Protein Folding Using Coarse-Grained Optimal Control and Molecular Dynamics

Yaman Arkun*. Mert Gür**

*Department of Chemical and Biological Engineering, Koc University, Istanbul, Turkey (Tel: 90-212-338-1313; e-mail: yarkun@ku.edu.tr).

**Center for Computational Biology and Bioinformatics, Koc University, Istanbul 34450 Turkey (e-mail: mgur@ku.edu.tr).

Abstract: Protein folding is formulated as a moving horizon optimal control problem. We present a novel method that combines coarse-grained optimal control and molecular dynamics to generate dynamical folding trajectories with full atomic details. The method is applied to a benchmark protein, Villin headpiece.

1. INTRODUCTION

After their synthesis in the cell, proteins fold into their unique native states in order to fulfill their biological functions in human body. Incorrectly folded proteins are known to be the causes of several diseases such as Creutzfeldt-Jakob, Parkinson’s and Alzheimer’s diseases. For this reason the problem of protein folding is an important and active research area (Pappu et al, 2009).

Proteins are chains of amino acids (residues) that are connected with peptide bonds. The particular sequence of the amino acids (i.e. the primary protein structure) determines the protein’s folding characteristics. Protein folding can be characterized as a dynamical process during which the protein starts from a random unfolded or denatured initial configuration and folds into its unique steady or native state as depicted in Fig.1. Significant amount of research has been devoted to the determination of the alternative folding routes (i.e. those that bridge the initial and native configurations shown in Fig. 1.) and characterization of the folding landscapes. It is now recognized that, as it folds to its final native state, the protein minimizes its energy while avoiding high-entropy loss routes (Dill and Chang, 1997; Arkun and Erman, 2010).

Protein folding is accomplished by the intermolecular forces among the amino acid groups. Determination of intermolecular forces and the resulting folding trajectories over large time scales using molecular dynamics (MD) simulations becomes computationally very demanding as the number of residues increases. Therefore one usually resorts to less detailed or coarse-grained (CG) models which facilitate the computations. When coarse-grained models are combined with more refined atomistic models, useful insight into the problem of protein folding can be obtained (Clementi, 2007).

Figure 1. The protein’s three dimensional configuration before and after folding.

The objective of this paper is: given an initial and final configuration (like in Fig.1), determine the optimal trajectory that connects them. Specifically we are interested in developing a method that can easily compute optimal folding trajectories and at the same time closely represent the real protein.

2. METHOD

Our proposed method combines the best of two worlds of Coarse-Grained (CG) modeling and Molecular Dynamics (MD). Performing dynamic optimization using a CG model provides simplicity and speed whereas MD supplements accuracy. We formulate a coarse-grained model based dynamic optimization to compute the folding trajectories. This is coupled with MD which fills in the atomistic details and refines the folding trajectory.
2.1 Protein Model

In our work we use a coarse-grained model where each amino acid group is lumped to and represented by its alpha carbon. Each of these alpha carbons is represented with a spherical bead in space and the position of the \( i \)th bead is denoted by the vector \( r_i \) with respect to a fixed reference frame. Beads are connected with each other with springs. Beads-and-springs representation of the protein is common in the literature (e.g. see Erman and Dill, 2000). The dynamic model is governed by Newton’s equation of motion, and for a system with \( N \) beads its state space representation is given by (Güner et al., 2006):

\[
\ddot{r}_{CG} = Ar_{CG} + u_{CG} \tag{1}
\]

where

\[
r_{CG} = [r_1, r_2, ..., r_N] \in \mathbb{R}^{3N}
\tag{2}
\]

where CG stands for Coarse-Grained; matrix \( A \) is the known connectivity matrix of amino acids; and \( u_{CG} \in \mathbb{R}^{3N} \) is the vector of individual forces acting on each bead. It is these intermolecular attractive and repulsive forces that cooperatively fold the protein.

2.2 Optimization

In CG dynamic optimization the following LQR (Linear Quadratic Regulator) problem is solved subject to (1).

\[
\min_{\vec{u}_{CG}} \int_{0}^{\infty} \left[ \mathbf{r}_{CG}^T \mathbf{Q}_{CG} \mathbf{r}_{CG} + \rho \mathbf{u}_{CG}^T \mathbf{R} \mathbf{u}_{CG} \right] dt \tag{3}
\]

Both the initial condition \( r_{CG}(0) \) (i.e. initial alpha carbon positions) and final steady states \( r_{CG}^{\text{native}} \) and \( u_{CG}^{\text{native}} \) are known. Symbol “-” denotes the deviation from the known native state.

The objective function in (3) is directly derived from the thermodynamics of the problem (Arkun and Erman, 2010). The first term under the integral is the harmonic coarse-grained approximation of energy which we want to minimize; and the second term penalizes excess entropy losses. We set \( R=I \) without loss of generality, and \( \rho \) is a tuning parameter that is used to trade-off energy vs. entropy.

\( Q \) is a positive definite matrix set by the topology of the protein (Arkun and Erman, 2010). Optimization (3) computes the optimal force field \( u^{*}_{CG}(t) \) and the optimal coarse-grained folding trajectory \( r^{*}_{CG}(t) \) that connects the initial CG conformation \( r_{CG}(0) \) with the final state \( r_{CG}^{\text{native}} \).

The CG model (1) and objective function in (3) represent a simplified version of the real world. As such the resulting CG trajectory \( r^{*}_{CG}(t) \) is suboptimal for the real protein represented through MD which includes all the atomistic details (in addition to alpha carbons). Therefore the optimal CG trajectory obtained from (3) needs to be refined by making the necessary corrections to the alpha carbon positions based on the information acquired from MD. This multi-graining is achieved by a concerted use of CG dynamic optimization and MD through the whole folding process as shown in Figures 2 and 3 and described below.

In Fig. 2 in MD \( r_{AA}(0) \) stands for the initial position vector including all the atoms. This is available to us. CG optimal control-MD update algorithm starts by picking a CG conformation from the coarse-grained optimal trajectory \( r^{*}_{CG}(t) \) at a particular time \( t = t_k \). This sampled conformation \( r^{*}_{CG}(t_k) \) is accepted as the first “target” structure for MD and is supplied to MD. Next atomic details are added to this structure by MD which performs an all-atom simulation using its built-in force field \( F(t) \) (see the dynamic model inside MD box in Fig. 2). First MD brings the alpha carbon positions of the initial all-atom conformation \( r^{*}_{AA}(0) \) to the optimal target structure \( r^{*}_{CG}(t_k) \) by performing targeted molecular dynamics (TMD) simulation (Ferrera et al., 2000). This is followed by equilibration of potential and kinetic energies. The resulting structure is an all-atom detailed structure closest to the targeted CG structure. It is denoted by \( r_{AA}(t_k) \) with its corresponding alpha carbon positions \( r_{CG}(t_k) \). This information is fed back by MD to CG optimization.

In order to construct the next conformation on the remaining “optimal” trajectory from \( t_k \) to \( t_f \), time is advanced to \( t_k \); alpha carbon positions in the CG model are set equal to the corrected ones supplied by MD i.e. \( r_{CG}(t_k) \) and CG optimization is repeated starting at time \( t_k \). This sequence of optimal trajectory calculations and their update via MD is summarized in Fig. 3.

The above method implements a moving horizon optimal controller which adapts to the real world. In each time
horizon an “open loop” CG model-based optimization is performed. The resulting trajectory’s states (i.e. alpha carbon positions) are corrected based on the feedback information from MD and optimization is repeated at the next sampling time. This cycle of dynamic optimization and feedback correction is repeated until the end of folding. At the end one obtains an optimal folding trajectory that consists of $N$ conformations with full atomic details:

$$\{r_{AA}(0), r_{AA}(t_1), r_{AA}(t_{k+1}), \ldots, r_{AA}(t_{k+N-2})\}.$$ We should note that after each MD simulation, in addition to the state, weighting matrix $Q$ in the objective function (3) is also updated. By definition the entries of $Q$ are determined by the amino acids (alpha carbons) that make contact (i.e. in proximity of 7 Å). Since these contacts change as the protein folds, $Q$ needs updating before each optimal control calculation. Therefore, after each MD, new contacts are recomputed and $Q$ is recomputed and updated.

### 3. RESULTS

Villin headpiece, Protein Data bank code 1VII.pdb, was selected as an example. Villin has 36 amino acids and it is the smallest protein that folds autonomously. All molecular dynamics simulations were performed using NAMD 2.7b package (Phillips et al., 2005) with CHARMM27 force field. The equilibrated unfolded Villin structure was chosen as the initial conformation. Using its alpha carbon positions as starting coordinates, the first coarse grained optimal folding trajectory was constructed. 50 coarse grained structures along this folding trajectory were sampled. The first sampled coarse grained structure was selected as the target structure for MD. A 0.01ns long TMD simulation and a subsequent 0.05ns long equilibration simulation were performed. The final structure was recorded as the first all-atom structure of the folding trajectory. Coarse grained optimization is repeated starting with the alpha carbon positions of this new structure and optimization-MD update is repeated until the protein folds successfully to its final native state.

**Figure 2. Computation of the folding trajectory using CG Dynamic Optimization and MD.**

At $t = 0$:

$$r_{AA}(0) \xrightarrow{\text{CG Optimization}} r_{CG}(t_1) \xrightarrow{\text{MD}} r_{AA}(t_1)$$

Advance the time to $t_1$ and repeat:

$$r_{AA}(t_1) \xrightarrow{\text{CG Optimization}} r_{CG}(t_{k+1}) \xrightarrow{\text{MD}} r_{AA}(t_{k+1})$$

**Figure 3. Calculation sequence**

Figure 4 is a plot of RMSD (RootMeanSquareDeviation) values of alpha carbon positions from the native state during folding. Monotonic decrease of RMSD with time indicates that initial unfolded configuration folds to its native state as time progresses.

Figure 5 illustrates how the energy of the protein (as evaluated by NAMD) behaves along its folding trajectory. It is well-known that the proteins’ energy landscapes are usually rugged (Dill and Chang, 1997). Several energy barriers and local minima in Fig. 4 confirm this. However the protein is able to approach its minimum-energy native state. The energy difference between the initial and final structures is -351.48 kcal/mol.
6. CONCLUSIONS
This paper presents a new method that combines coarse grained optimal control and atomic scale molecular dynamics simulation to construct optimal protein folding trajectories. While coarse graining facilitates the optimal trajectory calculation, MD supplements accuracy. The method is implemented as a moving horizon optimal control algorithm with MD learning.

REFERENCES


