *BMC Syst Biol*, *10*(Suppl 2).

Engh, R. A., & Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr*, *A47*, 392–400.

Enosh, A., Raveh, B., Furman-Schueler, O., Halperin, D., & Ben-Tal, N. (2008). Generation, comparison, and merging of pathways between protein conformations: gating in K-channels. *Biophys J*, *95*(8), 3850–3860.

Fattebert, J.-L., Richards, D. F., & Glosli, J. N. (2012). Dynamic load balancing algorithm for molecular dynamics based on voronoi cells domain decompositions. *Comput Phys Communic*, *183*(12), 2608–2615.

Fenwick, R. B., van den Bedem, H., Fraser, J. S., & Wright, P. E. (2014). Integrated description of protein dynamics from room-temperature X-ray crystallography and NMR. *Proc Natl Acad Sci USA*, *111*(4), E445–E454.

Fernández-Medarde, A., & Santos, E. (2011). Ras in cancer and developmental diseases. *Genes Cancer*, *2*(3), 344–358.

Fersht, A. (2013). Profile of martin karplus, michael levitt, and arieh warshel, 2013 nobel laureates in chemistry. *Proc Natl Acad Sci USA*, *110*(49), 19656–19657.

Fersht, A. R. (1999). *Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding* (3 edition). W. H. Freeman and Co., New York, NY.

Feynman, R. P., Leighton, R. B., & Sands, M. (1963). *The Feynman Lectures on Physics*. Addison-Wesley, Reading, MA.

Fischer, S., & Karplus, M. (1992). Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem Phys Lett*, *194*(3), 252–261.

Fox, N., & Streinu, I. (2013). Towards accurate modeling of noncovalent interactions for protein rigidity analysis. *BMC Bioinf*, *14*(Suppl 18), S3.

Gall, A., Ilioaia, C., Krüger, T. P., Novoderezhkin, V. I., Robert, B., & van Grondelle, R. (2015). Conformational switching in a light-harvesting protein as followed by single-molecule spectroscopy. *Biophys J*, *108*(11), 2713–2720.

Gibbs, J. B., Schaber, M. D., Allard, W. J., Sigal, I. S., & Scolnick, E. M. (1988). Purification of Ras GTPase activating protein from bovine brain. *Proc Natl Acad Sci USA*, *85*(14), 5026–5030.

Gipson, B., Hsu, D., Kavraki, L. E., & Latombe, J.-C. (2012). Computational models of protein kinematics and dynamics: Beyond simulation. *Annu Rev Anal Chem*, *5*, 273–291.

Gipson, B., Moll, M., & Kavraki, L. E. (2013). SIMS: A hybrid method for rapid conformational analysis. *PLoS One*, *8*(7), e68826.

Gorfe, A. A., Grant, B. J., & McCammon, J. A. (2008). Mapping the nucleotide and isoform-dependent structural and dynamical features of Ras proteins. *Structure*, *16*(6), 885–896.

Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S., & Walker, R. C. (2012). Routine microsecond molecular dynamics simulations with amber on GPUs. 1. Generalized Born. *J Chem Theory Comput*, *8*(5), 1542–1555.

Grant, B. J., Gorfe, A. A., & McCammon, J. A. (2009). Ras conformational switching: Simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. *PLoS Comput Biol*, *5*(3), e1000325.

Greenleaf, W. J., Woodside, M. T., & Block, S. M. (2007). High-resolution, single-molecule measurements of biomolecular motion. *Annu Rev Biophys Biomol Struct*, *36*, 171–190.

Gremer, L., Gilsbach, B., Ahmadian, M. R., & Wittinghofer, A. (2008). Fluoride complexes of oncogenic Ras mutants to study the Ras-RasGap interaction. *Biol Chem*, *389*(9), 1163–1171.

Gsponer, J., Christodoulou, J., Cavalli, A., Bui, J. M., Richter, B., Dobson, C. M., & Vendruscolo, M. (2008). A coupled equilibrium shift mechanism in calmodulin-mediated signal transduction. *Structure*, *16*(5), 736–746.

Han, K. F., & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA*, *93*(12), 5814–5818.

Han, L., & Amato, N. M. (2001). A kinematics-based probabilistic roadmap method for closed chain systems. In Donald, B. R., Lynch, K. M., & Rus, D. (Eds.), *Algorithmic and Computational Robotics: New Directions*, pp. 233–246. AK Peters, MA.

Harvey, M. J., Giupponi, G., & de Fabritiis, G. (2009). ACEMD: Accelerating biomolecular dynamics in the microsecond timescale. *J Comput Theor Chem*, *5*(6), 1632–1639.

Haspel, N., Moll, M., Baker, M. L., Chiu, W., & E., K. L. (2010). Tracing conformational changes in proteins. *BMC Struct Biol*, *10*(Suppl1), S1.

Haspel, N., Tsai, C., Wolfson, H., & Nussinov, R. (2003). Reducing the computational complexity of protein folding via fragment folding and assembly. *Protein Sci*, *12*(6), 1177–1187.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hermans, J., Berendsen, H. J. C., van Gunsteren, W. F., & Postma, J. P. M. (1984). A consistent empirical potential for water-protein interactions. *Biopolymers*, *23*(8), 1513–1518.

Hoang, T. H., Trovato, A., Seno, F., Banavar, J. R., & Maritan, A. (2007). Geometry and symmetry presculpt the free-energy landscape of proteins. *Proc Natl Acad Sci USA*, *101*(21), 7960–7964.

Hoard, B., Jacobson, B., Manavi, K., & Tapia, L. (2016). Extending rule-based methods to model molecular geometry and 3d model resolution. *BMC Syst Biol*, *10*(Suppl 2), 48.

Hoelder, S., Clarke, P. A., & Workman, P. (2012). Discovery of small molecule cancer drugs: Successes, challenges and opportunities. *Mol Oncol*, *6*(2), 522–524.

Hohlbein, J., Craggs, T. D., & Cordes, T. (2014). Alternating-laser excitation: single-molecule FRET and beyond. *Chem Soc Rev*, *43*(4), 1156–1171.

Hori, N., Chikenji, G., & Takada, S. (2009). Folding energy landscape and network dynamics of small globular proteins. *Proc Natl Acad Sci USA*, *106*(1), 73–78.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., & Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Struct Funct Bioinf*, *65*(3), 712–725.

Hsu, D., Kindel, R., Latombe, J.-C., & Rock, S. (2002). Randomized kinodynamic motion planning with moving obstacles. *Intl J Robot Res*, *21*(3), 233–255.

Hsu, D., Sánchez-Ante, G., & Sun, Z. (2005). Hybrid PRM sampling with a cost-sensitive adaptive strategy. In *Intl Conf Robot Autom (ICRA)*, pp. 3885–3891, Barcelona, Spain.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics. *J Mol Graph Model*, *14*(1), 33–38. http://www.ks.uiuc.edu/Research/vmd/.

Jaillet, L., Corcho, F. J., Perez, J.-J., & Cortés, J. (2011). Randomized tree construction algorithm to explore energy landscapes. *J Comput Chem*, *32*(16), 3464–3474.

Jaillet, L., Cortés, J., & Siméon, T. (2008). Transition-based RRT for path planning in continuous cost spaces. In *Intl Conf Intell Robot Sys (IROS)*, pp. 22–26, Stanford, CA. IEEE/RSJ.

Jaillet, L., Yershova, A., LaValle, S. M., & Siméon, T. (2005). Adaptive tuning of the sampling domain for dynamic-domain RRTs. In *Intl Conf Intell Robot Sys (IROS)*, pp. 4086–4091. IEEE/RSJ.

Jayachandran, G., Vishal, V., & Pande, V. S. (2006). Using massively parallel simulation and Markovian models to study protein folding: examining the dynamics of the villin headpiece. *J Chem Phys*, *124*(16), 164902–164914.

Jenzler-Wildman, K., & Kern, D. (2007). Dynamic personalities of proteins. *Nature*, *450*(7172), 964–972.

Jorgensen, W. L., Maxwell, D. S., & Tirado-Reves, J. (1988). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Amer Chem Soc*, *118*(45), 11225–11236.

Kalisiak, M., & van de Panne, M. (2006). RRT-blossom: RRT with a local flood-fill behavior. In *Intl Conf Robot Autom (ICRA)*, pp. 1237–1242, Orlando, FL. IEEE.

Kamerlin, S. C., Haranczyk, M., & Warshel, A. (2009). Progresses in *ab initio* QM/MM free energy simulations of electrostatic energies in proteins: Accelerated QM/MM studies of pka, redox reactions and solvation free energies. *J Phys Chem B*, *113*(5), 1253–1272.

Karam, P., Powdrill, M. H., Liu, H. W., Vasquez, C., Mah, W., Bernatchez, J., Götte, M., & Cosa, G. (2014). Dynamics of hepatitis C virus (HCV) RNA-dependent RNA polymerase NS5B in complex with RNA. *J Biol Chem*, *289*(20), 14399–14411.

Karaman, S., & Frazzoli, E. (2011). Sampling-based algorithms for optimal motion planning. *Intl J Robot Res*, *30*(7), 846–894.

Karnoub, A. E., & Weinberg, R. A. (2008). Ras oncogenes: split personalities. *Nat Rev Mol Cell Biol*, *9*(7), 517–531.

Kavraki, L. E., Švestka, P., Latombe, J. C., & Overmars, M. H. (1996). Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans Robot Automat*, *12*(4), 566–580.

Kay, L. E. (1998). Protein dynamics from NMR. *Nat Struct Biol*, *5*(2-3), 513–517.

Kay, L. E. (2005). NMR studies of protein structure and dynamics. *J Magn Reson*, *173*(2), 193–207.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., & Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, *181*(4610), 662–666.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., & Shore, V. C. (1960). Structure of myoglobin: A three-dimensional Fourier synthesis at 2Å resolution. *Nature*, *185*(4711), 422–427.

Khaliullin, R. Z., VandeVondele, J., & Hutter, J. (2013). Efficient linear-scaling density functional theory for molecular systems. *J Chem Theory Comput*, *9*(10), 4421–4427.

Kiesel, S., Burns, E., & Ruml, W. (2012). Abstraction-guided sampling for motion planning. In *Symp Combinat Search (SOCS)*, pp. 162–163, Niagara Falls, Canada.

Kim, K. M., Jernigan, R. L., & Chirikjian, G. S. (2002a). Efficient generation of feasible pathways for protein conformational transitions. *Biophys J*, *83*(3), 1620–1630.

Kim, M. K., Chirikjian, G. S., & Jernigan, R. L. (2002b). Elastic models of conformational transitions in macromolecules. *J Mol Graph Model*, *21*(2), 151–160.

Kirillova, S., Cortés, J., Stefaniu, A., & Siméon, T. (2008). An NMA-guided path planning approach for computing large-amplitude conformational changes in proteins. *Proteins: Struct Funct Bioinf*, *70*(1), 131–143.

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

Ladd, A. M., & Kavraki, L. E. (2004). Fast tree-based exploration of state space for robots with dynamics. In *Workshop Algorithm Found Robot*, pp. 297–312. Springer, Utrecht/Zeist, Netherlands.

Ladd, A. M., & Kavraki, L. E. (2005). Motion planning in the presence of drift, underactuation and discrete system changes. In *Robot: Sci and Sys*, pp. 233–241, Boston, MA.

LaValle, S. M., & Kuffner, J. J. (2001). Randomized kinodynamic planning. *Intl J Robot Res*, *20*(5), 378–400.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y. E., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popovi, Z., & et. al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, *487*, 545–574.

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

Lee, H. M., M., K. S., Kim, H. M., & Suh, Y. D. (2013). Single-molecule surface-enhanced Raman spectroscopy: a perspective on the current status. *Phys Chem Chem Phys*, *15*, 5276–5287.

Lee, J., Kwon, O., Zhang, L., & Yoon, S.-E. (2014). A selective retraction-based RRT planner for various environments. *IEEE Trans Robotics*, *30*(4), 1002–1011.

Levitt, M., & Lifson, S. (1969). Refinement of protein conformations using a macromolecular energy minimization procedure. *J Mol Biol*, *46*(2), 269–279.

Levitt, M., & Warshel, A. (1975). Computer simulation of protein folding. *Nature*, *253*(5494), 94–96.

Levy, Y., Jortner, J., & Becker, O. M. (2001). Solvent effects on the energy landscapes and folding kinetics of polyalanine. *Proc Natl Acad Sci USA*, *98*(5), 2188–2193.

Li, D., Yang, H., Han, L., & Huo, S. (2008). Predicting the folding pathway of engrailed homeodomain with a probabilistic roadmap enhanced reaction-path algorithm. *Biophys J*, *94*(5), 1622–1629.

Lifson, S., & Warshel, A. (1968). A consistent force field for calculation on conformations, vibrational spectra and enthalpies of cycloalkanes and n-alkane molecules. *J Phys Chem*, *49*, 5116–5129.

Lin, T., & Song, G. (2011). Efficient mapping of ligand migration channel networks in dynamic proteins. *Proteins: Struct Funct Bioinf*, *79*(8), 24752490.

Lindorff-Larsen, K., Piana, S., Dror, R. O., & Shaw, D. E. (2011). How fast-folding proteins fold. *Science*, *334*(6055), 517–520.

Lois, G., Blawzdziewicz, J., & O'Hern, C. S. (2010). Protein folding on rugged energy landscapes: Conformational diffusion on fractal networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, *81*(5 Pt 1), 051907.

Lotan, I., van den Bedem, H., Deacon, A. M., & Latombe, J.-C. (2004). Computing protein structures from electron density maps: The mising loop problem. In Erdman, M., Hsu, D., Overmars, M., & van der Stappen, F. (Eds.), *Algorithmic Foundations of Robotics VI*, pp. 153–168. Springer STAR Series.

Ma, B., Kumar, S., Tsai, C., & Nussinov, R. (1999). Folding funnels and binding mechanisms. *Protein Eng*, *12*(9), 713–720.

Ma, J., & Karplus, M. (1997). Molecular switch in signal transduction: reaction paths of the conformational changes in ras p21. *Proc Natl Acad Sci USA*, *94*(22), 11905–11910.

Maisuradze, G. G., Liwo, A., & Scheraga, H. A. (2009). Principal component analysis for protein folding dynamics. *J Mol Biol*, *385*(1), 312–329.

Malmstrom, R. D., Lee, C. T., Van Wart, A. T., & Amaro, R. E. (2014). Application of molecular-dynamics based Markov state models to functional proteins. *J Chem Theory Comput*, *10*(7), 2648–2657.

Manocha, D., & Canny, J. (1994). Efficient inverse kinematics for general 6r manipulator. *IEEE Trans Robot Autom*, *10*(5), 648–657.

Manocha, D., & Zhu, Y. (1994). Kinematic manipulation of molecular chains subject to rigid constraints. In Altman, R. B., Brutlag, D. L., Karp, P. D., Lathrop, R. H., & Searls, D. B. (Eds.), *Intl Conf Intell Sys Mol Biol (ISMB)*, Vol. 2, pp. 285–293, Stanford, CA. AAAI.

Manocha, D., Zhu, Y., & Wright, W. (1995). Conformational analysis of molecular chains using nano-kinematics. *Comput. Appl. Biosci.*, *11*(1), 71–86.

Maragakis, P., & Karplus, M. (2005). Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. *J Mol Biol*, *352*(4), 807–822.

Matysiak, S., & Clementi, C. (2004). 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*(8), 235–248.

Matysiak, S., & Clementi, C. (2006). Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J Mol Biol*, *363*(1), 297–308.

Maximova, T., Moffatt, R., Ma, B., Nussinov, R., & Shehu, A. (2016). Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput Biol*, *12*(4), e1004619.

Maximova, T., Plaku, E., & Shehu, A. (2015). Computing transition paths in multiple-basin proteins with a probabilistic roadmap algorithm guided by structure data. In *Intl Conf on Bioinf and Biomed (BIBM)*, pp. 35–42, Washington, D.C. IEEE.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. *Nature*, *267*(5612), 585–590.

McLachlan, A. D. (1972). A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr A*, *26*(6), 656–657.

McMahon, T., Thomas, S., & Amato, N. M. (2015). Reachable volume RRT. In *Intl Conf Robot Autom (ICRA)*, pp. 2977–2984, Seattle, WA. IEEE.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *J Chem Phys*, *21*(6), 1087–1092.

Miao, Y., Sinko, W., Pierce, L., Bucher, D., Walker, R. C., & McCammon, J. A. (2014). Improved reweighting of accelerated molecular dynamics simulations for free energy calculation. *J Chem Theory Comput*, *10*(7), 2677–2689.

Michalet, X., Weiss, S., & Jäger, M. (2006). Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem Rev*, *106*(5), 1785–1813.

Moerner, W. E., & Fromm, D. P. (2003). Methods of single-molecule fluorescence spectroscopy. *Rev Scientific Instruments*, *74*(8), 3597–3619.

Moffat, K. (2003). The frontiers of time-resolved macromolecular crystallography: movies and chirped X-ray pulses. *Faraday Discuss*, *122*(79-88), 65–77.

Moll, M., Schwartz, D., & Kavraki, L. E. (2008). Roadmap methods for protein folding. In Zaki, M., & Bystroff, C. (Eds.), *Protein Structure Prediction*, Vol. 413 of *Methods Mol Biol*, pp. 219–239. Springer.

Molloy, K., Clausen, R., & Shehu, A. (2016). A stochastic roadmap method to model protein structural transitions. *Robotica*, *34*(8), 1705–1733.

Molloy, K., Saleh, S., & Shehu, A. (2013). Probabilistic search and energy guidance for biased decoy sampling in *ab-initio* protein structure prediction. *IEEE/ACM Trans Bioinf and Comp Biol*, *10*(5), 1162–1175.

Molloy, K., & Shehu, A. (2013). Elucidating the ensemble of functionally-relevant transitions in protein systems with a robotics-inspired method. *BMC Struct Biol*, *13*(Suppl 1), S8.

Molloy, K., & Shehu, A. (2015). Interleaving global and local search for protein motion computation. In Harrison, R., Li, Y., & Mandoiu, I. (Eds.), *LNCS: Bioinformatics Research and Applications*, Vol. 9096, pp. 175–186, Norfolk, VA. Springer International Publishing.

Molloy, K., & Shehu, A. (2016). A general, adaptive, roadmap-based algorithm for protein motion computation. *IEEE Trans. NanoBioSci.*, *2*(15), 158–165.

Mukherjee, S., & Warshel, A. (2011). Electrostatic origin of the mechanochemical rotary mechanism and the catalytic dwell of F1-ATPase. *Proc Natl Acad Sci USA*, *108*(51), 20550–20555.

Mukherjee, S., & Warshel, A. (2012). Realistic simulations of the coupling between the protomotive force and the mechanical rotation of the F0-ATPase. *Proc Natl Acad Sci USA*, *109*(3), 14876–14881.

Mukherjee, S., & Warshel, A. (2013). Electrostatic origin of the unidirectionality of walking myosin v motors. *Proc Natl Acad Sci USA*, *110*(43), 17326–17331.

Na, H., & Song, G. (2015). Quantitative delineation of how breathing motions open ligand migration channels in myoglobin and its mutants. *Proteins: Struct Funct Bioinf*, *83*(4), 757770.

Neudecker, P., Robustelli, P., Cavalli, A., Walsh, P., Lundstrm, P., Zarrine-Afsar, A., Sharpe, S., Vendruscolo, M., & Kay, L. E. (2012). Structure of an intermediate state in protein folding and aggregation. *Science*, *336*(6079), 362–366.

Nevo, R., Brumfeld, V., Kapon, R., Hinterdorfer, P., & Reich, Z. (2005). Direct measurement of protein energy landscape roughness. *EMBO Rep*, *6*(5), 482–486.

Nielsen, C. L., & Kavraki, L. E. (2000). A two level fuzzy PRM for manipulation planning. In *Intl Conf Intell Robot Sys (IROS)*, Vol. 3, pp. 1716–1721, Takamatsu, Japan. IEEE/RSJ.

Noé, F., Doose, S., Daidone, I., Löllmann, M., Sauer, M., Chodera, J. D., & Smith, J. C. (2011). Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments. *Proc Natl Acad Sci USA*, *108*(12), 4822–4827.

Noé, F., & Fischer, S. (2008). Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opinion Struct Biol*, *18*(2), 154–162.

Olson, B., Hashmi, I., Molloy, K., & Shehu, A. (2012). Basin hopping as a general and versatile optimization framework for the characterization of biological macromolecules. *Advances in AI J*, *2012*(674832).

Olson, B., & Shehu, A. (2012). Evolutionary-inspired probabilistic search for enhancing sampling of local minima in the protein energy surface. *Proteome Sci*, *10*(Suppl 1), S5.

Olson, B. S., Molloy, K., Hendi, S.-F., & Shehu, A. (2012). Guiding search in the protein conformational space with structural profiles. *J Bioinf & Comput Biol*, *10*(3), 1242005.

Onuchic, J. N., Luthey-Schulten, Z., & Wolynes, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annu Rev Phys Chem*, *48*, 545–600.

Onuchic, J. N., & Wolynes, P. G. (2004). Theory of protein folding. *Curr Opinion Struct Biol*, *14*, 70–75.

Ovchinnikov, V., & Karplus, M. (2012). Analysis and elimination of a bias in targeted molecular dynamics simulations of conformational transitions: Application to calmodulin. *J Phys Chem B*, *116*(29), 85848603.

Ozenne, V., Schneider, R., Yao, M., Huang, J. R., Salmon, L., Zweckstetter, M., Jensen, M. R., & Blackledge, M. (2012). Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J Amer Chem Soc*, *134*(36), 15138–15148.

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

Palmieri, L., & Arras, K. O. (2015). Distance metric learning for RRT-based motion planning with constant-time inference. In *Intl Conf Robot Autom (ICRA)*, pp. 637–643, Seattle, WA. IEEE.

Pande, V. S., Beachamp, K., & Bowman, G. R. (2010). Everything you wanted to know about Markov state models but were afraid to ask. *Nat Methods*, *52*(1), 99–105.

Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., & Wolynes, P. G. (2004). Water in protein structure prediction. *Proc Natl Acad Sci USA*, *101*(10), 3352–3357.

Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G., & Noé, F. (2013). Identification of slow molecular order parameters for markov model construction. *J Chem Phys*, *139*(1), 015102.

Perilla, J. R., Goh, B. C., Cassidy, C. K., Liu, B., Bernardi, R. C., Rudack, T., Yu, H., Wu, Z., & Schulten, K. (2015). Molecular dynamics simulations of large macromolecular complexes. *Curr Opin Struct Biol*, *31*, 64–74.

Phillips, D. C. (1967). The hen egg-white lysozyme molecule. *Proc Natl Acad Sci USA*, *57*(3), 483–495.

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *J Comput Chem*, *26*(16), 1781–1802.

Piana, S., Lindorff-Larsen, K., Dirks, R. M., Salmon, J. K., Dror, R. O., & Shaw, D. E. (2012a). Evaluating the effects of cutoffs and treatment of long-range electrostatics in protein folding simulations. *PLoS ONE*, *7*(6), e39918.

Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2012b). Protein folding kinetics and thermodynamics from atomistic simulation. *Proc Natl Acad Sci USA*, *109*(44), 17845–17850.

Plaku, E. (2015). Region-guided and sampling-based tree search for motion planning with dynamics. *IEEE Transactions on Robotics*, *31*(3), 723–735.

Plaku, E., Stamati, H., Clementi, C., & Kavraki, L. E. (2007). Fast and reliable analysis of molecular motions using proximity relations and dimensionality reduction. *Proteins: Struct Funct Bioinf*, *67*(4), 897–907.

Plaku, E., Kavraki, L. E., & Vardi, M. Y. (2010). Motion planning with dynamics by a synergistic combination of layers of planning. *IEEE Transactions on Robotics*, *26*(3), 469–482.

Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Adv Protein Chem*, *66*, 27–85.

Porta, J. M., & Jaillet, L. (2013). Exploring the energy landscapes of flexible molecular loops using higher-dimensional continuation. *J Comput Chem*, *34*(3), 234–244.

Porta, J. M., Thomas, F., Corcho, F., Canto, J., & Perez, J. J. (2007). Complete maps of molecular-loop conformation spaces. *J Comput Chem*, *28*(13), 2170–2189.

Prinz, J. H., Keller, B., & Noé, F. (2011a). Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. *Phys Chem Chem Phys*, *13*(38), 16912–16927.

Prinz, J. H., Wu, H., Sarich, M., Keller, B., Senne, M., Held, M., Chodera, J. D., Schütte, C., & Noé, F. (2011b). Markov models of molecular kinetics: generation and validation. *J Chem Phys*, *134*(17), 174105.

Proctor, A. J., Lipscomb, T. J., Zou, A., Anderson, J. A., & Cho, S. S. (2012). Performance analyses of a parallel verlet neighbor list algorithm for GPU-optimized MD simulations. In *ASE/IEEE Intl Conf Biomed Comput (BioMedCom)*, pp. 14–19, Alexandria, VA. IEEE.

Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol*, *7*, 95–99.

Raveh, B., Enosh, A., Furman-Schueler, O., & Halperin, D. (2009). Rapid sampling of molecular motions with prior information constraints,. *PLoS Comput Biol*, *5*(2), e1000295.

Rodriguez, S., Thomas, S., Pearce, R., & Amato, N. (2006a). RESAMPL: A Region-Sensitive Adaptive Motion Planner. In *Workshop Algorithm Found Robot*, Vol. 47 of *Springer Tracts in Advanced Robotics*, pp. 285–300. Springer, New York, NY.

Rodriguez, S., Tang, X., Lien, J.-M., & Amato, N. M. (2006b). An obstacle-based rapidly-exploring random tree. In *Intl Conf Robot Autom (ICRA)*, pp. 895–900, Orlando, FL. IEEE.

Rohrdanz, M. A., Zheng, W., Maggioni, M., & Clementi, C. (2011). Determination of reaction coordinates via locally scaled diffusion map. *J Chem Phys*, *134*(12), 124116.

Rose, G. D., Fleming, P. J., Banavar, J. R., & Maritan, A. (2006). A backbone-based theory of protein folding. *Proc Natl Acad Sci USA*, *103*(45), 16623–16633.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Roy, R., Hohng, S., & Ha, T. (2008). A practical guide to single-molecule FRET. *Nat Methods*, *5*(6), 507–516.

Rychkova, A., Mukherjee, S., Bora, R. P., & Warshel, A. (2013). Simulating the pulling of stalled elongated peptide from the ribosome by the translocon. *Proc Natl Acad Sci USA*, *110*(25), 10195–10200.

Sánchez, G., & Latombe, J.-C. (2002). On delaying collision checking in PRM planning: Application to multi-robot coordination. *Intl J Robot Res*, *21*(1), 5–26.

Sapin, E., Carr, D. B., De Jong, K. A., & Shehu, A. (2016). Computing energy landscape maps and structural excursions of proteins. *BMC Genomics*, *17*(Suppl 4), 456.

Schlau-Cohen, G. S., Wang, Q., Southall, J., Cogdell, R. J., & Moerner, W. E. (2013). Single-molecule spectroscopy reveals photosynthetic LH2 complexes switch between emissive states. *Proc Natl Acad Sci USA*, *110*(27), 10899–10903.

Schotte, F., Lim, M., Jackson, T. A., Smirnov, A. V., Soman, J., Olson, J. S., Phillips, G. N., Wulff, M., & Anfinrud, P. A. (2003). Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. *Science*, *300*(5627), 1944–1947.

Schuyler, A. d., Jernigan, R. L., Wasba, P. K., Ramakrishnan, B., & Chirikjian, G. S. (2009). Iterative cluster-NMA (icnma): a tool for generating conformational transitions in proteins. *Proteins: Struct Funct Bioinf*, *74*(3), 760–776.

Shatsky, M., Nussinov, R., & Wolfson, H. J. (2002). Flexible protein alignment and hinge detection. *Proteins*, *48*(2), 242–256.

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., Jumper, J. M., Salmon, J. K., Shan, Y., & Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. *Science*, *330*(6002), 341–346.

Shehu, A. (2009). An *ab-initio* tree-based exploration to enhance sampling of low-energy protein conformations. In *Robot: Sci and Sys*, pp. 241–248, Seattle, WA, USA.

Shehu, A. (2013). Probabilistic search and optimization for protein energy landscapes. In Aluru, S., & Singh, A. (Eds.), *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series.

Shehu, A., Clementi, C., & Kavraki, L. E. (2006). Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins: Struct Funct Bioinf*, *65*(1), 164–179.

Shehu, A., Clementi, C., & Kavraki, L. E. (2007). Sampling conformation space to model equilibrium fluctuations in proteins. *Algorithmica*, *48*(4), 303–327.

Shehu, A., & Kavraki, L. E. (2012). Modeling structures and motions of loops in protein molecules. *Entropy J*, *14*(2), 252–290.

Shehu, A., Kavraki, L. E., & Clementi, C. (2007). On the characterization of protein native state ensembles. *Biophys J*, *92*(5), 1503–1511.

Shehu, A., Kavraki, L. E., & Clementi, C. (2008). Unfolding the fold of cyclic cysteine-rich peptides. *Protein Sci*, *17*(3), 482–493.

Shehu, A., Kavraki, L. E., & Clementi, C. (2009). Multiscale characterization of protein conformational ensembles. *Proteins: Struct Funct Bioinf*, *76*(4), 837–851.

Shehu, A., & Olson, B. (2010). Guiding the search for native-like protein conformations with an *ab-initio* tree-based exploration. *Intl J Robot Res*, *29*(8), 1106–11227.

Shkolnik, A., Walter, M., & Tedrake, R. (2009). Reachability-guided sampling for planning under differential constraints. In *Intl Conf Robot Autom (ICRA)*, pp. 2859–2865.

Shlens, J. (2003). A tutorial on principal component analysis. https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf.

Shukla, D., Hernández, C. X., Weber, J. K., & Pande, V. S. (2015). Markov state models provide insights into dynamic modulation of protein function. *Acc Chem Res*, *48*(2), 414–422.

Singh, A. P., Latombe, J.-C., & Brutlag, D. L. (1999). A motion planning approach to flexible ligand binding. In Schneider, R., Bork, P., Brutlag, D. L., Glasgow, J. I., Mewes, H.-W., & Zimmer, R. (Eds.), *Intl Conf Intell Sys Mol Biol (ISMB)*, Vol. 7, pp. 252–261, Heidelberg, Germany. AAAI.

Singhal, N., & Pande, V. S. (2005). Error analysis and efficient sampling in markovian state models for molecular dynamics. *J Chem Phys*, *123*(20), 204909–1–204909–13.

Singhal, N., Snow, C. D., & Pande, V. S. (2004). 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*(1), 415–425.

Socher, E., & Imperiali, B. (2013). FRET-CAPTURE: A sensitive method for the detection of dynamic protein interactions. *Chem Biochem*, *14*(1), 53–57.

Song, G., & Amato, N. M. (2000). A motion-planning approach to folding: From paper craft to protein folding. Tech. rep. TR00-001, Department of Computer Science, Texas A & M University.

Song, G., & Amato, N. M. (2004). A motion planning approach to folding: From paper craft to protein folding. *IEEE Trans Robot Autom*, *20*(1), 60–71.

Song, J., & Zhuang, W. (2014). Simulating the peptide folding kinetic related spectra based on the Markov state model. In *Protein Conformational Dynamics*, Vol. 805 of *Adv Exp Med Biol*, pp. 199–220. Springer.

Soto, C. (2008). Protein misfolding and neurodegeneration. *JAMA Neurology*, *65*(2), 184–189.

Stadler, P. (2002). Fitness landscapes. *Appl Math & Comput*, *117*, 187–207.

Stone, J. E., Phillips, J. C., Freddolino, P. L., Hardy, D. J., Trabuco, L. G., & Schulten, K. (2007). Accelerating molecular modeling applications with graphics processors. *J Comput Chem*, *28*(16), 2618–2640.

Sucan, I. A., & Kavraki, L. E. (2012). A sampling-based tree planner for systems with complex dynamics. *IEEE Trans Robotics*, *28*(1), 116–131.

Sun, Z., Hsu, D., Jiang, T., Kurniawati, H., & Reif, J. (2005). Narrow passage sampling for probabilistic roadmap planners. *IEEE Trans Robotics*, *21*(6), 1105–1115.

Tama, F., & Sanejouand, Y. H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein Eng*, *14*(1), 1–6.

Tama, F., Valle, M., Frank, J., & Brooks, C. L. (2003). Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc Natl Acad Sci USA*, *100*(16), 9319–9323.

Tang, X., Kirkpatrick, B., Thomas, S., Song, G., & Amato, N. (2005). Using motion planning to study rna folding kinetics. *J Comput Biol*, *12*(6), 862–881.

Tang, X., Thomas, S., Tapia, L., Giedroc, D. P., & Amato, N. (2008). Simulating rna folding kinetics on approximated energy landscapes. *J Mol Biol*, *381*(4), 1055–1067.

Tanner, D. E., Phillips, J. C., & Schulten, K. (2012). GPU/CPU algorithm for generalized born/solvent-accessible surface area implicit solvent calculations. *J Chem Theory Comput*, *8*(7), 2521–2530.

Tapia, L., Tang, X., Thomas, S., & Amato, N. (2007). Kinetics analysis methods for approximate folding landscapes. *Bioinformatics*, *23*, i539i548.

Tapia, L., Thomas, S., & Amato, N. (2010). A motion planning approach to studying molecular motions. *Commun Inf Sys*, *10*(1), 53–68.

Teknipar, M., & Zheng, W. (2010). Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model. *Proteins: Struct Funct Bioinf*, *78*(11), 2469–2481.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Teodoro, M., Phillips, G. N. J., & Kavraki, L. E. (2003). Understanding protein flexibility through dimensionality reduction. *J Comput Biol*, *10*(3-4), 617–634.

Thomas, S., Song, G., & Amato, N. M. (2005). Protein folding by motion planning. *J. Phys. Biol.*, *2*(4), 148.

Thomas, S., Tang, X., Tapia, L., & Amato, N. M. (2007). Simulating protein motions with rigidity analysis. *J. Comput. Biol.*, *14*(6), 839–855.

Thorpe, M. F., & Ming, L. (2004). Macromolecular flexibility. *Phil. Mag.*, *84*(13-16), 1323–31137.

Torella, J. P., Holden, S. J., Santoso, Y., Hohlbein, J., & Kapanidis, A. N. (2011). Identifying molecular dynamics in single-molecule FRET experiments with burst variance analysis. *Biophys J*, *100*(6), 1568–1577.

Tsai, C., Kumar, S., Ma, B., & Nussinov, R. (1999a). Folding funnels, binding funnels, and protein function. *Protein Sci*, *8*(6), 1181–1190.

Tsai, C., Ma, B., & Nussinov, R. (1999b). Folding and binding cascades: shifts in energy landscapes. *Proc Natl Acad Sci USA*, *96*(18), 9970–9972.

Uversky, V. N. (2009). Intrinsic disorder in proteins associated with neurodegenerative diseases. In *Protein Folding and Misfolding: Neurodegenerative Diseases*, Vol. 14 of *Focus on Structural Biology*, pp. 5188–5238. Springer.

van den Bedem, H., Lotan, I., Latombe, J.-C., & Deacon, A. M. (2005). Real-space protein-model completion: an inverse-kinematics approach. *Acta Crystallogr*, *D61*(1), 2–13.

van der Maaten, L. J. P., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *J Mach Learn Res*, *10*(1-41), 66–71.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J Comput Chem*, *26*(16), 1701–1718.

van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P., & Tironi, I. G. (1996). Biomolecular simulation: the gromos96 manual and user guide. http://www.gromos.net/.

Verlet, L. (1967). Computer "experiments" on classical fluids. i. thermodynamical properties of Lennard-Jones molecules. *Phys Rev Lett*, *159*, 98–103.

Warshel, A. (2003). Computer simulations of enzyme catalysis: Methods, progress, and insights. *Annu Rev Biophys Biomol Struct*, *32*, 425–443.

Warshel, A., & Levitt, M. (1976). Theoretical studies of enzymatic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Biol*, *103*(2), 227–249.

Weber, J. K., Jack, R. L., & Pande, V. S. (2013). Emergence of glass-like behavior in markov state models of protein folding dynamics. *J Amer Chem Soc*, *135*(15), 5501–5504.

Wells, S., Menor, S., Hespenheide, B., & Thorpe, M. F. (2005). Constrained geometric simulation of diffusive motion in proteins. *J Phys Biol*, *2*(4), 127–136.

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Struct Funct Bioinf*, *80*(7), 1715–1735.

Yang, H., Wu, H., Li, D., Han, L., & Huo, S. (2007). Temperature-dependent probabilistic roadmap algorithm for calculating variationally optimized conformational transition pathways. *J Chem Theory Comput*, *3*(1), 17–25.

Yang, L., Song, G., Carriquiry, A., & Jernigan, R. L. (2008). Close correspondence between the essential protein motions from principal component analysis of multiple HIV-1 protease structures and elastic network modes. *Structure*, *16*(2), 321–330.

Yang, Z., Mâjek, P., & Bahar, I. (2009). Allosteric transitions of supramolecular systems explored by network models: Application to chaperonin GroEL. *PLoS Comput Biol*, *5*(4), e1000360.

Yao, P., Dhanik, A., Marz, N., Propper, R., Kou, C., Liu, G., van den Bedem, H., Latombe, J. C., Halperin-Landsberg, I., & Altman, R. B. (2008). Efficient algorithms to explore conformation spaces of flexible protein loops. *IEEE/ACM Trans Comput Biol Bioinf*, *5*(4), 534–545.

Zagrovic, B., Snow, C. D., Shirts, M. R., & Pande, V. S. (2002). Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J Mol Biol*, *323*(5), 927–937.

Zhang, M., & Kavraki, L. E. (2002a). Finding solutions of the inverse kinematics problem in computer-aided drug design. In Florea, L., Walenz, B., & Hannenhalli, S. (Eds.), *Currents in Computational Molecular Biology*, pp. 214–215, Washington, D.C. ACM.

Zhang, M., & Kavraki, L. E. (2002b). A new method for fast and accurate derivation of molecular conformations. *Chem Inf Comput Sci*, *42*(1), 64–70.

Zhao, G., Perilla, J. R., Yufenyuy, E. L., Meng, X., Chen, B., Ning, J., Ahn, J., Gronenborn, A. M., Schulten, K., Aiken, C., & Zhang, P. (2013). Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, *497*(7451), 643–646.

Zheng, W., & Brooks, B. (2005). Identification of dynamical correlations within the myosin motor domain by the normal mode analysis of an elastic network model. *J Mol Biol*, *346*(3), 745–759.

Zheng, W., Brooks, B. R., & Hummer, G. (2007). Protein conformational transitions explored by mixed elastic network models. *Proteins: Struct Funct Bioinf*, *69*(1), 43–57.

Zheng, W., & Doniach, S. (2003). A comparative study of motor-protein motions by using a simple elastic-network model. *Proc Natl Acad Sci USA*, *100*(23), 13253–13258.

Zheng, W., Rohrdanz, M. A., & Clementi, C. (2013). Rapid exploration of configuration space with diffusion-map-directed molecular dynamics. *J Phys Chem B*, *117*(42), 12769–12776.

Zheng, W., Rohrdanz, M. A., Maggioni, M., & Clementi, C. (2011). Polymer reversal rate calculated via locally scaled diffusion map. *J Chem Phys*, *134*(14), 144109.

Zhou, H. (2014). Theoretical frameworks for multiscale modeling and simulation. *Curr Opinion Struct Biol*, *25*, 67–76.