Christopher Madden
Peter Bohnenkamp
Kazem Kazerounian
Horea T. Ilieş

Department of Mechanical Engineering,
University of Connecticut,
USA
peytah@gmail.com
{kazem, ilies}@engr.uconn.edu

Residue Level Three-dimensional Workspace Maps for Conformational Trajectory Planning of Proteins

Abstract

The function of a protein macromolecule often requires conformational transitions between two native conformations. Understanding these transitions is essential to the understanding of how proteins function, as well as to the ability to design and manipulate protein-based nanomechanical systems. In this paper we propose a set of 3D Cartesian workspace maps for exploring protein pathways. These 3D maps are constructed in the Euclidean space for triads of chain segments of protein molecules that have been shown to have a high probability of occurrence in naturally observed proteins based on data obtained from more than 38,600 proteins from the Protein Data Bank (PDB). We show that the proposed 3D propensity maps are more effective navigation tools than the propensity maps constructed in the angle space. We argue that the main reason for this improved efficiency is the fact that, although there is a one-to-one mapping between the 2D dihedral angle maps and the 3D Cartesian maps, the propensity distributions are significantly different in the two spaces. Hence, the 3D maps allow the pathway planning to be performed directly in the 3D Euclidean space based on propensities computed in the same space, which is where protein molecules change their conformations.

KEY WORDS—protein, kinematic pathway, trajectory planning, conformation, transition, energy landscape.

1. Introduction

Proteins are nature's nanorobots. The function of these natural nanorobots often requires conformational transitions between two or more native conformations[1] that are made possible by the intrinsic mobility of the proteins. Such functional transitions occur, for example, in allosteric transitions in enzymes (Kern and Zuiderweg 2003), force generations in motor proteins (Geeves and Holmes 1999) and the conformational changes induced by ligand binding to various enzymes and receptors (Liu et al. 2008). Understanding these transitions is essential to the understanding of how proteins function (Zheng et al. 2007), as well as to the ability to design and manipulate protein-based nanomechanical systems (Chirikjian et al. 2005; Madden et al. 2008).

Macroscopic robotic manipulators have relatively few components and, therefore, a rather small number of variables that define their spatial configurations. On the other hand, the transitional pathway of a protein moving between two native conformations is significantly more complex due to the very large number of "motion variables" and is therefore a computationally enormous task. Moreover, the stimuli external to the protein that cause these conformational changes are not understood completely. Nevertheless, the paramount importance of the conformational transitions in biological functions requires models based on first principles that could be both practical and valuable in understanding the transition pathways of proteins.

The transitional trajectories in the large majority of the functional proteins have not been observed and no experimental data is available for comparison. However, it is generally agreed that functional proteins have two or more native structures that are relatively close in terms of the corresponding potential energy values (Gibbs et al. 2001; Itoh and Sasai 2004;

The International Journal of Robotics Research
Vol. 28, No. 4, April 2009, pp. 450–463
DOI: 10.1177/0278364908098092

Figures 4, 9–11 appear in color online: http://ijr.sagepub.com

1. A *native conformation* or *native structure* of functional proteins is a conformation corresponding to either a global minimum potential energy state, or a stable, local minimum energy state that is observed in nature and is related to a specific function.

Scheraga et al. 2004; Miyashita et al. 2005). Furthermore, it takes a relatively small amount of energy[2] to trigger the transition from one conformation to another. It is common to assume that the pathway between native conformations is therefore a "valley" in the potential energy landscape. To understand such pathways, a detailed description of the kinematic motion of the molecule as well as the energy landscape corresponding to the kinematic structure is needed. Numerous methods have been developed to describe the conformational transition of the proteins along preferred energy pathways. A straightforward attempt is linear interpolation of the native conformations in either the angle space or the space of atom coordinates (Gerstein and Krebs 1998), possibly followed by subsequent energy minimization (Krebs et al. 2003). While these methods are useful in visualization of the conformational transition, they do not represent the physical motion of the protein. More complex approaches include the introduction of artificial potential forces in conjunction with molecular dynamics simulation (MDS) to force the protein motion from one conformation to another (Schlitter et al. 1994; Guilbert et al. 1995). However, the enormous computational requirement of the MDS (see Pande (2005)) severely limits the applicability of these methods, and the accuracy of assuming large potential forces for guiding the conformation has not been quantified.

Traditional engineering-based methods have proven very useful in studying the conformational pathways. One such method is normal modal analysis of the protein structures (Guilbert et al. 1996; Tama and Sanejouand 2001; Tama et al. 2002). While effective, due to the linear nature of modal analysis, such methods are, at best, local in the time domain. By developing elastic models of the protein as a network of mass and springs, the efficacy of normal modal analysis is expanded globally throughout the conformational transition (Haliloglu et al. 1997; Kim et al. 2003; Schuyler and Chirikjian 2005). One other class of approach searching for the protein conformational pathways, which is deeply rooted in engineering applications, is based on robot motion planning algorithms (Amato and Song 2002; Cortes et al. 2005; Wells et al. 2005; Tapia et al. 2007). Such approaches have been remarkably successful in navigating the environment surrounding a robot, partly because that environment is assumed to be known. In the context of protein motion, the challenges faced by these approaches stem from the facts that the energy landscape is essentially a high-dimensional space whose dimension is almost always much larger than the dimension of the spaces encountered in the typical robot motion planning problems.

In a recent work (Madden et al. 2008), we developed a new method for interactive planning of transitional pathways of a given protein by using dihedral angle combinations that have been shown to have a high probability of occurrence in naturally observed proteins. In Madden et al. (2008) we used statistical propensity torque maps[3] for pairs of dihedral angles that capture the probability of occurrence of specific dihedral angle combinations in nature. These maps were constructed in the angle space, similar to the Ramachandran charts, but are based on data obtained from more than 38,600 proteins from the Protein Data Bank (PDB) (Berman et al. 2002) so that each map corresponds to pairs of dihedral angles (φ_i, ψ_i), (φ_i, ψ_{i+1}), $(\varphi_i, \varphi_{i+1})$ and (ψ_i, ψ_{i+1}).

1.1. Objectives

In this work the information from the PDB (see http://www.wwpdb.org) is used to construct 3D workspace maps for three link segments of a protein chain (also called *triad chain segments* or simply triads). These maps live in the 3D Euclidean space and correspond to workspaces reachable by the "end-effector" of the triad. Since our maps are constructed based on PDB data relying on experimental observations of proteins, they are also statistical propensity maps for the natural protein molecules, and can be used as effective guidelines in exploring the pathways of the conformational proteins. Here we explore the effectiveness of these 3D propensity maps by performing interactive explorations of the energy landscape, but automating the space exploration procedure based on these maps is clearly a feasible approach.

One important advantage of the 3D maps compared with the 2D maps detailed by Madden et al. (2008) is that each point of such a 3D map corresponds (via inverse kinematics) to three angles of the protein chain so that larger segments of the proteins can be manipulated. Furthermore, the angles defining a protein conformation can be considered as a parametrization of the protein moving in 3D space. Since the parametric and Euclidean spaces do not share the same metric, a uniform distribution in the parametric space results in an often highly non-uniform distribution in the Euclidean space (e.g. conformations that are not an "equal distance" apart, for example, measured in terms of root mean square distance (RMSD) values or other shape similarity measures). A simple analogy is that of a parametric surface in 3D space: moving between equally spaced points in the parametric space will usually produce curve segments on the surface having non-equal lengths. Hence, the fact that proteins move in the 3D space suggests that this space may be more suitable for planning protein pathways. At the same time, the 3D maps have the highest dimension that can be visualized during the exploration of the energy landscape. Our numerical experiments indicate that the 3D maps are significantly more effective than the 2D torsion angle work envelopes in finding the low-energy pathways between two conformations of functional protein molecules. The path obtained from the torsion angle work envelope maps (Madden et al. 2008) is used as an initial estimate.

2. "Small" compared with the energy needed to perturb the conformation in other "directions".

3. In the robotics literature, a concept similar to our propensity maps is known as the end-effector work envelopes.

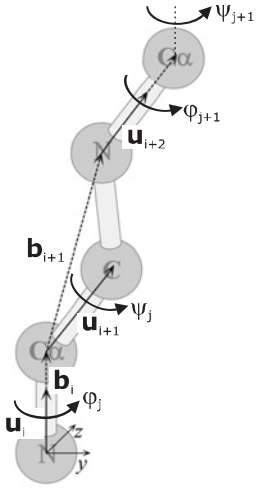


Fig. 1. A triad segment of an amino acid chain with dihedral angles as generalized coordinates.

2. Protein Model

2.1. Protein Molecule as a Kinematic Chain

The authors have successfully modeled the protein molecule as a kinematic chain of rigid bodies connected by revolute joints (Kazerounian et al. 2005a,b; Subramanian and Kazerounian 2006; Bohnenkamp et al. 2007; Subramanian and Kazerounian 2007a,b). Our model fully defines the kinematic structure of the backbone (main chain) of the protein polymer at the zero position as follows. The backbone of a protein chain that contains M residues (amino acids) is defined as a serial chain of $N+1$ solid links connected by N revolute joints (see, for example, Figure 1). Since each residue has two revolute joints $N = 2M$. The revolute joints in each amino acid are φ_i and $\psi_i, i = \overline{1, M}$. We introduce a uniform notation of these angles as θ_j, $j = \overline{1, N}$, where

$$\theta_j = \varphi_i, \tag{1}$$

$$\theta_{j+1} = \psi_i, \quad \text{for all} \quad j = 2(i-1) \quad \text{and} \quad i = \overline{1, M}. \tag{2}$$

The N revolute joints and solid links connecting them within the N degree of freedom serial linkage are kinematically defined by a set of unit vectors $\mathbf{u}$ and body vectors $\mathbf{b}$.

- $\mathbf{u}_{0j1}$: unit vectors along the dihedral joints θ_j. The first index indicates that we use the zero-position configuration, the second index corresponds to the joint number and the third index indicates the chain number (e.g. the index of the main chain is 1; side chains have indices larger than 1). Thus, $\mathbf{u}_{0j1}$ corresponds to the unit vector along the ith joint of the main chain. In this paper we only focus on the main chain and, therefore, the third index will always be equal to 1.

- $\mathbf{b}_{0j1}$: body vector connecting a point on the dihedral joint axis of the angle θ_j (specifically, a nitrogen atom if j is odd and α-carbon atom if j is even) to a point on the dihedral joint axis of the angle θ_{j+1} (an α-carbon if j is odd and a nitrogen atom of the next residue if j is even). The indices have the same meaning as defined above.

Note that in this configuration (the zero position, which is *not* a native conformation) all dihedral angles θ_j are defined to be zero. Furthermore, our selection for the biological reference conformation as our zero configuration implies that the pair of body vectors $\mathbf{b}_{0j1}$ and $\mathbf{b}_{0(j+1)1}$ are identical for all residues in the chain in this conformation. Thus, $\mathbf{u}_{0j1}$ and $\mathbf{b}_{0j1}$) need to be defined only for one residue.

The backbone structure of the protein in any non-zero configuration (when the values of the dihedral angles θ_j are non-zero) can be found by a series of successive rotations:

$$\mathbf{b}_{j1} = [R_{\theta_1,\mathbf{u}_{011}}][R_{\theta_2,\mathbf{u}_{021}}]\ldots[R_{\theta_j,\mathbf{u}_{0j1}}]\mathbf{b}_{0j1}, \tag{3}$$

where j denotes the jth solid link in the protein chain and R is the screw rotation matrix of angle θ_j about the axis $\mathbf{u}_{0j1}$. In every peptide plane, assumed to be a rigid body in our formulation, the above equations result in the computation of two body vectors $\mathbf{b}_{j1}$ and $\mathbf{b}_{(j+1)1}$. A similar procedure is also implemented for the side chains.

The reduction in the computational time between the linkage model that we use and the standard all-atom models is tremendous. If we assume that each amino residue in the chain, on average, has 10 atoms (most have more), then the number of optimization variables in the majority of *ab initio* methods is reduced from $10 \cdot 3 \cdot N$ to $2N$ variables. Considering that most optimization methods have a computational cost proportional to the square of the number of variables (global search based methods such as simulated annealing and genetic algorithms excluded), the expected reduction in computational time (number of cost evaluations) is a factor of 15^2 or 225. In addition, in evaluating the cost function, the bond and angle energies are no longer needed. This argument does not take into account the computational time of the direct kinematics.

2.2. Successive Kineto-static Fold Compliance

In our approach, known as the successive kineto-static fold compliance method, the conformational changes of the peptide chain are driven by an inter-atomic force field without the need for MDS. Instead, the chain complies under the kineto-static effect of the force field in such a manner that each rotatable joint changes by an amount proportional to the effective torque

on that joint. This process successively iterates until all of the joint torques have converged to zero. The resulting conformation is in a minimum potential energy state. PROTOFOLD, our own protein simulation platform, uses this methodology. It has been shown that PROTOFOLD is orders of magnitude more efficient and robust than traditional MDS (Su et al. 2007).

The main steps in successive kineto-static fold compliance are therefore described as follows.

1. At a given set of joint angles, calculate the Cartesian coordinates of all atoms in the protein molecule (direct kinematics using the zero-position formulation discussed above).

2. Calculate all of the inter-atomic forces in this conformation (using the AMBER force field model discussed below).

3. Calculate the equivalent joint torques (τ_j) using the well-known relation between the end-effector forces and the joint forces in robotics.

4. Calculate an effective change in each joint variable, proportional to that joint's equivalent torque τ_j ($\Delta\theta_j = k\tau_j$) and rotate each joint accordingly.

5. Go back to step 1 until all of the joint equivalent torques have diminished to zero (within some small prescribed error).

2.3. AMBER Potential and Force Field

PROTOFOLD employs the AMBER atomic force model (Cornell et al. 1996) to describe the system energy of a given conformation. AMBER describes the total potential energy of the protein chain as a sum of the electrostatic and van der Waals energies between all atoms. Note that there are several terms in the force field that represent the bond angle and length energy changes (spring like effects) that are automatically eliminated due to the rigid body assumption of the peptide planes. Therefore,

$$E_{\text{potential}} = \sum_{i<j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}, \right) \tag{4}$$

where A_{ij} and B_{ij} are the van der Waals and London dispersion terms; q_i and q_j are the partial atomic charges, and ε is the dielectric constant (see Cornell et al. (1996) and Duan et al. (2003)).

The force between any two atoms is the negative derivative of the potential energy between those same atoms. Thus,

$$F_{ij} = -\frac{12A_{ij}}{r_{ij}^{13}} + \frac{6B_{ij}}{r_{ij}^{7}} - \frac{q_i q_j}{\varepsilon r_{ij}^{2}}. \tag{5}$$

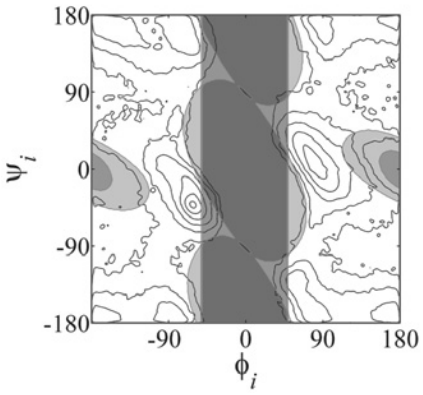


Fig. 2. Outline of sterically prohibited angle space for glycine overlaid onto dihedral angle density propensity from PDB.

The combined force applied to each atom i is then the sum of the forces exerted by all other atoms j:

$$F_i = \sum_{j=1}^{N} F_{ij}, \quad i \neq j. \tag{6}$$

3. Generating the 2D Data Maps

The protein backbone structure in the serial chain kinematic model is uniquely defined by the set of dihedral angles φ (the amino angle) and ψ (the carboxyl angle) for each amino acid (residue). The set of these angles for all residues in the chain constitute the generalized coordinates of the backbone. In this work, we have modeled 20 of the 22 amino acids known to exist in the nature (the discovery of the last two has only been announced recently).

The Ramachandran $\varphi - \psi$ maps (Ramachandran et al. 1963) have traditionally been generated using the union of the van der Waals rigid sphere collision models. Alternatively, such maps encoding the statistical distribution of angle combinations observed in nature could be generated from the experimentally observed data found in the PDB, which contains information on more than 40,000 proteins. While both techniques result in analogous map contours, PDB-based maps contain details of the population propensity. This is illustrated in Figure 2 where we superimpose the maps obtained via the two methods mentioned above for a glycine residue: the shaded areas are obtained from the collision model, and the contour curves (isocurves) have been computed based on PDB data.

Furthermore, observe that the typical Ramachandran chart (shaded areas in Figure 2) is inadequate for computing propensity regions for all of the dihedral angles in all amino acid sequences. Therefore, we have developed torque charts for the dihedral angle sets (φ_i, ψ_i), (φ_i, ψ_{i+1}), $(\varphi_i, \varphi_{i+1})$ and

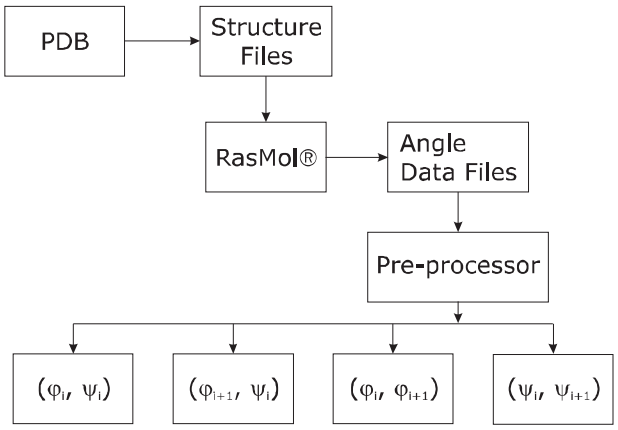


Fig. 3. Process overview.

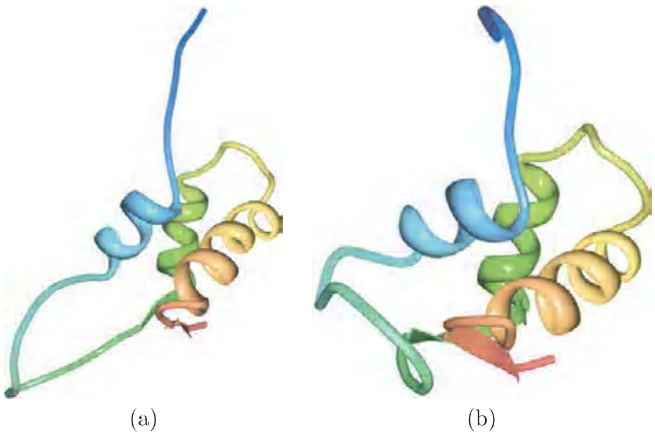


Fig. 4. 1FOX ribbon plot; (b) 2FOW ribbon plot.

(ψ_i, ψ_{i+1}). Since we modeled 20 amino acids, 400 possible sequences exists for a pair of residues. Therefore, the total number of maps in our master collection is 1,220 $(20 + 400 + 400 + 400)$ maps, which were developed based on 38,642 proteins from PDB.

Figure 3 illustrates the data parsing process. The protein structures available in the PDB at the time this work was in progress (38,642 at the time), have been downloaded and the structure files for DNA, RNA, and other nucleic acid/protein combinations were excluded. The main chain angles for each protein chain were computed and the dihedral angles were determined by using RasMol (Sayle and Milner-White 1995), an open-source protein visualization tool. During this process, some structure files failed to produce meaningful data, which we believe is mainly due to rare errors in the PDB files (54 files out of all downloaded files). Each entry in our database contains dihedral angle data on the structure file for each protein; this included multiple entries for proteins that had multiple models listed in the PDB. The angle data was organized into data distribution matrices, which consisted of 360×360 grids; each cell location representing a pairs of adjacent torsion angles from -180° to 180° As the data was sampled for the four sets of dihedral angle pairs (i.e. (φ_i, ψ_i), (φ_i, ψ_{i+1}), $(\varphi_i, \varphi_{i+1})$ and (ψ_i, ψ_{i+1})), the corresponding cells in the data distribution matrices were incremented. The value in each cell of 360×360 grid represents the frequency of occurrence of that particular combination of torsion angles in the PDB. Data for over six million angle pairs was collected in this fashion. In the final step, the raw data matrices were normalized and smoothed using a convolution algorithm from MATLAB. Figure 5 illustrates four sample dihedral angle propensity maps. Darker regions indicate higher population density observed in nature and, hence, more favorable energy conformations. The resolution of the final predicted trajectories is dependent on the grid resolution.

4. Dihedral Angle Propensity Maps as Navigation Guides

Our numerical experiments focused on the protein 1FOX, whose two known native conformations are shown in Figure 4. These two native conformations, known as 1FOX and 2FOW, have been computed based on data obtained from PDB.

Observe that PDB contains a set of dihedral angles for the backbone for each of the two native conformations. However, our dihedral angles are slightly different than those found in PDB due to both our serial kinematic chain model of connected rigid bodies as well as the improved numerical model of the peptide planes (Subramanian and Kazerounian 2006), but the differences are negligible for the purpose of this work. There are 76 residues in this protein molecule, which result in 152 dihedral angles for the backbone. Consequently, there are a total number of 213 torsion propensity maps.

A point in each of these maps corresponds to a specific set of values for the corresponding angles. Therefore, each of the two native conformations 1FOX and 2FOW are represented in each of our maps by one point in the torsion angle workspace. Each curve connecting these two points (for a given map), represents a pathway for that particular set of dihedral angles. Then the pathway taking the protein from one native conformation to the other will be obtained by combining all these angle level pathways. In (Madden et al. 2008) we argued that such a pathway can be found by using dihedral angle combinations that have been shown to have a high probability of occurrence in naturally observed proteins. We proposed there statistical propensity maps for tuples of dihedral angles that capture the probability of occurrence of specific dihedral angle combinations in nature.

Planning a path using these dihedral angle maps is a multistep process. First, we connect every two points (torsion angle pair) by a straight line, which is the basis of most of the existing visualization techniques for conformational tran-

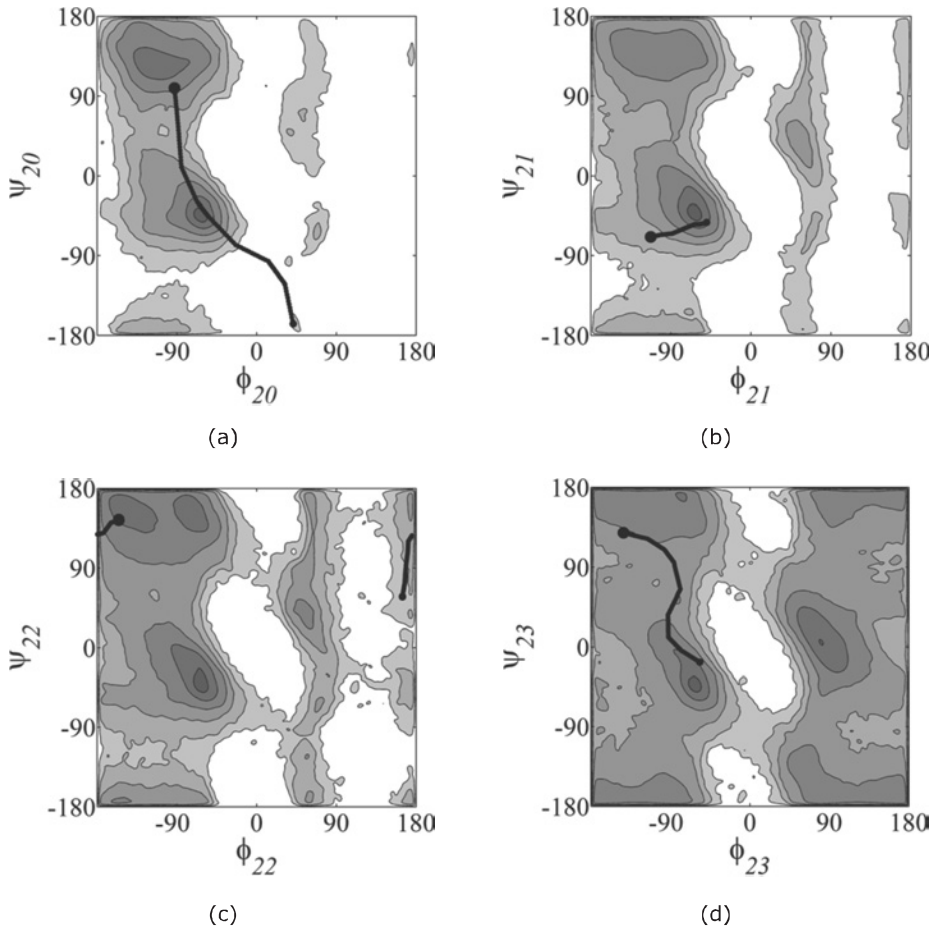


Fig. 5. Sample conformational energy favorable transition pathway of the dihedral angles obtained through navigating the torsion angle work envelopes for (1FOX to 2FOW).

sitions. Second, we select the direction of the linear paths in the angle space to obtain a linear trajectory that is more energy favorable. Through visual inspection of the charts, the user interactively forces the paths to go through lower energy conformations as indicated by higher population intensity (darker regions). In our computer implementation, the trajectory modification is possible through the introduction of additional trajectory points, or by dragging existing trajectory points to new locations. Automating this step is clearly feasible, and represents the next phase of this work.

The larger circle at one end of the path in each chart (such as those shown in Figure 5) indicates the initial conformation (i.e. 1FOX). It can be clearly seen that the portions of the path on higher energy domains (lighter or less-populated regions) indicate higher total potential energy for the transitional conformation at that point on the pathway. To fulfill the objective of obtaining a path of minimal energy between two low-energy end conformations, the interactively designed pathways described above need to be further optimized. The kineto-static compliance method implemented in PROTOFOLD (Kazerounian et al. 2005a,b) is applied to the curvilinear angle space pathways. PROTOFOLD can more finely tune the pathways by rotating the joint angles that will relieve the highest remaining joint torques. Figure 6 illustrates how the linearly interpolated, planned and optimized paths in angle space relate. The contours shown are very simple representation of the complex energy landscape between the start and end conformations. If the end goal is to minimize the overall energy rise over the entire conformational motion, then the path should follow an isoenergy contour. The final pathway in torsion angle space is shown in Figure 6.

Figure 7 shows the energy profile of the various pathways that we have generated by using these dihedral angle propensity maps. The potential energy is directly related to the equivalent joint torques calculated in PROTOFOLD. In Figure 7, the Euclidean norm of the torque vector (summation of the

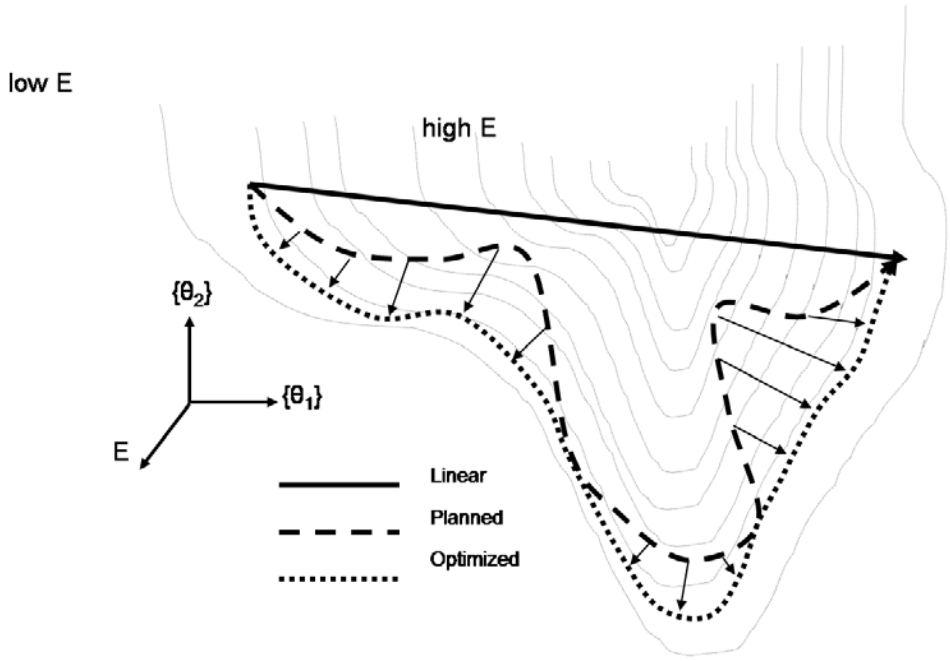


Fig. 6. 3D contour representation of the energy landscape with various angle space paths.

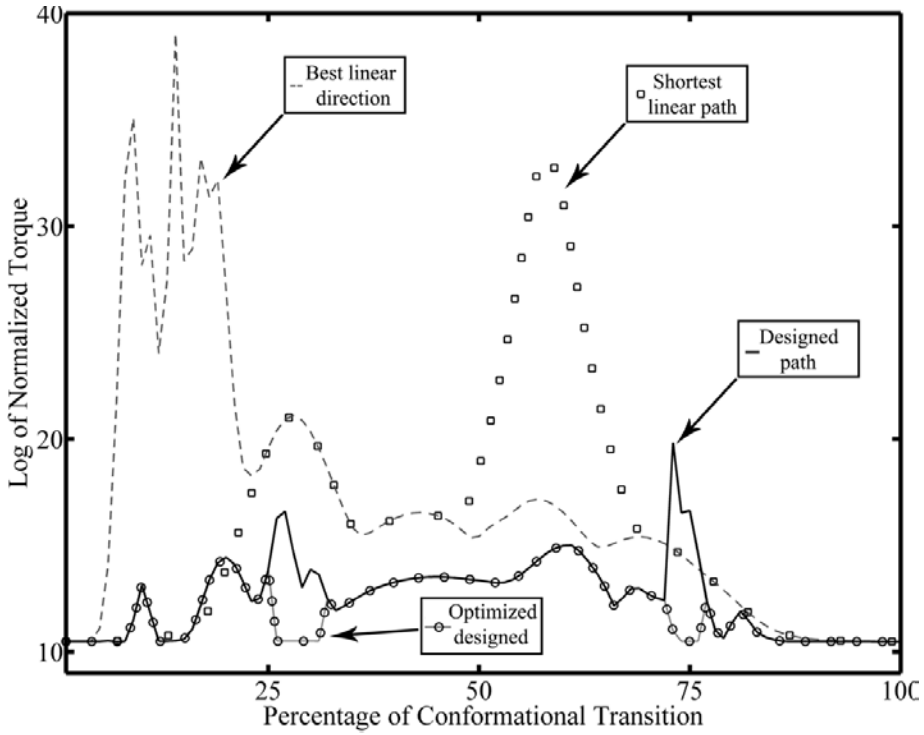


Fig. 7. Simulated pathways.

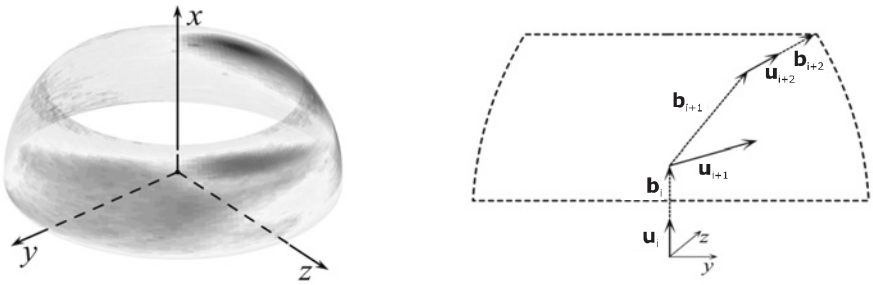


Fig. 8. Sample Cartesian workspace of a triad (three-link segment) of the peptide chain from Figure 1 and a cross section with a plane that contains the x-axis.

square of all joint torques) is evaluated at discrete points along the pathway. As seen in Figure 7, the best direction linear paths show some improvements over the shortest linear paths in most, but not all, segments of the pathway.

5. Three-dimensional Triad Workspaces

The 2D dihedral angle workspaces proved to be an effective tool in finding a low-energy path for the conformational transition of the protein chain. Path planning in this dihedral space amounts to navigating through the regions of low energy in the angle space. However, since these workspaces are formed due to the interference of the atoms in the 3D Cartesian space where the folding takes place, planning the path directly in this Cartesian space is not only more intuitive, but also more efficient. Hence, we propose a set of 3D Cartesian workspaces (or Cartesian maps) comparable to robotic manipulator work envelopes.

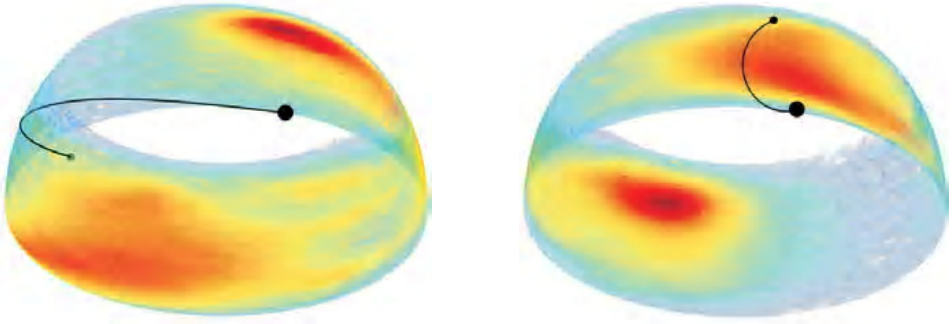
These workspaces are developed for triad segments of amino acid chains with three dihedral angles as generalized coordinates. In kinematics, a triad is a three-link serial chain connecting two spatial points. Figure 1 shows a repeating triad of a peptide chain. It is evident from this figure that the third joint value does not influence the position of the end-effector (Cα). In other words, the position of the end α-carbon is the same for all possible values of the angle around vector $\mathbf{b}_{j+2}$ (not shown). Following the zero-position notation used in PROTOFOLD, unit vectors $\mathbf{u}_j$, $\mathbf{u}_{j+1}$ and $\mathbf{u}_{j+2}$ indicate revolute joint directions (dihedral angles), and $\mathbf{b}_j$, $\mathbf{b}_{j+1}$ and $\mathbf{b}_{j+2}$ are body vectors connecting the points corresponding to the joints. As discussed in Section 2, in the zero-position configuration all dihedral angles are zero, and vectors $\mathbf{u}_{0j1}$ and $\mathbf{b}_{0j1}$ only need to be defined for one residue, whereas the corresponding vectors for the main chain body vectors are calculated using the bond lengths for and bond angles between each pair of consecutive atoms. Note that the bond length gives us the magnitude of the body vectors. We set the coordinates of the first nitrogen in the triad (in other words, the nitrogen contained within the amino group) to the origin. This is explained in detail in Kazerounian et al. (2005a,b). The backbone structure of the protein in any non-zero configuration (i.e. for non-zero dihedral angles θ_j) can be found by a series of successive rotations as described by Equation (3).

The workspace (or the "reach") of the end-effector (Cα) is determined by the values of the dihedral angles. The same PDB data used to develop the 2D maps in Madden et al. (2008) was used to develop 3D maps for triads of chain segments, and Figure 8 shows a sample of such a 3D map. The darker regions in this 3D workspace indicate regions with higher propensity for the placement of (Cα) with respect to the nitrogen atom. Informally, this workspace is the 3D equivalent of the dihedral angle maps illustrated in Figure 5: for each such point, the corresponding dihedral angles can be found by solving an inverse kinematic problem. Observe that this workspace is a curved surface in the Cartesian space generated by revolving a circular arc around the x-axis as shown in Figure 8.

6. Triad Trajectories of the 2D Angle Space Pathways

The simplest way to connect two points is by a straight line, which is the basis of most of the existing visualization techniques for conformational transitions. We draw such a geodesic in the dihedral angle space as illustrated in Figure 9 to obtain the resulting Cα trajectory in Cartesian workspace. This trajectory corresponds to the shortest distance as measured in the dihedral angle space between the two end conformations.

In order to improve the path in the 2D dihedral angle maps, we select the best direction of this line based on the corresponding energy of the conformations. In other words, if the shortest distance forces the protein through a high-energy field, we let the protein move on the same geodesic, but in the opposite direction as long as this direction is more energy favorable. The corresponding trajectory for Cα in a sample triad segment is shown in Figure 10(a) and (c). Note that the directions of both trajectories are switched to the opposite



(a) (b)

Fig. 9. The trajectory of Cα of a triad segment corresponding to the shortest distance line in the angle space: (a) (φ_3, ψ_3); (b) (φ_4, ψ_4).

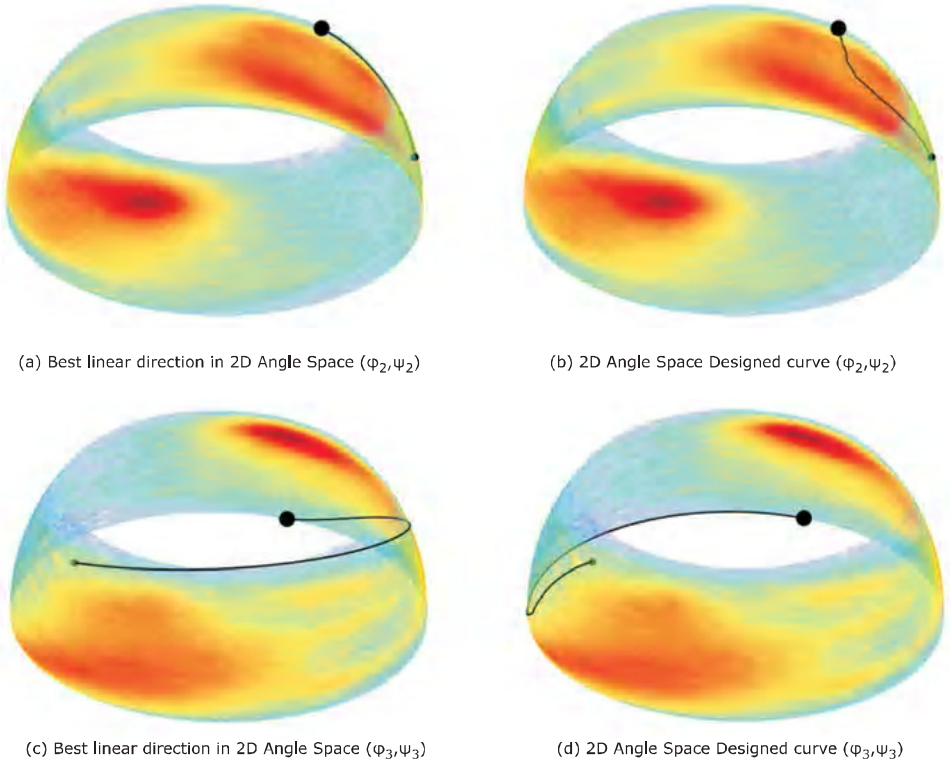


Fig. 10. The trajectory of Cα of a triad segment: (a) and (c) correspond to the best linear direction for (φ_2, ψ_2) and (φ_3, ψ_3); (b) and (d) correspond to the interactively designed path in the 2D angle space for (φ_2, ψ_2) and (φ_3, ψ_3).

direction as described above. To further improve the energy profile along the trajectory, the paths in the 2D maps were adjusted interactively so that they pass through higher propensity 2D regions. The corresponding trajectory in the Cartesian workspace is shown in Figure 10(b) and (d). Please note that these interactive adjustments were imposed in the dihedral angle spaces.

7. Triad Workspace Propensity Maps as Navigation Guides

Such sample pathways are a good starting point in the search for the best (minimum) energy conformational transition pathways. In this work, the user interactively examines each Cartesian workspace. The 20 amino acids that we modeled pro-

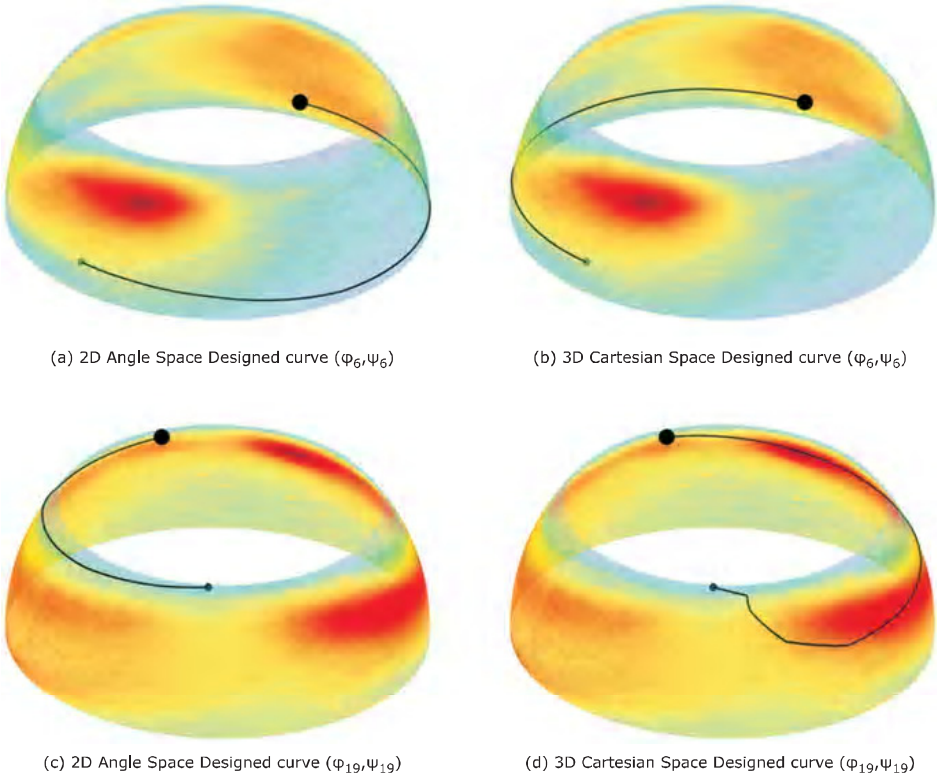


Fig. 11. The trajectory of Cα of a triad segment corresponding to designed trajectory in the 2D dihedral angle space (a) and (c) versus the designed paths in 3D Cartesian space (b) and (d).

duce 820 of these charts for every combination of amino acid residues in a peptide chain. For our numerical experiment that considers the transition from 1FOX to 2FOW, there are 147 3D Cartesian maps. The user examines each such map and modifies the path to pass through the highest propensity regions which are shown in darker shading. Figure 11 contains two sample "designed" paths, namely (b) and (d), which had been modified by the user from their original designed form ((a) and (c)). The original paths were obtained from the dihedral angle space maps. In order to transfer the adjustments made to the Cartesian Cα trajectories to corresponding changes in the angle space, one must solve a simple inverse kinematics problem for each triad as discussed in Appendix A.1. It should be noted again that further automation of the process is warranted.

8. Energy Minimization

The paths designed in the Cartesian space are used as initial conformations that are then used to find pathways corresponding to minimum energy conformations of the protein molecule. For the pathway planned in the dihedral angle space, energies are calculated for a set of intermediate conformations between the two known end conformations using the AMBER force field applied to the kinematic model. The successive kinetostatic fold compliance method in PROTOFOLD is used to move the kinematic model at each intermediate conformation to one of lower energy. This process is repeated for all intermediate conformations, under the restriction of minimal joint rotation to maintain continuity between adjacent intermediate conformations (that is, temporal and space coherence). The energy minimized pathway closely follows the planned input pathway in angle space, but is able to traverse a much lower energy profile.

This energy minimization is particularly effective for relieving local areas of intense atomic interaction that cause high energy, but are not detected through angle space planning alone. Applying this optimization for segments of the pathway that contain large jumps in energy is computationally effective because it is run for few and not all of the intermediate conformations.

9. Simulation Results

Figure 7 shows the energy reduction based on 2D angle space trajectory path planning discussed by Madden et al. (2008) for

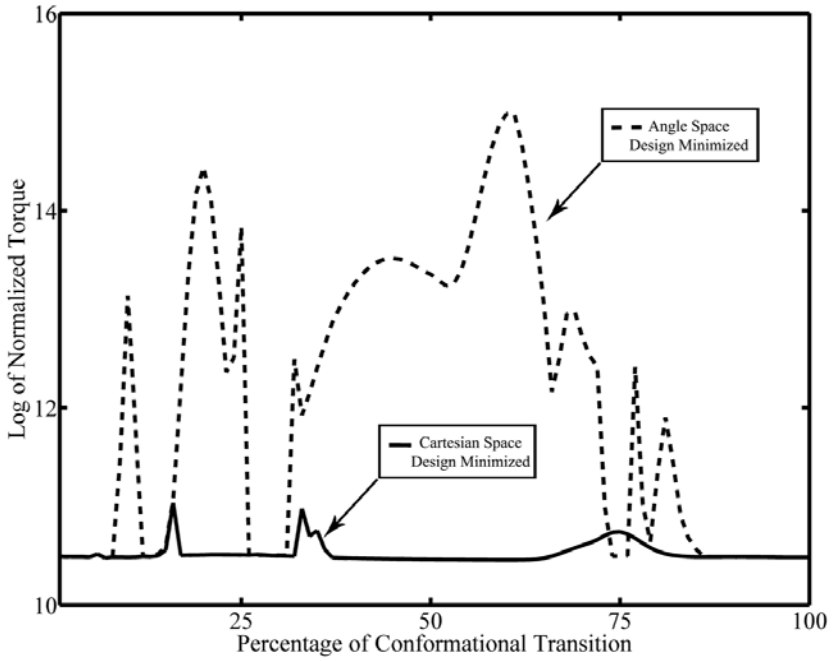


Fig. 12. Normalized joint torque for minimized angle space designed path versus the minimized Cartesian space designed path.

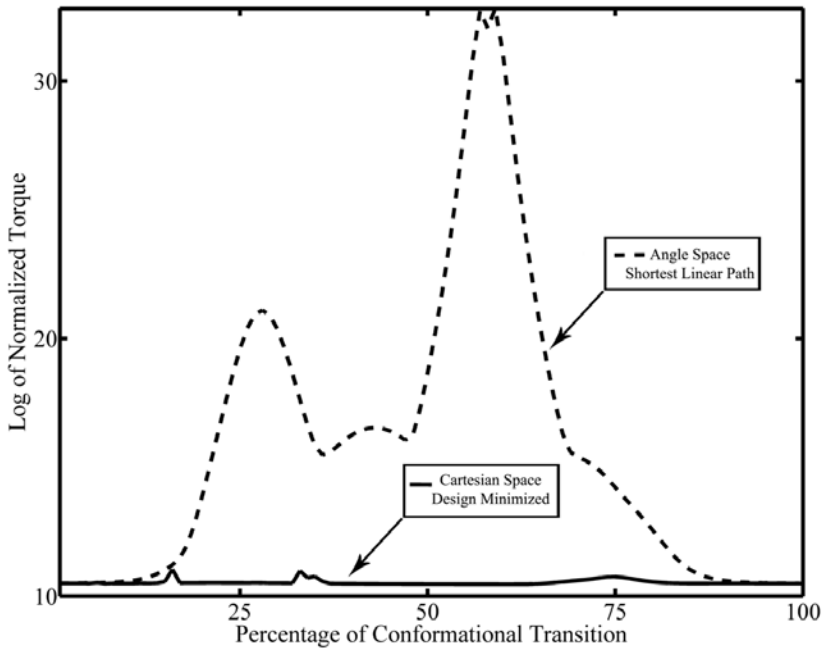


Fig. 13. Normalized joint torque for shortest linear path in angle space versus the minimized Cartesian space designed path.

the conformational transition from 1FOX to 2FOW. The pathway labeled "optimized designed" trajectory is the pathway designed interactively based on the 2D propensity maps followed by an energy optimization in PROTOFOLD.

The minimization of the pathway energy continued based on the 3D strategies discussed above. Figure 12 contains the energy profiles of two pathways: the trajectory shown with dotted line is the best energy path of Figure 7 (2D optimized

designed trajectory). In the same figure, the solid line pathway is the 3D optimized design pathway. The effectiveness of the 3D strategies proposed in this paper is apparent from Figure 12. Furthermore, Figure 13 shows the overall minimization achieved through the 2D and 3D strategies by comparing the linear pathway in the 2D angle space to the final optimized design path in 3D triad workspaces.

10. Conclusions

This paper extends the strategies developed by the authors to systematically develop pathways for the conformational transition of the functional protein molecules. The pathways are developed based on torsion angle propensity maps obtained from the data available on more than 38,000 protein chains in the PDB. The 2D dihedral angle propensity maps are useful navigation tools for interactive planning of energy favorable pathways. Clearly, there is a one-to-one mapping between the 2D dihedral angle maps and the 3D Cartesian maps proposed here, which means that any path planned in one set of these maps can be equivalently planned in the other set of propensity maps.

However, we argue that 3D propensity maps constructed in the Euclidean space for triads of chain segments (from nitrogen to α-carbon) are more efficient navigation tools than the propensity maps constructed in the angle space. This is so because the proposed 3D Cartesian maps naturally represent the propensity distributions in a space where the physical conformation changes takes place. Furthermore, these 3D maps allow the user to initiate the pathway planning directly in the 3D Euclidean space where the conformation of the protein molecules is actually changing, and where the actual inter-atomic collision/proximity takes place. Consequently, we show that the interactive path planning based on the proposed 3D maps can lead to lower state energies of the resulting pathway. Our numerical simulations indicate that these 3D workspaces provide a very effective strategy for charting minimum energy profile pathways.

This work is focused on developing and demonstrating the effectiveness of the 3D energy maps. For implementation as a production tool, automation of the process is warranted. Furthermore, we expect that automating the proposed planning strategies would lead to substantially more potent capabilities for exploring the energy landscape of proteins. This would allow a more efficient investigation of additional proteins with known stable conformations, which, in turn, would lead to a better understanding of the performance and efficiency of our method. In addition, the proposed 3D maps provide a superior alternative to the 2D dihedral angle maps in path planning methods, artificial potential techniques or modal analysis approaches reported in the literature.

A. Appendix

A.1. Inverse kinematics formulation

The derivation of the inverse kinematics for the amino acid triad is relatively straightforward. The glossary of the symbols used in these equations is:

- b_{ix} , b_{iy} and b_{iz} are the components of the zero-position body vector $\mathbf{b}_i$ for a given triad along the coordinate axes;
- u_{ix}, u_{iy} and u_{iz} are the components of the zero-position unit vectors $\mathbf{u}_i$ for a given triad along the coordinate axes;
- x, y and z are the Cartesian coordinates of the last α-carbon in the triad;
- φ and ψ are the dihedral angles corresponding to the first α-carbon in the triad.

To simplify expressions, the constant A is calculated using the zero-position vectors for the structure:

$$\begin{aligned} A &= (b_{(i+1)x} + b_{(i+2)z})u_{(i+1)x}^2 \\ &+ (u_{(i+1)y}b_{(i+1)y} + u_{(i+1)y}b_{(i+2)y})u_{(i+1)x}. \end{aligned} \tag{7}$$

The cosine of the dihedral angle ψ can be determined using the definition of A from Equation (7), along with the x component of each of the zero-position body vectors for the structure. The only variable in this equation is x, the Cartesian space X coordinate of the last α-carbon in the triad:

$$\cos(\psi) = \frac{A - x - b_{ix}}{A - b_{(i+1)x} - b_{(i+1)z}}, \tag{8}$$

which results in two values of the angle (and the corresponding multiples). The next section outlines a simple procedure to deal with this issue. After the value of the angle ψ has been determined, it is used with the following equations to obtain a value for the angle φ:

$$C_1(a,b) = \frac{u_{(i+1)x}u_{(i+1)y}a(1-\cos(\psi))+u_{(i+1)y}\sin(\psi)b}{a^2+b^2}, \tag{9}$$

$$C_2(a,b) = \frac{u_{(i+1)y}^2a(1-\cos(\psi))+a\cos(\psi)-u_{(i+1)x}\sin(\psi)b}{a^2+b^2}, \tag{10}$$

$$\begin{aligned} \sin(\varphi) &= C_1(z,y)(b_{(i+1)x} + b_{(i+2)x}) \\ &+ C_2(z,y)(b_{(i+1)y} + b_{(i+2)y}), \end{aligned} \tag{11}$$

$$\begin{aligned} \cos(\varphi) &= C_1(y,-z)(b_{(i+1)x} + b_{(i+2)x}) \\ &+ C_2(y,-z)(b_{(i+1)y} + b_{(i+2)y}), \end{aligned} \tag{12}$$

$$\varphi = \arctan 2(\sin(\varphi), \cos(\varphi)). \tag{13}$$

A.2. Procedure for ψ Selection

As addressed in the previous section there are two possibilities for the dihedral angle ψ:

$$\psi = \pm \arccos(\cos(\psi)). \tag{14}$$

Selecting either option will result in a corresponding value for φ, obtained using the formulation in the previous section. The formulation indicates that for any point in the triad's workspace there are two unique angle combinations that will move the tip of the triad to that point. However, this description does not hold true for instances where ψ is equal to $n\pi$, where n is any integer number. In these cases there is only one possible solution for ψ.

The inverse kinematics formulation used here is implemented in the design interface for the Cartesian space α-carbon trajectories. It is used to obtain updated values for the dihedral angles of each triad as their trajectories are manipulated. In order to select appropriate values for ψ, it is assumed that any manipulation of the trajectories is to be of small magnitude (less than 1 Ångström), which simplifies the ψ selection process. It is not very computationally demanding to check both possibilities for ψ. Therefore, after any Cartesian space trajectory manipulation (for all cases where $\cos(\psi)$ is not equal to 0 or -1), (φ, ψ) pairs are obtained using both values in Equation (14)). These (φ, ψ) pairs are compared with the previous (φ, ψ) pair for that triad and the new pair that is closest to the original is selected.

References

Amato, N. M. and Song, G. (2002). Using motion planning to study protein folding pathways. *Journal of Computational Biology*, **9**(2): 149–168.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. and Zardecki, C. (2002). The Protein Data Bank. *Acta Crystallographa D Biological Crystallography*, **58**(Part 6 No 1): 899–907.

Bohnenkamp, P., Kazerounian, K. and Ilieş, H. (2007). Strategies to avoid energy barriers in *ab initio* protein folding. *Proceedings of the 12th IFToMM (International Federation of the Theory of Mechasims and Machines) World Congress*, Besan con, France.

Chirikjian, G. S., Kazerounian, K. and Mavroidis, C. (2005). Analysis and design of protein based nanodevices: challenges and opportunities in mechanical design. *Journal of Mechanical Design*, **127**(4): 695–698.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. and Kollman, P. A. (1996). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *Journal of the American Chemical Society*, **118**(9): 2309–2309.

Cortes, J., Simeon, T., de Angulo, V. R., Guieysse, A. D., Remaud-Simeon, M. and Tran, V. (2005). A path planning approach for computing large-amplitude motions of flexible molecules. *Bioinformatics*, **21**: I116–I125.

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G. M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. M. and Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, **24**(16): 1999–2012.

Geeves, M. A. and Holmes, K. C. (1999). Structural mechanism of muscle contraction. *Annual Review of Biochemistry*, **68**: 687–728.

Gerstein, M. and Krebs, W. (1998). A database of macromolecular motions. *Nucleic Acids Research*, **26**(18): 4280–4290.

Gibbs, N., Clarke, A. R. and Sessions, R. B. (2001). Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. *Proteins—Structure Function and Genetics*, **43**(2): 186–202.

Guilbert, C., Pecorari, F., Perahia, D. and Mouawad, L. (1996). Low frequency motions in phosphoglycerate kinase. A normal mode analysis. *Chemical Physics*, **204**(2–3): 327–336.

Guilbert, C., Perahia, D. and Mouawad, L. (1995). A method to explore transition paths in macromolecules—applications to hemoglobin and phosphoglycerate kinase. *Computer Physics Communications*, **91**(1–3): 263–273.

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

Itoh, K. and Sasai, M. (2004). Coupling of functioning and folding: photoactive yellow protein as an example system. *Chemical Physics*, **307**(2–3): 121–127.

Kazerounian, K., Latif, K. and Alvarado, C. (2005a). Protofold: a successive kinetostatic compliance method for protein conformation prediction. *Journal of Mechanical Design*, **127**(4): 712–717.

Kazerounian, K., Latif, K., Rodriguez, K. and Alvarado, C. (2005b). Nano-kinematics for analysis of protein molecules. *Journal of Mechanical Design*, **127**(4): 699–711.

Kern, D. and Zuiderweg, E. R. P. (2003). The role of dynamics in allosteric regulation. *Current Opinion in Structural Biology*, **13**(6): 748–757.

Kim, M. K., Jernigan, R. L. and Chirikjian, G. S. (2003). An elastic network model of HK97 capsid maturation. *Journal of Structural Biology*, **143**(2): 107–117.

Krebs, W. G., Tsai, J., Alexandrov, V., Junker, J., Jansen, R. and Gerstein, M. (2003). Tools and databases to analyze protein flexibility; approaches to mapping implied features onto sequences. *Macromolecular Crystallography D*, **374**: 544.

Liu, J., Zhang, J., Yang, Y., Huang, H., Shen, W., Hu, Q., Wang, X., Wu, J. and Shi, Y. (2008). Conformational change upon ligand binding and dynamics of the PDZ domain from leukemia-associated Rho guanine nucleotide exchange factor. *Protein Science*, 073416508.

Madden, C., Bohnenkamp, P., Kazerounian, K. and Ilieş, H. T. (2008). Predicting protein conformational transitions by trajectory planning through torsion angle propensity maps. *Interdisciplinary Applications of Kinematics*, Kecskeméthy, A., (ed.). Amsterdam, Elsevier.

Miyashita, O., Wolynes, P. G. and Onuchic, J. N. (2005). Simple energy landscape model for the kinetics of functional transitions in proteins. *Journal of Physical Chemistry B*, **109**(5): 1959–1969.

Pande, V. S. (2005). Folding@home: using desktop grid computing to overcome fundamental barriers in biomolecular simulation. *Abstracts of Papers of the American Chemical Society*, **230**: U1295–U1295.

Ramachandran, G., Ramakrishnan, C. and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, **7**: 95–99.

Sayle, R. and Milner-White, E. J. (1995). Rasmol: biomolecular graphics for all. *Trends in Biochemical Sciences*, **20**(9): 374–376.

Scheraga, H. A., Liwo, A., Oldziej, S., Czaplewski, C., Pillardy, J., Ripoll, D. R., Vila, J. A., Kazmierkiewicz, R., Saunders, J. A., Arnautova, Y. A., Jagielska, A., Chinchio, M., and Nanias, M. (2004). The protein folding problem: global optimization of force fields. *Frontiers in Bioscience*, **9**: 3296–3323.

Schlitter, J., Engels, M. and Kruger, P. (1994). Targeted molecular-dynamics—a new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics*, **12**(2): 84–89.

Schuyler, A. D. and Chirikjian, G. S. (2005). Efficient determination of low-frequency normal modes of large protein structures by cluster-NMA. *Journal of Molecular Graphics and Modelling*, **24**(1): 46–58.

Su, H.-J., Parker, J., Kazerounian, K. and Ilieş, H. (2007). A comparison of kinetostatic and multibody dynamics models for simulating protein structures. *Proceedings of the ASME IDETC Mechanisms and Robotics Conference*, Las Vegas, NV, September 2007.

Subramanian, R. and Kazerounian, K. (2006). Improved molecular model of a peptide unit for proteins. *Proceedings of the ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, DETC 2006-99315, Philadelphia, PA.

Subramanian, R. and Kazerounian, K. (2007a). Kinematic mobility analysis of peptide based nano-linkages. *Mechanism and Machine Theory*, **42**(8): 903–918.

Subramanian, R. and Kazerounian, K. (2007b). Residue level inverse kinematics of peptide chains in the presence of observation inaccuracies and bond length changes. *Journal of Mechanical Design*, **129**(3): 312–319.

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

Tama, F., Wriggers, W. and Brooks, C. L. (2002). Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *Journal of Molecular Biology*, **321**(2): 297–305.

Tapia, L., Tang, X. Y., Thomas, S. and Amato, N. M. (2007). Kinetics analysis methods for approximate folding landscapes. *Bioinformatics*, **23**(13): I539–I548.

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