

Folding pathways explored with artificial potential functions

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2009 Phys. Biol. 6 036008

(<http://iopscience.iop.org/1478-3975/6/3/036008>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 193.140.201.95

This content was downloaded on 10/11/2014 at 22:30

Please note that [terms and conditions apply](#).

Folding pathways explored with artificial potential functions

B Ulutaş^{1,3}, T Haliloglu² and I Bozma¹

¹ Intelligent Systems Laboratory, Electric Electronic Engineering Department, Bogaziçi University, 34342 Istanbul, Turkey

² Polymer Research Center, Department of Chemical Engineering, Bogaziçi University, 34342 Istanbul, Turkey

E-mail: halilogt@boun.edu.tr and bozma@boun.edu.tr

Received 6 February 2009

Accepted for publication 14 April 2009

Published 14 May 2009

Online at stacks.iop.org/PhysBio/6/036008

Abstract

This paper considers the generation of trajectories to a given protein conformation and presents a novel approach based on artificial potential functions—originally proposed for multi-robot navigation. The artificial potential function corresponds to a simplified energy model, but with the novelty that—motivated by work on robotic navigation—a nonlinear compositional scheme of constructing the energy model is adapted instead of an additive formulation. The artificial potential naturally gives rise to a dynamic system for the protein structure that ensures collision-free motion to an equilibrium point. In cases where the equilibrium point is the native conformation, the motion trajectory corresponds to the folding pathway. This framework is used to investigate folding in a variety of protein structures, and the results are compared with those of other approaches including experimental studies.

1. Introduction

This paper addresses conformational changes related to the protein-folding pathways which are believed to reflect the nature of folding [1]. The aim is to understand how a protein folds to a given native structure. A simplistic formulation of the folding problem under known native topology can be stated as follows: a three-dimensional workspace contains an ordered set of bodies (amino acids), where each is virtually connected to the previous and next bodies in the sequence—if they exist. Each body has a goal location and the ensemble of these locations specifies the native conformation. Once folding starts, the bodies move from their initial conformation to their respective goal positions in this native conformation while adhering to simple physical principles such as mutual avoidance. We propose a novel approach for modeling the folding mechanism that is inspired by work on the coordinated navigation of multiple robots that promises to provide simultaneously computationally

simple and yet comparably reliable depiction of the folding process.

1.1. Related literature

Experimentally determined structures of different proteins have become available over the last several years through studies using x-ray crystallography, nuclear magnetic resonance spectroscopy and cryo-electron microscopy. However, despite recent advances, direct experimental studies are still limited in elucidating these mechanisms. Moreover, there is no consensus as regards to whether experimentation alone can be sufficient in obtaining a full mechanistic description. The development of theoretical and computational models of conformational changes has then impacted the nature of protein-folding research [14]. These models, based on the principle of minimal frustration, use, in general, the energy landscape theory that associates the protein-folding problem with a globally funneled energy surface whose global minima correspond to the native state. It is observed that the proposed approaches are governed by two design criteria:

³ Present address: University of Victoria, Victoria, British Columbia, W8W 3P22, Canada.

- the construction of an associated energy landscape. The choice is between complex but precise models and simplified but approximate models.
- the strategy employed to generate the folding pathways. The choice is between dynamical simulation and sampling-based approaches.

1.2. Complex energy models

Complex energy models treat the protein chain at an atomistic level and are observed to be used with two alternative motion strategies. The traditional approaches such as molecular dynamics simulations use true but extremely complex dynamical models based on the solution of Newtonian or Langevin's equations to simulate the folding process⁴ [14, 16, 28]. Although the generated paths are extremely reliable, they are computationally very expensive. Moreover, it is extremely difficult to utilize them in large structures. *Ab initio* structure prediction approaches also use relatively complex energy models, but their motion strategies employ statistical techniques such as sampling [24]. Although the computational complexity of these methods is lower, the generated folding pathways may not be necessarily realistic.

1.3. Simplified energy models

Simplified energy models are based on reduced models of proteins, which can be categorized as follows: general models of protein-like polymers, knowledge-based models, physics-based models and models biased toward the native structure [34]. In general, the desired locations are modeled by attractive potentials, and obstructions are modeled by repulsive potentials. These two terms are then additively combined to generate the equations of motion for the protein. In most constructions, simple 'bowl-like' gradients are added to Newtonian inverse square type that become unbounded as one approaches the boundary of a forbidden region. Even with a minimalist, native structure-based energy function, it has proved difficult to capture many folding and unfolding events at equilibrium conditions using standard dynamics [12]. Such constructions have then been used in conjunction with one of the two alternative motion strategies in order to generate the folding pathways. One approach is to approximate the true dynamics via using simpler energy models as is done in statistical mechanical approaches [2, 9]. These have been successful in predicting folding kinetics of smaller proteins [29]. Monte Carlo methods also use these simpler energy models where the energy function is formulated by additively combining attractive and repulsive terms [13]. However, instead of adopting a dynamical model, the folding process is simulated through statistical sampling techniques. Recently, observing the similarity between the protein folding and robotic path planning problems, the probabilistic roadmap (PRM) approach has been applied [4, 5, 6, 35, 36]. Here, folding sequences are automatically produced by generating a roadmap and moving on it—both

of which are accomplished probabilistically based on the energy landscape. While all the sampling-based techniques including the PRM models have proved to be valuable for investigation of folding pathways, their path quality is sampling dependent. Moreover, as the dimension of conformation space increases the required number of samples (for protecting the same path quality) increases dramatically—an issue known as Levinthal's paradox [25]. Furthermore, for none of these constructions can it be guaranteed that forbidden regions corresponding to the physically infeasible case of overlapping bodies will be avoided (except for point-by-point checking) much less than the desired goal point will be achieved. Hence, additional logical or numerical compositional operators requiring many new parameters need to be introduced in order to sieve out infeasible conformations or weaken the nominal strength of the unbounded repelling fields [32, 20], which obviously further complicate the algorithms and affect their robustness.

1.4. Contributions

The contribution of this paper is to present an approach based also on simplified energy models and dynamics, but with the novelty that—motivated by work on robotic navigation [10, 19–23]—a nonlinear compositional scheme of constructing the energy model is adapted instead of an additive formulation. The construction of the energy function is slightly different compared to previously presented models—as is defined on the configuration space of torsion (dihedral) angles associated with a protein chain. The advantages can be summarized as follows:

- Here, in contrast to previous work, the avoidance of colliding conformations is guaranteed as is also demonstrated by proposition 2.1.
- Furthermore, as the approach is based on dynamics as opposed to sampling, there is no exponential increase in the required computational resources as the size of the protein and hence the complexity of the folding process increase. The convergence time is simply determined by the topology of the underlying surface—similar to any dynamics-based approach.

Let it be noted that in all protein-folding approaches based on simplified energy models and dynamics, no guarantees of convergence hold as is the case here. Here, a mathematical issue of concern is the existence of cycles that may lead to oscillatory type of behavior. This question can be answered definitively; only if the energy function is shown to have certain properties (i.e. being a navigation function, etc). However, such an analysis has proved to be very difficult in general, and guarantees ensuring that the moving robot will avoid obstacles and arrive at the desired destination have been possible only for very simple robot navigation problems [10, 19, 21]. Hence, prior examples of correct algorithms along with our extensive simulations strongly motivate that in future, the theoretical results can be extended to the current problem.

⁴ Let it be noted that these methods are based on first principles and hence can potentially address folding problems with unknown native conformations.

2. Approach

2.1. Ordered constrained bodies

Consider an ordered set $P = \{1, \dots, p\}$ of p bodies where each body, $i \in P$, has radius $\rho_i \in R$. Let $Q = \{(m, j) | m, j \in P, m < j\}$ denote the set of body pairs in P with cardinality $|Q| = p(p-1)/2$. Assume that the Euclidean distance between any consecutive pair $(m, m+1)$ is fixed with the length l . The whole set can then be defined by $p-2$ links, where the first link is defined by $\theta_i \in S^1$ and the remaining links are defined by $w_i = [\theta_i, \phi_i]^T \in S^2$ where the superscript T indicates the transpose operator. For ease of notation, let $\theta = [\theta_1, \dots, \theta_{p-2}]^T$ and $\phi = [\phi_2, \dots, \phi_{p-2}]^T$. Hence, the aggregate conformation space is defined as $W = S^1 \times \underbrace{S^2 \times \dots \times S^2}_{p-3}$. Let $w_g \in W$ denote the native configuration

associated with this ordered set. Given a conformation w , the corresponding relative Cartesian coordinates $b_i(w) \in R^3$ for each body, $i \in P, i \neq 1$, can be computed via a series of coordinate transformations as

$$b_i(w) = \sum_{k=2}^i \left(\prod_{j=1}^{k-2} T_j I_k \right). \quad (1)$$

Each T_i denotes the transformation matrix for a frame centered on the body of the index $i-1$, such that its x -axis is along the direction of the i th body and y -axis is defined such that if $\frac{1}{2}\pi < \phi_{i-1} < \frac{3}{2}\pi$ then the $(i+1)$ th body is in the positive y -axis. For $i=1$, T_1 is as follows:

$$T_1 = \begin{bmatrix} -\cos(\theta_1) & \sin(\theta_1) & 0 \\ \sin(\theta_1) & \cos(\theta_1) & 0 \\ 0 & 0 & -1 \end{bmatrix}. \quad (2)$$

Here, as there is no ϕ_1 angle, we define the y -axis such that the second body is in the positive y direction. For $i > 1$, T_i is defined as follows:

$$T_i = \begin{bmatrix} -\cos(\theta_i) & \sin(\theta_i) & 0 \\ -\sin(\theta_i) \cos(\phi_i) & -\cos(\theta_i) \cos(\phi_i) & -\sin(\phi_i) \\ -\sin(\theta_i) \sin(\phi_i) & -\cos(\theta_i) \sin(\phi_i) & \cos(\phi_i) \end{bmatrix}. \quad (3)$$

I_i denotes the position of the i th-ordered body with respect to this frame as

$$I_i^T = [l \quad 0 \quad 0]. \quad (4)$$

The relative Cartesian coordinates of the native conformation w_g are also computed similarly. Let it be remarked that given a conformation w , the fixed distance constraint between consecutive bodies can be expressed as

$$\delta_{m(m+1)}(w) = l, \quad \forall (m, m+1) \in Q, \quad (5)$$

where $\delta_{mn}(w) = \|b_m(w) - b_n(w)\|$ denotes the Euclidean distance between the pair $(m, n) \in Q$. Finally, let $G \subset W$ denote the free conformation space as

$$G = \{w \in W | \forall (m, n) \in Q, \delta_{mn}(w) \geq \rho_{mn}, \delta_{m,m+1}(w) = l\},$$

where $\rho_{mj} = \rho_m + \rho_j$. It should be observed that free conformation space G contains conformations physically feasible with respect to excluded volume while satisfying the fixed consecutive body distance constraint.

2.2. Artificial potential functions

The whole system is associated with a single artificial potential function, $\varphi : G \rightarrow [0, 1]$, that aims to achieve the collision-free navigation of the bodies to the goal conformation. It encodes both the intrinsic geometry of the goal and collision-free conformations. Its construction follows a methodology previously presented in [10, 19–23]:

$$\varphi(w) = \sigma_d \circ \sigma \circ \check{\varphi}(w). \quad (6)$$

The function $\check{\varphi} : G \rightarrow [0, \infty)$ is constructed as the ratio of two functions $\gamma_T : G \rightarrow [0, \infty)$ and $\beta : G \rightarrow [0, \infty)$ as

$$\check{\varphi}(w) = \frac{1000\gamma_T^k(w)}{\beta(w)} \quad k \in Z^+. \quad (7)$$

The numerator encodes the distance from the goal in terms of torsion angles as

$$\gamma_T(w) = \|w - w_g\|^2. \quad (8)$$

The denominator encodes the distance from the free-space boundary by using Cartesian coordinates as

$$\beta(w) = \prod_{(m,n) \in Q} \beta_{mn}(w), \quad (9)$$

where $\forall (m, n) \in Q, \beta_{mn}(w) = \delta_{mn}(w)^2 - \rho_{mn}^2$ and $\rho_{mn} = \rho_m + \rho_n$. Following a construction as presented earlier, the intrinsic potential function can be made admissible and has a non-degenerate critical point at the goal.

2.3. Mutual avoidance: additive versus multiplicative formulation

In the proposed approach, the denominator encodes the distance from the freespace boundary as a multiplicative collision function as

$$\beta(w) = \prod_{(m,n) \in Q} \beta_{mn}(w), \quad (10)$$

where it should be recalled that β_{mn} are non-negative-valued functions that encode the distance between the respective pairs $(m, n), \forall (m, n) \in Q$. Let $Z_M = \{w | \beta(w) = 0\}$ be the zeros of this function. Hence, the function is maximal on Z_M .

An alternative approach, which is used more generally, is to measure this distance using an additive collision function:

$$\beta_A(w) = \sum_{(m,n) \in Q} \kappa_{mn} \beta_{mn}(w), \quad (11)$$

where $\forall (m, n) \in Q, \kappa_{mn} > 0$. Now, let $S_A = \{w | \beta_A(w) = 0\}$ be the zeros of this function.

The two constructions can be compared by relatively considering the sets Z_M and S_A . For consistency, let us assume that the form of the potential function is the same—namely, the ratio of an attracting function and a collision function. The following proposition suggests that, in general, $Z_A \subseteq Z_M$ which means that the zeros of the additive formulation in general will contain configurations with interresidue collisions.

Proposition 2.1. $Z_A \subseteq Z_M$.

Proof. First note that if $w \in Z_A$ iff $\forall(m, n) \in Q, \beta_{mn}(w) = 0$. Hence $\beta(w) = 0 \rightarrow w \in Z_M$. Conversely, pick $w \in Z_M$ where $\exists(m, n), (k, l) \in Q$ such that $\beta_{mn}(w) = 0$ and $\beta_{lk}(w) > 0$. Hence, $\beta_A(w) > 0 \rightarrow w \notin Z_A$. \square

This proposition suggests that an additive formulation can miss configurations with interresidue collisions. Hence, a simple strategy such as evaluating the function suffices in the multiplicative formulation while in the additive formulation, this will not suffice. Each term of the formulation must be checked separately for the zero value. Furthermore, coupled with a gradient-based dynamical system, the progression of the folding trajectory is ensured to slide away from these maximal points.

2.4. Folding dynamics

The folding dynamics is governed by the following gradient system:

$$\dot{w} = -D_w \varphi(w), \quad (12)$$

with the initial condition as determined by the initial conformation w^0 . Let us note that for simplification purposes, a first-order model is adapted here. However, the formulation can easily be extended to higher-order models. Note that in this manner, the protein-folding dynamics considers the contribution of each robot's potential function in driving the overall conformation geometry. This dynamic system inherits the critical qualitative behavior of gradient trajectories. Suppose the integral curve of \dot{w} through the initial condition w^0 is denoted by w^t . If $-D_w \varphi(w) = 0$ implies full rankness, then the limit set, $\lim_{t \rightarrow \infty} w^t$, is some isolated singularity. In particular, the goal w_g lies in the limit set.

2.5. Relation between constrained bodies and proteins

The multibody framework is easily mapped to the protein structure as follows:

- The C_α components of proteins correspond to the spherical bodies. Hence, a protein structure with p number of residues is represented by the torsion angles $\theta \in S^{p-2}$ and $\phi \in S^{p-3}$.
- The protein structure is encoded by the fixed Euclidean distance between any two consecutive bodies.
- The spherical coordinate system allows the analysis of folding dynamics regardless of the absolute location of the protein structure in Cartesian workspace.

In this framework, given a protein with p residues, the native conformation $w_g \in G$ can be calculated via first extracting the Cartesian coordinates b_i of the amino acids $i \in P$ of its native conformation from the Protein Data Bank (PDB) [37]. These are then transformed to the spherical coordinate system as follows. Each θ_{g_i} angle is computed as

$$\theta_{g_i} = \pi - \arccos(v_i^T v_{i+1}), \quad (13)$$

where the vector $v_i \in R^3$ is defined as $v_i = b_i(w_g) - b_{i-1}(w_g)$. Each ϕ_{g_i} is as follows:

$$\phi_{g_i} = \begin{cases} \cos^{-1}((v_{i-1} \times v_i)^T (v_i \times v_{i+1})) \\ \quad \text{if } v_i^T ((v_{i-1} \times v_i) \times (v_i \times v_{i+1})) > 0 \\ 2\pi - \arccos((v_{i-1} \times v_i)^T (v_i \times v_{i+1})) \\ \quad \text{otherwise.} \end{cases} \quad (14)$$

Hence, in summary, the mathematical formulation of the energy function is based on a protein model where the backbone of a protein chain that contains p residues (amino acids) is modeled as a serial chain of $p - 2$ virtual links connected by revolute joints. Here, there are three parameters that determine the geometry of each link—the radii of the pair of residues (namely the values ρ_i and ρ_{i+1}) and the link length l . Intuitively, the link length should be larger than $(l > \rho_i + \rho_{i+1})$, but less than $2(\rho_i + \rho_{i+1})$ in order to hinder another residue going in between. As expected, as each residue gets larger, the distance between it and its adjacent residues will also need to get larger, which implies a larger l value.

3. Simulation results

This section presents a series of simulation studies using protein's native structure from PDB. The simulations are done using a Java-based program designed and developed in the lab. The dynamic system is iterated with a differential solver, and the iteration number corresponding to an event of concern is recorded as the 'step time'. The program is capable of visually showing the temporal progression of the 3D protein folding from the initial conformation to the native conformation as well as that of the associated contact map. The distance between any adjacent pair of residues (virtual bond length) is taken to be around 3.79 and $\rho_i = 1, \forall i$. Let it be noted that larger radii values can easily be considered by changing the three parameters accordingly as explained in section 2.5. For example, a radius, $\rho_i = \rho_{i+1} = 2.5$, can easily be considered by also changing the link length l to $5 < l < 10.0$.

The following measures are considered.

3.1. Contact time

Given a $w \in G$ and the C_α pair $(m, n) \in Q$, let $c : Q \times G \rightarrow \{0, 1\}$ denote the Boolean-valued contact map defined as

$$c(m, n, w) = \begin{cases} 1 & \text{if } \delta_{mn}(w) < 7\text{\AA} \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

In the dynamical system as defined by equation (12), let the contact time function $t_c : Q \rightarrow R^+$ be the first time formation of the associated contact as defined by

$$t_c(m, n) = t, \quad \text{where } \delta_{mn}(w^t) < 7\text{\AA} \quad \text{and} \quad \delta_{mn}(w^{t-1}) \geq 7\text{\AA}. \quad (16)$$

3.2. Secondary structure formation

Let us define the index set of secondary structures in a protein chain by $S = \{1, \dots, N_s\}$, where N_s is the number of secondary structures. A secondary structure

Table 1. Sample average contact time and time evolution graphs for two protein A folding simulations that serve to demonstrate that the two measures capture different features of the folding process.

Secondary Structures (Regions)	α 1 (9-18)	α 2 (24-38)	α 3 (41-57)
α 1 (9-18)	167.5	N/A	202
α 2 (24-38)	N/A	173	201
α 3 (41-57)	202	201	165.6

Secondary Structures (Regions)	α 1 (9-18)	α 2 (24-38)	α 3 (41-57)
α 1 (9-18)	163.2	N/A	202
α 2 (24-38)	N/A	171.3	201
α 3 (41-57)	202	201	164

$i \in S$ is defined by $S_i = \{s_i, \dots, f_i\}$ where $s_i, f_i \in P$ denote the start and the end residue numbers of secondary structures on protein. Define $Q_i = \{(m, n) \in Q : m, n \in S_i, m \neq n, n \neq m - 1, n \neq m + 1\}$. For each secondary structure $i \in Q$, the average time $T_A : Q \rightarrow R^+$ is defined by

$$T_A(i) = \frac{1}{|Q_i|} \sum_{(m,n) \in Q_i} t_c(m, n). \quad (17)$$

Let it be noted that the longer the secondary structure is, the longer time is taken for all the natural contacts to form. Thus, the average formation time measure does not differentiate between native structures with different lengths.

3.3. Tertiary structure formation

The average time for tertiary structure formation is also computed in a similar manner. For each pair (i, j) of secondary structures where $i, j \in S$, the average time $T_A : Q \times Q \rightarrow R^+$ of tertiary structure formation is defined as

$$T_A(i, j) = \frac{1}{|S_i||S_j|} \sum_{m \in S_i} \sum_{n \in S_j} t_c(m, n) \quad (18)$$

contacts.

3.4. Time-evaluation graphs

Time-evolution graphs show the completion of the formation of the secondary and tertiary structures with respect to time as measured by step times. The amount of completion is measured by the percentage of the natural contacts completed at the particular moment with respect to the total number of natural contacts. Given $t \in R^+$, recall that w^t denotes the

conformation at the time step t . For each secondary structure $i \in Q$, define the number of natural contacts formed as follows:

$$h_i(w^t) = \frac{1}{|Q_i|} \sum_{(m,n) \in Q_i} c(m, n, w^t). \quad (19)$$

For each tertiary structure $i, j \in Q$,

$$h_{ij}(w^t) = \frac{1}{|S_i||S_j|} \sum_{m \in S_i} \sum_{n \in S_j} c(m, n, w^t). \quad (20)$$

As the system starts to fold, at each time instant t , for each secondary or tertiary structure, we use a blue–green–red coding scheme $[R_{ij}(w), G_{ij}(w), B_{ij}(w)]$ to show the completion amount. For a tertiary structure $i, j \in S$, the color-code functions are computed as follows:

$$\begin{aligned} R_{ij}(w) &= \begin{cases} 2h_{ij}(w^t) - 1 & \text{if } h_{ij}(w^t) > 0, 5 \\ 0 & \text{otherwise} \end{cases} \\ G_{ij}(w) &= \begin{cases} 2 - 2h_{ij}(w^t) & \text{if } h_{ij}(w^t) > 0, 5 \\ 2h_{ij}(w^t) & \text{otherwise} \end{cases} \\ B_{ij}(w) &= \begin{cases} 1 - 2h_{ij}(w^t) & \text{if } h_{ij}(w^t) < 0, 5 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \quad (21)$$

Let it be remarked that the average contact time and time evaluation graphs present two different features of the folding process. While the former indicates an average time at which the secondary or the tertiary structure formation is completed, the latter shows the time at which the structure begins to form. As the completion of the whole structure may take time, an early start in the formation may not necessarily imply a smaller average contact time. A sample case is as shown in table 1, which shows the results of two different folding

instances for a protein starting from two different random initial conformations. In the first simulation, it is observed that $\alpha 3$ helix starts to form earlier and is completed before $\alpha 1$ helix. However, in the second simulation, although its structure starts to form prior to $\alpha 1$ helix, its completion as measured by the average contact time occurs later than that of $\alpha 1$ helix. Nevertheless, average contact times are very similar for these two helices. The average contact time for $\alpha 3$ helix, the middle helix, is longer. It is worthy of note here that $\alpha 2$ and $\alpha 3$ helices are relatively longer helices and that should be reflected in the average contact times. With this, $\alpha 3$ helix may be more prone to folding than implied by its average contact time.

3.5. Sample cases: proteins A, G and L

We first present detailed results for three relatively small proteins: protein A (*Staphylococcus Aureus* Protein A, immunoglobulin-binding B domain), protein G (streptococcal protein G, immunoglobulin-binding domain B1) and protein L (binding protein(immunoglobulin 1 chain)). The PDB files [37] used for the proteins were 1bdd.pdb (Protein A), 1gb1.pdb (Protein G) and 2ptl.pdb (Protein L), respectively. The native structure and the contact maps are as shown in figure 1:

- Protein A is relatively a small structure comprised of 60 residues with the native conformation and corresponding contact map. It has three α helices; $\alpha 1$, $\alpha 2$ and $\alpha 3$. In terms of torsion angles, 60 residues are associated with 115° of freedom.
- Protein G is a protein with 56 residues (107° of freedom) and one α helix and four β structures (2 β hairpins) with the native conformation.
- For protein L, a 62 residue sequence starting with the 16th residue is used. With this, there are 107° of freedom with one α helix and four β structures (2 β hairpins).

These proteins are selected as they have been extensively studied by both experimental and computational means including a previously proposed, planning-based approach. Protein A and protein G are topologically two different structures, whereas proteins G and L have similarity with respect to secondary and tertiary structures. It will be important to reproduce the differences in the formation order of secondary versus tertiary contacts between protein A and protein G, and the difference in hairpin formation order for protein G and protein L. The initial conformations are taken to be the extended chain conformation. It should be remarked that the initial conformation can be selected to be any ‘physically feasible’ conformation which means mutual avoidance in our case.

Statistical results from a series of simulations are presented for both proteins. In each case, approximately 250 random initial conformations are generated. The generation of random conformations is realized as follows:

- The set of all transient conformations from an extended initial conformation to the native conformation as generated in section 3.5 is considered.
- The set of intermediate conformations generated in the first 40% of a sample run is taken.

- A perturbation of total magnitude 20π radians is assumed.
- The number of angles to be perturbed is randomly selected to be between 1 and $2p - 5$ as denoted by N_p .
- N_p angles from all the degrees of freedom are randomly selected where each is assigned a random weight.
- Each selected angle is subjected to a perturbation where the perturbation amount is based on the ratio of its associated random weight with respect to the sum of all weights and the total perturbation magnitude.
- If the generated conformation is not in G , it is discarded.

For each study, the variation of the generated random initial conformations with respect to their Euclidean proximity to the native conformation is provided using $\gamma : G \rightarrow [0, \infty)$ defined as

$$\gamma(w) = \|b(w) - b(w_g)\|^2. \quad (22)$$

3.6. Folding of protein A

Snapshots taken from one of the simulations describe the folding pathway with the associated contact maps in figure 2. As is seen, first the gradual growth of the secondary structures is observed, which is followed by simultaneous stabilization of the helices coupled with the formation of the tertiary contacts. The corresponding functions γ (measuring the overall proximity to the goal conformation), β (measuring the overall proximity between the residues), the artificial potential function $\tilde{\varphi}$ (APF) and k parameter are shown in figure 3. The function γ has a decreasing trend which implies that the Euclidean distance between the current conformation and the native topology decreases in a continual manner. It is observed that β is also decreasing which indicates that the native conformation is more packed as compared to the transition conformations, and this packedness is achieved only toward the later stages of the folding process. The high jump in $\tilde{\varphi}$ is a direct consequence of the increase in k value, which tends to balance the relative magnitudes of the attraction to the goal conformation and the repulsion existing between the residues.

Table 2 shows the contact times and the time-evolution graph for this simulation. It is observed that $\alpha 3$ helix is formed first, followed by $\alpha 1$ helix and $\alpha 2$ helix. Although the average time of the natural contact formation for each helix is similar, the helices close to the termini start to fold earlier than the middle helix as seen in table 2. However, their formation times are very close to each other. Let it be noted that these results agree with the results of hydrogen exchange experiments [27], where, as presented therein, all helices fold simultaneously by ‘pulse labeling’. Once the formation of the secondary structures is close to completion, the tertiary structures evolve. The tertiary contacts between $\alpha 2$ helix and $\alpha 3$ helix form first followed by those between $\alpha 1$ helix and $\alpha 3$ helix. As is seen, the contacts between $\alpha 2$ helix and $\alpha 3$ helix start to form before the contacts of the individual helices are fully formed. Again here, the length of the helices may affect the average contact times; that could emphasize the early formation of this C-terminal hairpin of $\alpha 2$ and $\alpha 3$ helices

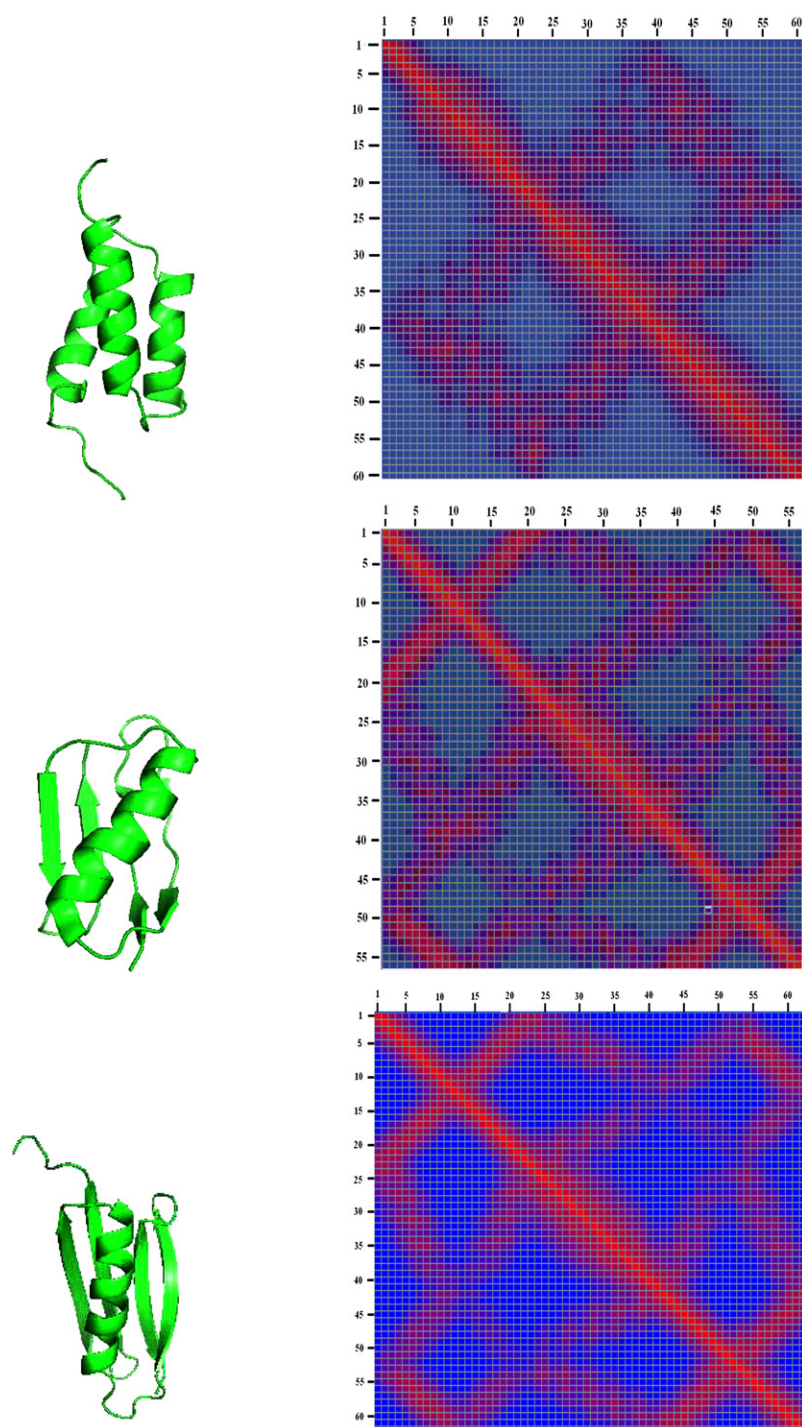


Figure 1. Native structure (left) and contact map (right): protein A (top), protein G (middle) and protein L (bottom).

more than reflected with its average contact time. The early formation of $\alpha 2$ helix and $\alpha 3$ helix contacts agrees with the equilibrium hydrogen experiments [8] and recent mutational analysis [38]. Also, several recent computational studies [18, 31] agree with this early contact formation between $\alpha 2$ helix and $\alpha 3$ helix, although $\alpha 1$ helix and $\alpha 2$ helix hairpin formation was suggested in another study [40]. Nevertheless, there also exist some variations in the results for the order of the formation of $\alpha 1$ helix and $\alpha 3$ helix between various computational studies. For example, the results from [3, 18,

31] suggest relatively early folding of $\alpha 3$ helix, whereas the results of [11] favor the formation of $\alpha 1$ helix. Our results also agree with the observed paths by [5], except the order of the individual helices.

The results for protein A, in general, suggest that the secondary structure folds first which, in turn, indicates that the formation of short-range contacts along the sequence is favored first. This agrees also with other computational studies [6, 31, 40], but disagrees with the study of [18] where simultaneous formation of the secondary and tertiary contacts

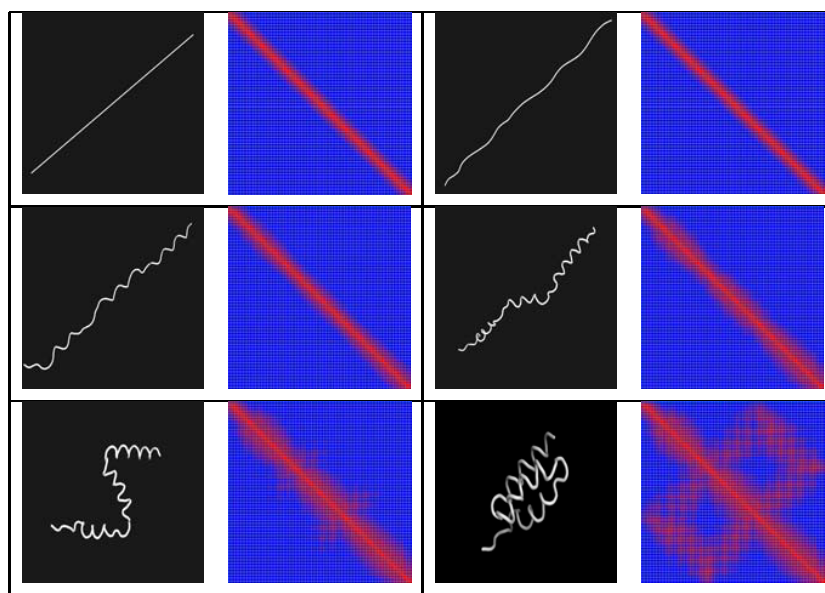


Figure 2. Snapshots and the corresponding contact maps from a sample protein A folding from extended conformation.

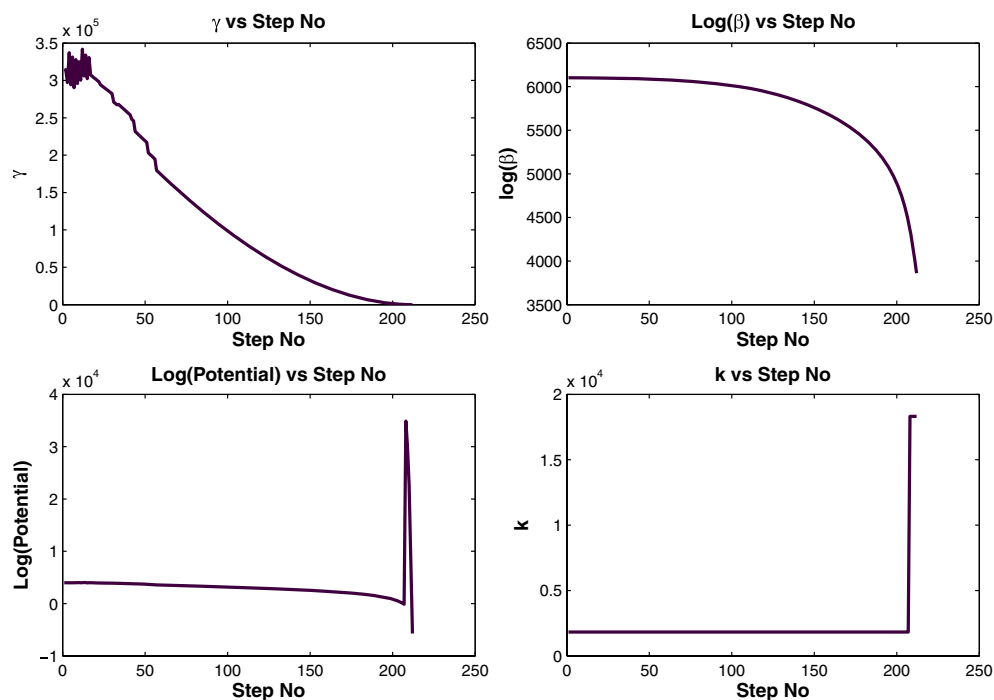


Figure 3. Time evolution of the γ , β , φ and k functions in folding of protein A.

Table 2. Results from a folding simulation of protein A from extended conformation. Left: contact times T_A and right: color-coded time evolution.

Structures (Regions)	α 1 (9-18)	α 2 (24-38)	α 3 (41-57)
α 1 (9-18)	177.6	N/A	212
α 2 (24-38)	N/A	183.2	211
α 3 (41-57)	212	211	175.7

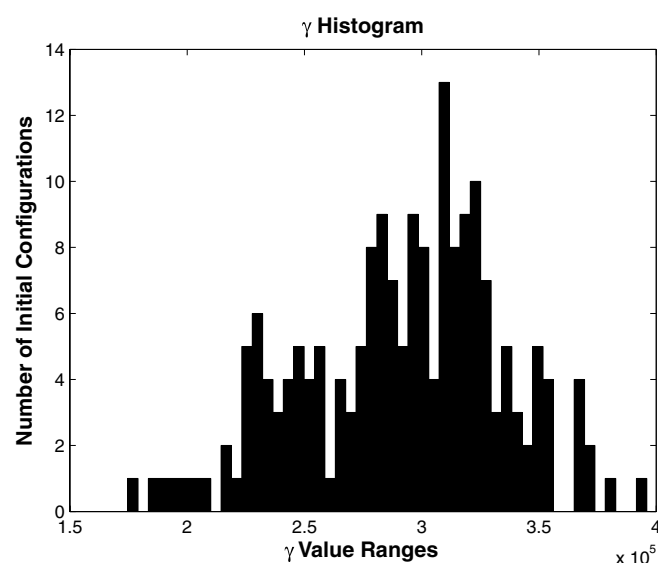


Figure 4. The histogram of the initial conformations as a function of γ for protein A.

is suggested. Although simultaneous formation of secondary and tertiary contacts to some extent is observed in the present simulations, the present results for protein A possibly suggest the zipper process [15]. The formation of a helix-turn-helix motif during the folding for three helix bundles was suggested to be an universal mechanism [40], although the details of which hairpin is formed may vary. The variations between various computational studies may lie in the existence of multiple pathways.

We also conducted an extensive statistical study where the initial conformations are generated as discussed in section 3.5. The variation of the generated random initial conformations with respect to their proximity to the native conformation is presented in figure 4. It is observed that the set of initial conformations is away from the native conformation with an approximately mean value of 3×10^5 , which is equal to the squared sum of the angle deviations. Table 3 presents the statistical results obtained for protein A, where the formation order is quite robust with two competing alternatives. These two are identical except the first-formed secondary structure. In 60% of the simulations, $\alpha 1$ forms first while $\alpha 3$ is the first-forming secondary structure in the remaining 40%. As is seen, the analysis of multiple simulation trajectories allows seeing parallel pathways. The early formation of either of terminus helices is possible. The contact formation of helices $\alpha 2$ and $\alpha 3$ consistently occurs in the same order.

Let it be noted that the iteration times are relatively unaffected by the value of the ρ_i parameters. For example, for protein A, if $1 < \rho_i < 2$, the iteration number remains nearly the same. For if $2 < \rho_i < 2.5$, it increases very slightly (about 5%).

3.7. Folding of protein G

Sampled snapshots taken from one of the folding simulations and the corresponding contacts maps are presented in figure 5. In comparison to protein A, here the formation

Table 3. The percentages of sequences in statistical runs of protein A with 239 random initial conformations.

Sequence of events	Percentage
$\alpha 1, \alpha 3, \alpha 2, \alpha 2-\alpha 3, \alpha 1-\alpha 3$	60
$\alpha 3, \alpha 1, \alpha 2, \alpha 2-\alpha 3, \alpha 1-\alpha 3$	40

of some tertiary contacts is observed with the growth of the secondary structures. The time evolutions of the γ , β , ϕ and k parameters in protein G simulation are shown in figure 6. It is observed that γ has a decreasing trend (with some little ripples) which again implies that the distance between the current conformation and the native topology is decreasing continuously.

The time evolution of the native contacts for the secondary and tertiary structures for the simulation run of figure 6 and the simulation average times of the contact formation between the secondary structures are presented in table 4. The time evolution of contacts indicates that $\beta 1$ and $\beta 4$ start to fold earlier. Simultaneously, α helix starts to fold and at the same time to form tertiary contacts with $\beta 2$. The $\beta 2$ and $\beta 3$ contacts complete around similar times. The formations of tertiary contacts between $\beta 1-\beta 2$ and $\beta 3-\beta 4$ are of very similar times, yet the C-terminal hairpin is observed earlier. These observations agree with the combined results of ‘out exchange’ and ‘pulse labeling’ [27] as α , $\beta 1$ and $\beta 4$ are the secondary structures that start to form their local contacts at relatively earlier times than $\beta 2$ and $\beta 3$. Additionally, the contacts between α and $\beta 2$ start to form before the complete formation of their secondary structures. The effects of several mutations suggested that the helix’s C-terminus is better defined than the rest of the helix at the folding transition state ensemble [42].

These results also agree with those of computational studies: α helix and hairpin of $\beta 3$ and $\beta 4$ are major intermediates [43, 44]; the most local contacts occur in α helix and the contacts between $\beta 1-\beta 2$ and $\beta 3-\beta 4$ form at the last stage of the folding [39]; the formation of the N-terminal and C-terminal hairpins of $\beta 1-\beta 2$ and $\beta 3-\beta 4$, respectively, and then the packing of α helix and the two hairpins [31]. Our results also agree with the findings of [6].

We also conducted an extensive statistical study for this protein whose methodology is as discussed in section 3.5. The variation of the generated random initial conformations with respect to their proximity to the native conformation is presented in figure 7. It is observed that each set of initial conformations is away from the native conformation with an approximately mean value of 3×10^5 . In a similar manner, table 5 presents the results for protein G. The contact formation between $\beta 1$ and $\beta 4$ occurs always last after the N- and C-terminal hairpins of $\beta 1$ and $\beta 2$, and $\beta 3$ and $\beta 4$. The latter hairpin is observed earlier than the former. The early appearance of local native contacts in $\beta 1$ and $\beta 4$ is observed in all runs, yet the formation of native contacts of α , $\beta 2$ and $\beta 3$ change—a phenomenon possibly related to the simultaneous interactions between α and $\beta 2$.

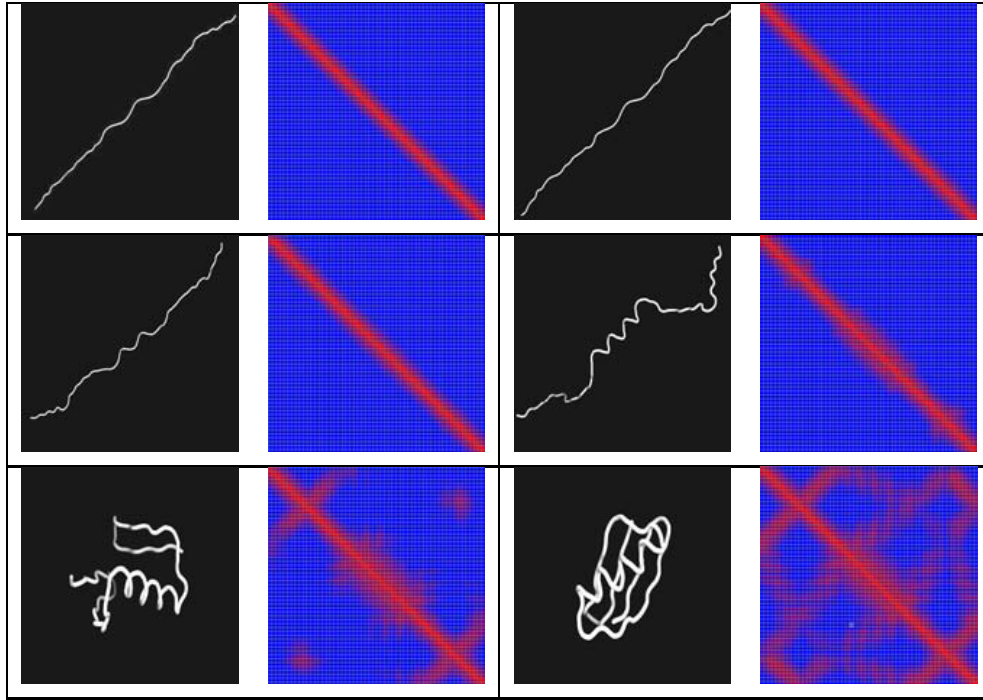


Figure 5. Snapshots and the corresponding contact maps from a sample protein G folding from extended conformation.

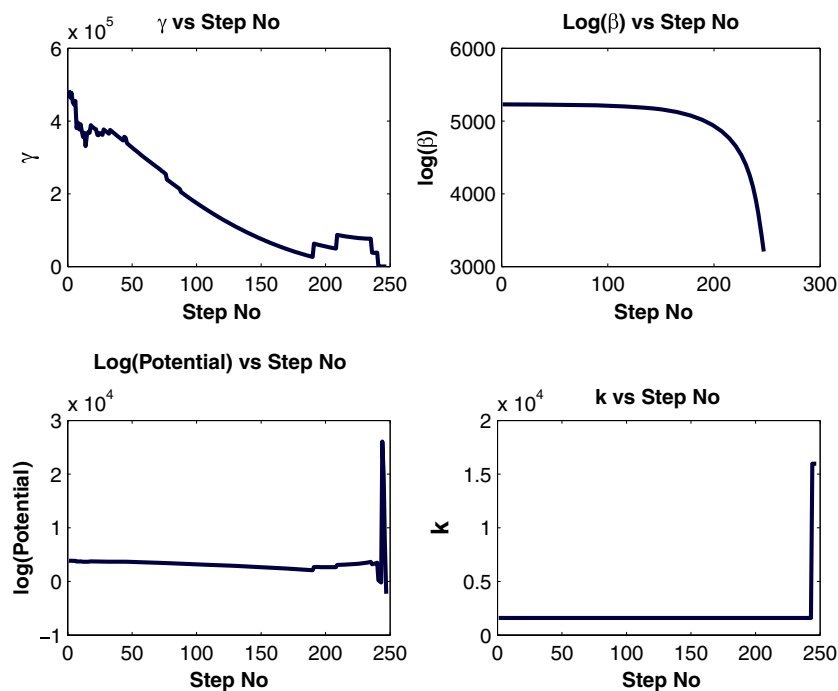


Figure 6. Time evolution of the γ , β , ϕ and k functions in folding of protein G.

3.8. Folding of protein L

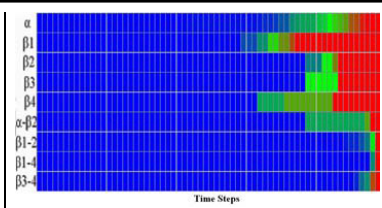
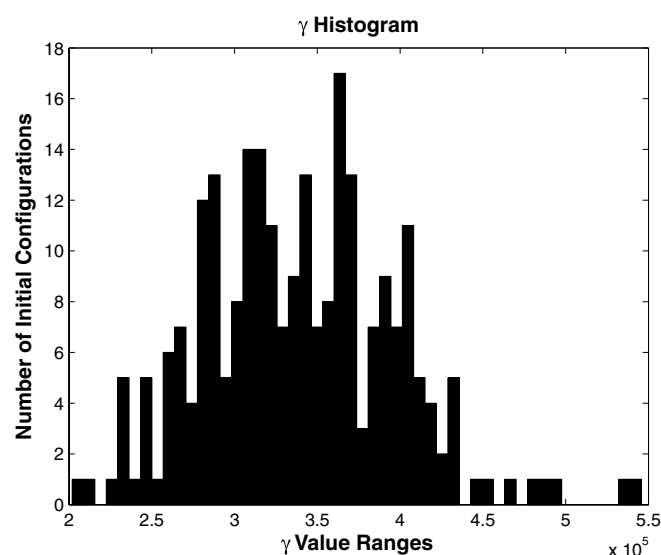
The time evolution of native contacts of the secondary and tertiary structures is presented in table 6 (left) where the average formation times are listed. As is seen in table 6 (right), the time evolution of the native contacts of β_4 and β_1 starts to appear earlier. Following α helix starts to fold. Table 7 presents the statistical results of the sequence of events for this

protein. As is seen therein, the formations of tertiary contacts between of β_1 – β_2 form first followed by those of β_3 – β_4 .

Finally, let us remark that a comparative analysis of these three proteins reveals the following. First, as is seen in figure 8, in all three proteins, the evolution of native contacts suggests a cooperative behavior, yet seems to deviate from single exponential behavior, which should be a characteristic behavior for two state cooperative folders such as those studied in this work. This might be due to the potential function that

Table 4. Results from a folding simulation of protein G from a random conformation. Left: contact times T_A and right: color-coded time evolution.

Structures (Residues)	Contact times
α (22-36)	161.7
β 1 (1-9)- β 2 (12-20)	189.5
β 3 (42-46)- β 4 (51-56)	188.6


**Figure 7.** The histogram of the initial conformations as a function of γ in the simulations for protein G.

has mainly a native-centric construct. The incorporation of residue-specific intra-chain interactions in addition to chain connectivity and excluded volume constraints may contribute to the generic cooperative features of these proteins. Next, these three proteins reveal different folding patterns. In protein A, three α helices considerably form first before packing into the final tertiary structure. In contrast, protein G (domain B1) seems to form the secondary structure gradually on the way to the tertiary structure as also pointed out in [7]. In protein L, the formation of the secondary structures is reversed. The hairpin of β 1- β 2 appears slightly earlier than the hairpin of β 3- β 4 in protein L, where the order is reversed in protein G.

Proteins L and G are two single-domain proteins that have little sequence identity but very similar fold topologies, consisting of a central α helix with a four-strand β sheet composed of two β hairpins. Experimental evidence suggests that protein L folds through a transition state involving a native-like N-terminal β hairpin [41]. Protein G, on the other hand, folds through formation of C-terminal β hairpin [42]. To this end, it should be noted that since the present proposed methodology does not have any sequence-dependent potential function, for the cases when sequence is relatively more important with respect to the effects of native-state topology for the folding, the contribution of the sequence may lack. Here, the contribution of the sequence could only be through its impact on the protein's topological features. That may

Table 5. The percentages of sequences in statistical runs of protein G with 248 random initial conformations.

Sequence of events	Percentage
α , α - β 2, β 3- β 4 , β 1- β 2 , β 1- β 4	100

imply that the difference in the order of formation of β -hairpins between protein L and protein could be more pronounced than observed in the present simulations.

3.9. Comparative study: APF versus PRM

This section presents a comparative study of artificial potential function based folding of six different protein structures with the results obtained with a previously proposed method based on probabilistic roadmaps [4]. The comparison is based on the order of secondary and tertiary structures occurrence. Six different protein structures as given in table 8 are considered. The results are presented in table 8, where each entry gives the structures that are ordered with respect to average contact times. Let it be noted that APF results present the average contact times for all the secondary and tertiary structures while with the other approach that is not necessarily the case as contact times only for a subset of the structures has been provided in [5]. Note that the first three proteins correspond to protein A, protein G and protein L, respectively; the detailed analysis of each is presented in the previous section. For the remainder protein structures, the following results hold:

- *SRC tyrosine kinase SH3 domain (1SRL)*. The sequence of events is β 4- β 5, β 3- β 4, β 2- β 3, β 1- β 5 and β 1- β 2. Experimental data are not available.
- *SH3 domain from tyrosine kinase (1NYF)*. The order turns out to be β 3- β 4, β 2- β 3, β 1- β 2. Experimental data are not available.
- *α -spectrin SH3 domain (1SHG) 3*. The sequence of events is β 3- β 4, β 2- β 3, β 1- β 5 and β 1- β 2. Let us note that this is slightly different from that of PRM which indicates that the tertiary structures β 3- β 4 and β 2- β 3 occur about the same time followed by β 1- β 5 and β 1- β 2 with contact times also very similar. Experimental data are not available.
- *Ubiquitin (1UBQ)*. The order turns out to be α 1, β 3- β 4, β 1- β 2, β 3- β 5 and β 1- β 5. This also agrees with the experimental data.

It is observed that there is a significant similarity between our results and PRM-based predictions [5]. Additionally, the

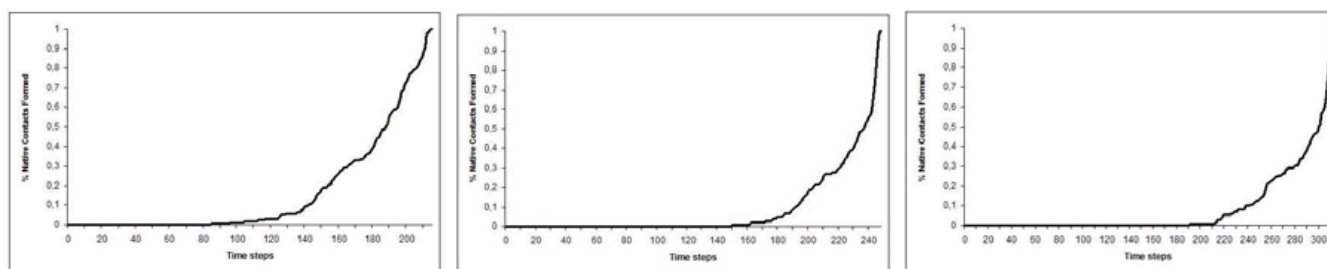


Figure 8. % Native contacts formed versus time steps. Left: protein A, center: protein G and right: protein L from initial extended chain conformation.

Table 6. Results from a folding simulation of protein L from a random conformation. Left: contact times T_A and right: color-coded time evolution.

Structures (Residues)	Contact Times
α (25-37)	258
β 1 (5-9)- β 2 (16-19)	288.1
β 3 (45-47) - β 4 (55-60)	288.6

formation order of the secondary and tertiary structures agrees with the experimental observation of [27], whenever available.

In general, the formation order of the secondary and the tertiary structures formed during the folding pathways is in agreement with the corresponding results of the PRM method as well as those of other computational and experimental results. These findings indicate that APF is a potentially valuable tool since in contrast to planning-based approaches, its computational complexity does not grow exponentially with the size of the proteins.

3.10. Summary

In summary, the following remarks can be made:

- A simple energy model using a nonlinear compositional scheme of encoding the goal conformation and forbidden regions (motivated by work in robot navigation) also seems promising for studying protein folding with known native conformations.
- In contrast to previous formulations where the forbidden conformations are encoded via additive terms, here the approach uses the ratio of proximity to the goal and distance among the C_α residues where—in contrast to previous work—mutual avoidance of bodies guaranteed.
- The dynamical nature of the formulation enables the natural generation of realistic conformational states between the initial and the final native topologies.
- The formation orders thus obtained turn out to be similar to the reported experimental and computational results, while the simulation times do not grow exponentially with the protein size.

4. Conclusion and outlook

This paper presents a new approach to the study of protein folding dynamics with known native structures based on

Table 7. The percentages of sequences in statistical runs of protein L with 28 random initial conformations.

Sequence of events	Percentage
α , β 1- β 2, β 3- β 4, β 1- β 4, α - β 1	100

artificial potential functions. The artificial potential function is basically a simplified energy function, but with the novelty that—motivated by work on robotic navigation—a nonlinear compositional scheme of construction is adapted instead of an additive formulation. Here, the artificial potential function is constructed on the conformation space as the ratio of proximity to the native goal conformation and distance among the C_α residues. In contrast to previous work, the avoidance of the forbidden regions corresponding to the physically infeasible conformations is guaranteed by construction as opposed to explicit checking. The protein folding pathway is then generated via a dynamic system built based on the negative gradient of the artificial potential function. The comparison of results from extensive simulations with similar results from other approaches including experimental data indicates that this approach is successful for coarsely approximating the folding dynamics. As it does not require the sampling of the conformation space as is required by previous methods, it is relatively fast. Hence it promises to be a valuable tool for studying protein structure formation for proteins whose native structures are known as well exploring the intermediate conformations between two or more known structures of a protein. It is also a potential tool for studying the binding phenomena, where both conformational and configurational schemes need to be considered. Currently, we are working on developing a web-based server for the usage of the community.

Table 8. Comparison of APF with PRM [6] and experimental results [27].

PDB code	p (#DOF)	APF-based method	PRM-based method [6]	Exp [27]
1BDD	60 (115)	$\alpha 1, \alpha 3, \alpha 2, \alpha 2-\alpha 3, \alpha 1-\alpha 3$	$\alpha 2, \alpha 3, \alpha 1, \alpha 2-\alpha 3, \alpha 1-\alpha 3$	Agreed
1GB1	56 (107)	$\alpha 1, \beta 3-\beta 4, \beta 1-\beta 2, \beta 1-\beta 4$	$\alpha 1, \beta 3-\beta 4, \beta 1-\beta 2, \beta 1-\beta 4$	Agreed
1SRL	56 (107)	$\beta 4-\beta 5, \beta 3-\beta 4, \beta 2-\beta 3, \beta 1-\beta 5, \beta 1-\beta 2$	$\beta 4-\beta 5, \beta 3-\beta 4, \beta 2-\beta 3, \beta 1-\beta 5, \beta 1-\beta 2$	N/A
1NYF	58 (111)	$\beta 3-\beta 4, \beta 2-\beta 3, \beta 1-\beta 2$	$\beta 3-\beta 4, \beta 2-\beta 3, \beta 1-\beta 2$	N/A
1SHG	57 (109)	$\beta 3-\beta 4, \beta 2-\beta 3, \beta 1-\beta 5, \beta 1-\beta 2$	$(\beta 2-\beta 3 \beta 3-\beta 4), (\beta 1-\beta 2 \beta 1-\beta 5)$	N/A
1UBQ	78 (147)	$\alpha 1, \beta 3-\beta 4, \beta 1-\beta 2, \beta 3-\beta 5, \beta 1-\beta 5$	$\alpha 1, \beta 3-\beta 4, \beta 1-\beta 2, \beta 3-\beta 5, \beta 1-\beta 5$	Agreed
2PTL	62 (119)	$\alpha 1, \beta 1-\beta 2, \beta 3-\beta 4, \beta 1-\beta 4$	$\alpha 1, \beta 1-\beta 2, \beta 3-\beta 4, \beta 1-\beta 4$	Agreed

Acknowledgments

This work is supported by TÜBİTAK MAG 106M077.

References

- [1] Alm E and Baker D 1999 Matching theory and experiment protein folding *Curr. Opin. Str. Biol.* **6** 189–96
- [2] Alm E and Baker D 1999 Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures *Proc. Natl. Acad. Sci. USA* **96** **20** 11305–10
- [3] Alonso D O V and Daggett V 2000 Staphylococcal protein A: unfolding pathways, unfolded states, and differences between the B and E domains *Proc. Natl. Acad. Sci. USA* **97** 133–8
- [4] Amato N M and Song G 2002 Using motion planning to study protein folding pathways *J. Comput. Biol.* **9** 149–68
- [5] Amato N M, Dill K A and Song G 2002 Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)* pp 2–11
- [6] Amato N M, Dill K A and Song G 2003 Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures *J. Comput. Biol.* **10** 239–55
- [7] Baldwin R L and Rose G D 1999 Is protein folding hierarchic? II. Folding intermediates and transition states *Trends Biochem. Sci.* **24** 77–83
- [8] Bail Y, Karimi A, Dyson H J and Wright P E 1997 Absence of a stable intermediate on the folding pathways of protein A *Protein Sci.* **6** 1449–57
- [9] Baker D 2000 A surprising simplicity to protein folding *Nature* **405** 39–42
- [10] Bozma H I and Koditschek D E 2001 Assembly as a noncooperative game of its pieces: analysis of 1d sphere assemblies *Robotica* **19** 93–108
- [11] Brooks C L 2002 Viewing protein folding from many perspectives *Proc. Natl. Acad. Sci. USA* **99** 1099–100
- [12] Chavez L L, Gosavi S, Jennings P A and Onuchic J N 2006 Multiple routes lead to the native state in the energy landscape of the beta-trefoil family *Proc. Natl. Acad. Sci. USA* **103** 10254–8
- [13] Covell D G 1992 Folding protein-carbon chains into compact forms by Monte Carlo methods *Proteins: Struct. Funct. Genet.* **14** 409–20
- [14] Daggett V and Levitt M 1993 Realistic simulation of naive-protein dynamics in solution and beyond *Annu. Rev. Biophys. Biomol. Struct.* **22** 353–80
- [15] Dill K A, Fiebig K M and Chan H S 1993 Cooperativity in protein-folding kinetics *Proc. Natl. Acad. Sci. USA* **90** 1942–6
- [16] Duan Y and Kollman P A 1998 Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution *Science* **282** 740–4
- [17] Haile J M 1992 *Molecular Dynamics Simulation: Elementary Methods* (New York: Wiley)
- [18] Ghosh A, Elber R and Scheraga A 2002 Atomically detailed study of folding pathways of protein A with stochastic difference equation *Proc. Natl. Acad. Sci. USA* **99** 10394–8
- [19] Karagöz C S, Bozma H I and Koditschek D E 2004 Feedback-based event-driven parts moving *IEEE Trans. Robot. Autom.* **20** 1012–8
- [20] Koditschek D 1987 Exact robot navigation by means of potential functions: some topological considerations *Proc. IEEE Conf. on Robotics and Automation* pp 1–7
- [21] Koditschek D E and Rimon E 1990 Robot navigation functions on manifolds with boundary *Adv. Appl. Math.* **11** 412–42
- [22] Koditschek D E 1994 An approach to autonomous robot assembly *Robotica* **12** 137–55
- [23] Koditschek D E 1991 The control of natural motion in mechanical systems *ASME, J. Dyn. Syst., Meas. Control* **113** 547–51
- [24] Kolinski A and Skolnick J 1994 Monte Carlo simulations of protein folding *Proteins Struct. Funct. Genet.* **18** 338–52
- [25] Levinthal C 1968 Are there pathways for protein folding? *J. Chim. Phys. Phys.-Chim. Biol.* **65** 44–5
- [26] Levitt M 1983 Protein folding by restrained energy minimization and molecular dynamics *J. Mol. Biol.* **170** 723–64
- [27] Li R and Woodward C 1999 The hydrogen exchange core and protein folding *Protein Sci.* **8** 1571–91
- [28] Mossa A and Clementi C 2007 Supersymmetric Langevin equation to explore free-energy landscapes *Phys. Rev. E* **75** 046707
- [29] Mufoz V and Eaton W A 1999 A simple model for calculating the kinetics of protein folding from three dimensional structures *Proc. Natl. Acad. Sci. USA* **96** 11311–6
- [30] Munoz V 2007 Conformational dynamics and ensembles in protein folding *Annu. Rev. Biophys. Biomol. Struct.* **36** 395–412
- [31] Ozkan S B, Wu G A, Chodera J D and Dill K A 2007 Protein folding by zipping and assembly *Proc. Natl. Acad. Sci. USA* **104** 11987–92
- [32] Newman W S and Hogan N 1987 High speed robot control and obstacle avoidance using dynamic potential functions *Proc. IEEE Conf. on Robotics and Automation* pp 14–24
- [33] Rimon E and Koditschek D E 1992 Exact robot navigation using artificial potential functions *IEEE Trans. Robot. Autom.* **8** 501–18
- [34] Scheraga H, Khalili M and Liwo A 2007 Protein-folding dynamics: overview of molecular simulation techniques *Annu. Rev. Phys. Chem.* **58** 57–83
- [35] Song G and Amato N M 2004 A motion planning approach to folding: from paper craft to protein folding *IEEE Trans. Robot. Automat.* **20** 60–71
- [36] Song G, Thomas S L, Dill K A, Scholtz J M and Amato N M 2003 A path planning-based study of protein folding with a

- case study of hairpin formation in protein G and L *Proc. Pacific Symp. Biocomputing (PSB)* pp 240–51
- [37] *The Protein Data Bank* 2006 <http://www.rcsb.org/pdb/>
- [38] Sato S, Religa T L, Dagget V and Fersht A R 2004 From the cover: testing protein-folding simulations by experiment: B domain of protein A *Proc. Natl. Acad. Sci. USA* **101** 6952–6
- [39] Sutto L, Tiana G and Broglia R A 2006 Sequence of events in folding mechanism: beyond the Go model *Protein Sci.* **15** 1638–53
- [40] Yang J S, Wallin S and Shakhnovich E I 2008 Universality and diversity of folding mechanics for three-helix bundle *Proc. Natl. Acad. Sci. USA* **105** 895–900
- [41] Gu H D, Kim D and Baker D 1997 Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein *J. Mol. Biol.* **274** 588–96
- [42] McCallister E L, Alm E and Baker D 2000 Critical role of b-hairpin formation in protein G folding *Nat. Struct. Biol.* **7** 669–73
- [43] Shimada J and Shakhnovich E I 2002 The ensemble of folding kinetics of protein G from an all-atom Monte Carlo simulation *Proc. Natl. Acad. Sci.* **99** 11175–80
- [44] Kmiecik A and Kolinski A 2008 Folding pathway of b1 domain of protein G explored by multiscale modeling *Biophys. J.* **94** 726–36