Efficient Generation of Feasible Pathways for Protein Conformational Transitions

Moon K. Kim,* Robert L. Jernigan,† and Gregory S. Chirikjian*

*Department of Mechanical Engineering, The Johns Hopkins University, Baltimore, Maryland 21218; and †Molecular Structure Section, Laboratory of Experimental and Computational Biology, CCR, NCI, NIH, Bethesda, Maryland 20892-5677

ABSTRACT We develop a computationally efficient method to simulate the transition of a protein between two conformations. Our method is based on a coarse-grained elastic network model in which distances between spatially proximal amino acids are interpolated between the values specified by the two end conformations. The computational speed of this method depends strongly on the choice of cutoff distance used to define interactions as measured by the density of entries of the constant linking/contact matrix. To circumvent this problem we introduce the concept of using a cutoff based on a maximum number of nearest neighbors. This generates linking matrices that are both sparse and uniform, hence allowing for efficient computations that are independent of the arbitrariness of cutoff distance choices. Simulation results demonstrate that the method developed here reliably generates feasible intermediate conformations, because our method observes steric constraints and produces monotonic changes in virtual bond and torsion angles. Applications are readily made to large proteins, and we demonstrate our method on lactate dehydrogenase, citrate synthase, and lactoferrin. We also illustrate how this framework can be used to complement experimental techniques that partially observe protein motions.

INTRODUCTION

Proteins are well known to be intrinsically flexible structures. Many proteins have been determined to have multiple conformations (in some cases called "open" and "closed" forms; Berman et al., 2000). Conformational transitions between two forms are often important for understanding the relationship between structure and function. In other words, such motions are involved in many basic functions such as catalysis, regulation, transportation, and aggregation (Subbiah, 1996). Hence, comprehending conformational transitions can be useful for understanding biological mechanisms, especially for protein machines.

However, it is also an important topic in molecular graphics to visualize conformational transitions. Obviously, one of the best ways is through animations, such as digital movie files (e.g., AVI or MPEG). Animations usually are produced by inserting images of intermediate conformations between the two conformations. These hypothetical intermediate conformations are visualized in sequence for an animation.

There have been several previous efforts in this area. Vonrhein et al. (1995) produced movies of conformational transitions by linear interpolation between the atomic coordinates of the two end conformations in Cartesian space. One problem with that method is that the bond lengths and angles of the intermediate conformations can be unrealistic, and in several cases protein chains actually pass through one another. To overcome this problem, Gerstein and Krebs (1998) applied proper restraints and minimized the energy of each intermediate conformation to correct for molecular stereochemistry and enforce rules of molecular structure.

An alternative interpolation approach is to use internal coordinates such as bond lengths, bond angles, and torsion angles instead. Kleywegt and Jones (1996) implemented this approach to construct intermediate conformations with their LSQMAN program. Ideally, this approach produces realistic bond lengths and torsional angles, but this method also has some problems. If one constructs intermediate conformations by interpolating torsional angles between two end conformations while holding bond lengths and angles fixed, one will often obtain infeasible pathways for several reasons. First, it may not even be possible for the conformation from one end to reach the other end in Cartesian space because the two conformations do not have identical values of internal variables such as bond lengths and bond angles. Therefore, one must either refine the two end conformations until they have a consistent set of internal variables, except for the torsion angles, before interpolating over torsion angles, or instead interpolate all of the internal variables simultaneously to avoid this problem. A second limitation is that in the process of generating intermediate conformations some residues can come too close to each other in order to not break the smoothness of the simulated pathway and this can produce unfavorable states in the sense of high-energy interactions and steric clashes. Fig. 1 shows that a particular pair of α-carbons can come too close to each other during conformational transitions using internal coordinate interpolation, which would give rise to exceedingly high repulsive energy peaks because of van der Waals forces between nonbonded atoms. A third problem occurs for specific values of internal rotation an-

Submitted January 3, 2002, and accepted for publication May 7, 2002.

Address reprint requests to Gregory S. Chirikjian, Dept. of Mechanical Engineering, The Johns Hopkins University, Baltimore, MD 21218. E-mail: gregc@jhu.edu.

R. L. Jernigan's present address is Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011-3020.

0006-3495/02/09/1620/11 $2.00

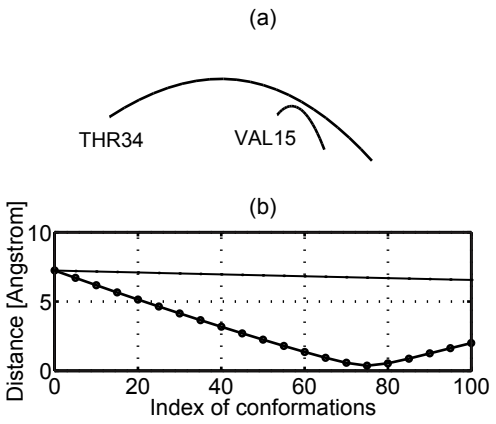

FIGURE 1 An example of torsional angle interpolation between two lac repressor headpiece structures (named 1LCC and 1LCD, Protein Data Bank). (*a*) During the conformational transition from 1LCC to 1LCD, the α-carbons of Val-15 and Thr-34 unrealistically come too close to each other, $\leq$1 Å. (*b*) This unrealistic relative distance between two atoms (*solid-dot line*) is compared to the result of the "distance interpolation method" developed in this paper (*solid line*). The conformation from one end does not reach the other end in Cartesian space because the two sets of crystallographic coordinates do not yield identical values of bond lengths and bond angles. One must either refine the two end conformations until they have a consistent set of internal variables, except for the torsion angles, before interpolating over torsion angles, or interpolate the full set of internal variables simultaneously. However, our method appears to conform well to steric constraints when reaching the other end without refinement of internal variables.

gles. Most are not equi-energetic for all values. Consequently, some forms have higher energies, and the intermediate forms generated could have inordinately high values, even if other lower energy pathways do exist.

There has been substantial work on the development of algorithms to generate plausible "reaction pathways." Elber and Karplus (1987) proposed to make a trial path connecting reactant and product structures. This path minimizes the functional

$$T = \frac{1}{L}\int_{\mathbf{r}_R}^{\mathbf{r}_P} U(\mathbf{r})dl(\mathbf{r}), \qquad (1)$$

where $l(\mathbf{r})$ is the reaction path connecting reactant ($\mathbf{r}_R$) and product ($\mathbf{r}_P$) structures, $U(\mathbf{r})$ is the potential energy, and $L = \int dl(\mathbf{r})$ is the path length (Czerminski and Elber, 1990). This algorithm demonstrates a good reaction path for the alanine dipeptide and the tetrapeptide IAN. However, the path that minimizes the functional above may not be the path with the maximum rate of transition between reactants and products. Jónsson et al. (1998) developed a "nudged elastic band method," which is a modified path method to find minimum energy paths by constructing a set of intermediate conformations between reactant and product structures. A spring interaction between adjacent conformations is added to ensure the continuity of the path. Minimization of the force acting on the conformations yields the minimum energy path. Another path method is the MaxFlux method proposed by Huo et al. (1997). It computes the reaction pathway of maximum diffusive flux by minimizing a discretized form of the line integral in Eq. 1 with added restraints such as constant mean-square distance between adjacent intermediates, repulsive interactions between adjacent intermediates along the path, and linear and angular momentum conservation for the system. This method was used to find the optimal pathway of the coil-to-helix transition in a short polyalanine chain, but these path methods have not been applied to a large protein because it is a computationally demanding task to find the global minimum value of the objective function in a high-dimensional space out of all possible reaction paths.

Some methods have considered probabilistic models of pathways. Olender and Elber (1996) integrated classical Newtonian dynamical equations of motion to compute long-time molecular dynamics trajectories based on the stochastic path integral. The activated dynamical transition path method developed by Dellago et al. (1997) generates and samples an ensemble of transition paths, which evolve according to stochastic dynamics (either Metropolis Monte Carlo or Brownian dynamics) and conserve the Boltzmann distribution. Again, these stochastic methods have been tested for simple cases such as the alanine dipeptide and two-dimensional Lennard-Jones clusters, but they also are computationally too expensive to be applied to a large protein.

Molecular dynamics (MD) simulations, a powerful method for the study of details of molecular motion, and normal mode analysis (NMA) using all-atom empirical potentials, are often used to follow the dynamics of proteins (Brooks and Karplus, 1985; Xu et al., 1997; Xu and Sigler, 1998). However, the use of atomic approaches becomes computationally inefficient with the increased size of a system.

To reduce such a computational burden, many recent papers have demonstrated the utility of coarse-grained protein models by including, for example, only α-carbons as point masses representing residues and using a simplified potential for considering internal interactions between neighboring residues. Such models are suitable to describe the global motions of complex systems of small proteins or single proteins having more than several thousand residues (Atilgan et al., 2001; Bahar and Jernigan, 1998; Bahar et al., 1999; Jaaskelainen et al., 1998; Tama and Sanejouand, 2001; Tirion and Ben-Avraham, 1993, 1998).

In this paper we develop a new interpolation method for generating feasible pathways for conformational transitions using the simplest potential and coarse-grained protein models. The key idea is to interpolate uniformly the distances between spatially proximal residues in both conformations within the context of the elastic network model. The present approach can be referred to as a "distance

interpolation method," which is completely different from position interpolation in Cartesian space. Because we interpolate relative distance between spatially close residues, unrealistic conformations and steric clashes become far less likely. This method offers a reasonable compromise between oversimplified linear interpolation in Cartesian or internal coordinates and computationally expensive methods such as MD simulations. We discuss this efficient modeling technique for large proteins. Our method computes sets of interpolated conformations within reasonable times in cases where full-atom computations such as MD or even NMA may be infeasible. Unlike dynamics-based methods in which the size of the timestep used is limited by the stiffest part of the structure, our method is purely geometric and so the number of required animation frames is dictated only by the difference in shape between the two conformations. An added advantage to this kind of model is that having virtual bonds attaching each α-carbon to all those that are within a sequential distance of three units induces stiffnesses in the virtual bond angles and torsion angles while retaining the ease of using Cartesian coordinates.

METHODS

Incremental construction of intermediate conformations

We derive here an incremental formulation to generate intermediate conformations. The key idea is to interpolate between two values for the distances between spatially proximal α-carbons, which are artificially connected with springs in the elastic network model (Atilgan et al., 2001; Bahar et al., 1997). Although the relationship between molecular conformations and the distances between atoms in conformations has been studied extensively (Crippen and Havel, 1988), our goal is to generate intermediate conformations by finding small changes in α-carbon positions that result from inducing correspondingly small changes in inter-residue distances.

Suppose that we have sets of α-carbon coordinates for the two end conformations of the same protein denoted by $\{\mathbf{x}_i\}$ and $\{\boldsymbol{\chi}_i\}$, respectively. One can build two elastic network models, one for each of these conformations. We introduce a cost function as follows

$$C(\boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}\{\|\mathbf{x}_i + \boldsymbol{\delta}_i - \mathbf{x}_j - \boldsymbol{\delta}_j\| - l_{i,j}\}^2. \quad (2)$$

Here $\boldsymbol{\delta} = [\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_n^T]^T$ is a $3n$-dimensional vector of displacements, with n being the number of residues. An intermediate conformation is defined by the value of $\boldsymbol{\delta}$ that minimizes this cost when all other parameters are held constant. The linking ("contact") matrix k is an $n \times n$ matrix containing the information about which amino acid residue is either connected to, or in contact with, any other. It is formed as the "union" of the two linking matrices for $\{\mathbf{x}_i\}$ and $\{\boldsymbol{\chi}_i\}$, so that $k_{i,j}$ can have value 1 whenever residues i and j are judged to be in contact in either conformation and 0 otherwise. The value $l_{i,j}$ is the desired distance between i and j, which can be chosen as

$$l_{i,j} = (1 - \alpha)\|\mathbf{x}_i - \mathbf{x}_j\| + \alpha\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|, \quad (3)$$

where α is the coefficient specifying how far a given state is along the transition from $\{\mathbf{x}_i\}$ to $\{\boldsymbol{\chi}_i\}$. For example, when $\alpha = 0.5$, the desired conformation is the one with inter-residue distances at the average values of those for conformations $\{\mathbf{x}_i\}$ and $\{\boldsymbol{\chi}_i\}$. Using the "union" linking matrices confines the intermediate conformations to the interval between the two end conformations.

The cost function in Eq. 2 can be related to the classical pairwise potential functions of biophysics in the following way. Suppose that a coarse-grained (C-alpha) potential function between any two residues i and j is $V_{i,j}(\|\mathbf{x}_i - \mathbf{x}_j\|)$. If from the crystal data we know that there are two conformations, we seek to establish a series of "artificial" equilibrium states by perturbing this potential at artificial time t. This results in an artificial potential of the form

$$V(\boldsymbol{\delta}, t) = \sum_{i,j} V_{i,j}(\|\mathbf{x}_i + \boldsymbol{\delta}_i - \mathbf{x}_j - \boldsymbol{\delta}_j\| - l_{i,j}(t)). \quad (4)$$

The goal at each value of time is then to let the system relax so that the values of the small displacements cause the new conformations to settle at new artificial equilibria. Because the i + 1st state is always calculated relative to the artificial equilibrium established for the ith state, a Taylor series expansion of each $V_{i,j}(r)$ up to quadratic order will result in an expression of the form in Eq. 2. Here r is the inter-residue distance. The linear term in the Taylor series expansion

$$V_{i,j}(r_0 + \epsilon) \approx V_{i,j}(r_0) + V'_{i,j}(r_0)\epsilon + \frac{1}{2}V''_{i,j}(r_0)\epsilon^2 \quad (5)$$

(where r_0 is the previous equilibrium value of inter-residue distance and ϵ is the small change) vanishes when summed over all values of i and j because of the definition of an equilibrium state. Even if the potential function is singular, such as in a six-twelve potential, the Taylor series expansion in a small neighborhood of an equilibrium is valid because the function will always be analytic locally. Hence, starting from an arbitrary potential function, one will always arrive at Eq. 2 when small incremental deviations from equilibrium are made. The only influence the potential will have is on the values of $k_{i,j}$, which are held constant over time and over all values of i and j for simplicity in the elastic network model, but could easily be allowed to vary within the framework given below to reflect any potential function.

Our goal is to find values of $\boldsymbol{\delta}$ that minimize Eq. 2, which itself can be linearized for small values of $\|\boldsymbol{\delta}_i\|$ and $\|\boldsymbol{\delta}_j\|$ with a Taylor series approximation. If we consider an individual term in Eq. 2

$$C_{i,j} = \tfrac{1}{2}k_{i,j}\{\|\mathbf{x}_i + \boldsymbol{\delta}_i - \mathbf{x}_j - \boldsymbol{\delta}_j\| - l_{i,j}\}^2, \quad (6)$$

then this can be written as

$$C_{i,j} \approx \tfrac{1}{2}k_{i,j}(C_{i,j}^{(1)} + C_{i,j}^{(2)} + C_{i,j}^{(3)}), \quad (7)$$

where

$$C_{i,j}^{(1)} = (\boldsymbol{\delta}_i - \boldsymbol{\delta}_j)^T\left[\mathbb{E}_3 - l_{i,j}\frac{A(\mathbf{x}_i - \mathbf{x}_j)}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right](\boldsymbol{\delta}_i - \boldsymbol{\delta}_j), \quad (8)$$

$$C_{i,j}^{(2)} = 2\left(1 - \frac{l_{i,j}}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right)(\mathbf{x}_i - \mathbf{x}_j)^T(\boldsymbol{\delta}_i - \boldsymbol{\delta}_j), \quad (9)$$

and

$$C_{i,j}^{(3)} = (\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j) - 2l_{i,j}\|\mathbf{x}_i - \mathbf{x}_j\| + l_{i,j}^2. \quad (10)$$

In Eq. 8 we use $\mathbb{E}_3$ to denote the 3×3 identity matrix, and

$$A(\mathbf{x}) = \mathbb{E}_3 - \frac{\mathbf{x}\mathbf{x}^T}{\|\mathbf{x}\|^2}.$$

If we take $G'_{i,j} \in \mathbb{R}^{3\times3}$ such that

$$G'_{i,j} = k_{i,j}\left[\mathbb{E}_3 - l_{i,j}\frac{A(\mathbf{x}_i - \mathbf{x}_j)}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right], \quad (11)$$

then

$$\frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}C^{(1)}_{i,j} = \frac{1}{2}\boldsymbol{\delta}^T\Gamma\boldsymbol{\delta} \quad (12)$$

for some $3n \times 3n$ matrix Γ, which can be divided into 3×3 blocks $\Gamma_{i,j}$. Generally, if $i \neq j$,

$$\Gamma_{i,j} = -G'_{i,j}. \quad (13)$$

When $i = j$, the result is

$$\Gamma_{i,i} = \sum_{k=1}^{i-1} G'_{k,i} + \sum_{k=i+1}^{n} G'_{i,k} = \sum_{k\neq i} G'_{k,i}. \quad (14)$$

Let $\mathbf{v}_{i,j} \in \mathbb{R}^{1\times3}$ be

$$\mathbf{v}_{i,j} = 2k_{i,j}\left(1 - \frac{l_{i,j}}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right)(\mathbf{x}_i - \mathbf{x}_j)^T. \quad (15)$$

Then

$$\frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}C^{(2)}_{i,j} = \frac{1}{2}\boldsymbol{\gamma}\boldsymbol{\delta}, \quad (16)$$

where

$$\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n] \in \mathbb{R}^{1\times3n} \quad (17)$$

and

$$\boldsymbol{\gamma}_i = -\sum_{k=1}^{i-1}\mathbf{v}_{k,i} + \sum_{k=i+1}^{n}\mathbf{v}_{i,k} = \sum_{k\neq i}\mathbf{v}_{i,k}. \quad (18)$$

Let B be

$$B = \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} k_{i,j}C^{(3)}_{i,j}. \quad (19)$$

When retaining terms to quadratic order, the result will have the form

$$C(\boldsymbol{\delta}) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_{i,j} \approx \frac{1}{2}\boldsymbol{\delta}^T\Gamma\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\gamma}\boldsymbol{\delta} + B, \quad (20)$$

where B is a constant. The matrix Γ is a $3n \times 3n$ matrix akin to a stiffness matrix that relates the relative cost of displacing any particular residue from its current position as compared to all other residues. We minimize $C(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$, which results in the following constraint equation:

$$\frac{\partial C(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} = \Gamma\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\gamma}^T = 0. \quad (21)$$

We note that $\Gamma \in \mathbb{R}^{3n\times3n}$ always has three zero eigenvalues corresponding to translation modes, because a translated version of $\boldsymbol{\delta}$ satisfying Eq. 21 can also minimize the cost function. That is, the solution of Eq. 21 is not unique. To solve this problem, one can either assume a particular point is fixed in space so that Γ can be reduced to a nonsingular (invertible) matrix, or add the constraint of linear momentum conservation such that

$$\sum_{i=1}^{n} m_i\boldsymbol{\delta}_i = 0. \quad (22)$$

In this case we take $m_i = 1$. Viewing Γ as a stiffness matrix, $\boldsymbol{\gamma}$ is like a generalized force that is used to push the system along the simplest realizable path connecting the two conformations.

Computational complexity

We have previously observed that the dynamic behavior and computational complexity of elastic network models of proteins vary with distance cutoff values defining interactions. Namely, large cutoff values yield increased numbers of interacting pairs. Consequently, the system becomes stiff, the amplitudes of fluctuations diminish with larger cutoff values, and the matrices describing the system become less sparse. Also for relatively short cutoff values, it is possible to get more than six zero eigenvalues corresponding to rigid-body modes in normal mode analysis, and there can be extremely large amplitude fluctuations along particular directions for particular residues (Atilgan et al., 2001). Likewise, our interpolation method, which is basically derived from a matrix similar to a stiffness matrix, is sensitive to cutoff values and the geometry of a given protein structure. Extremely short cutoff values will connect the residues only with their local neighbors. This can cause unrealistic results that lead to discontinuous animations. Adoption of larger cutoff values eliminates such behavior. However, denser linking matrices can increase tremendously the computation time required for generating intermediate transitions in large protein models composed of thousands of residues. The denser the matrix is, the longer the computation time is. Later we will demonstrate this relationship quantitatively. Fig. 2 illustrates how the sparsity pattern of the union linking matrices of lactoferrin (1LFG and 1LFH) depends on the cutoff values. When the cutoff value is 10 Å in (*a*), some residues have poor connections, so that the resulting pathway cannot be realistic. A larger cutoff value of 15 Å substantially increases the density of the linking matrix.

We address a new way to produce uniformly sparse linking matrices. The method reduces computational costs for the whole interpolation process and also guarantees realistic results. For this purpose, a linking matrix can be created by imposing a cutoff on the number of residue contacts rather than on the cutoff distances. We can connect one residue to its neighboring residues by increasing the cutoff distance until the limiting number of contacts is reached, regardless of the actual distance of the last connection. This enables the linking matrix to remain sparse and uniformly dense because all residues will have the same number of connections. One can see that this method creates a suitable linking matrix based on a contact number of 20 with a sparseness resembling one based on a cutoff distance of 10 Å in Fig. 2, but which no longer has any weakly connected parts. This is a smoothed, more uniform representation of protein structure.

Visualization

Animations of conformational transitions are more comprehensible than a series of static pictures, and are particularly useful for teaching (Booth, 2001). We incrementally generate 99 transient conformations between the two end conformations using the present distance interpolation method. In the implementation, we calculate $\boldsymbol{\delta}$ to minimize our cost function in Eq. 2 when $\alpha = 0.01$. Then we obtain the first intermediate conformation denoted by $\{\mathbf{x}_i^1\}$, which is 1% along the path between $\{\mathbf{x}_i\}$ and $\{\boldsymbol{\chi}_i\}$. That is,

$$\mathbf{x}_i^1 = \mathbf{x}_i + \boldsymbol{\delta}_i \quad (23)$$

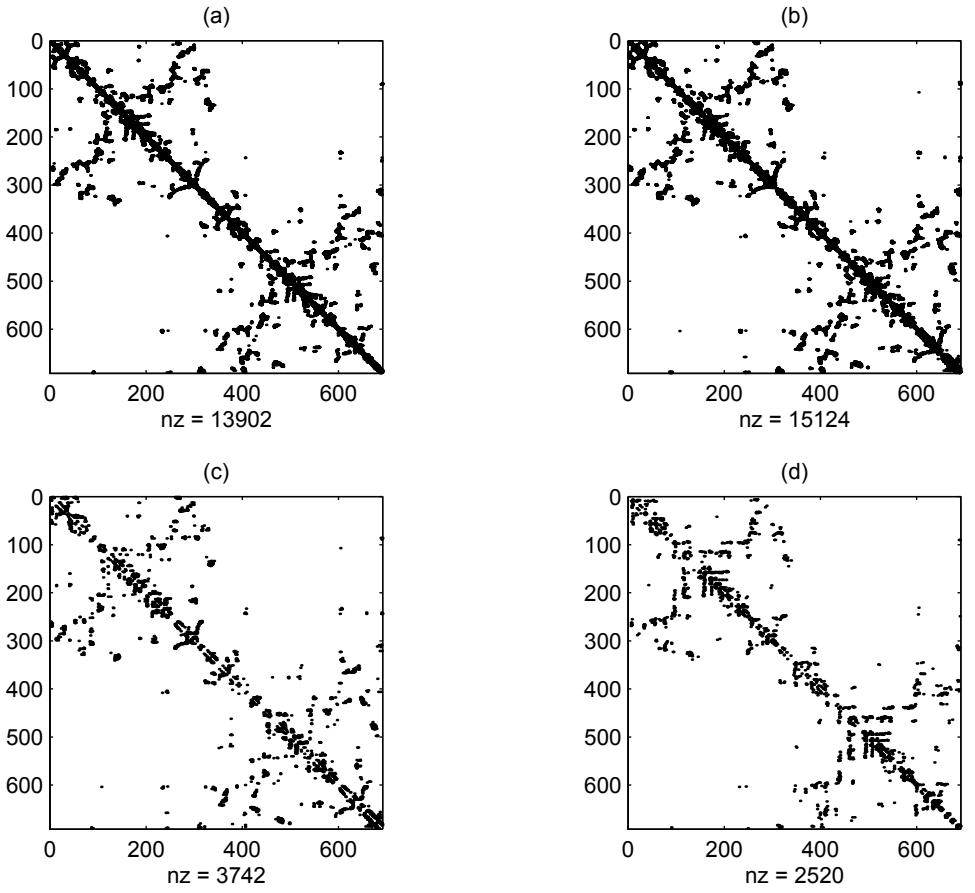

FIGURE 2 The sparsity patterns of the linking matrices. We apply two different cutoff methods to a model protein lactoferrin (1LFG and 1LFH). (*a*) The cutoff distance is 10 Å and this matrix contains 2.91% nonzero values. (*b*) The contact number of 20 is used as a cutoff regardless of neighbor distance. The density of this matrix is similar to the one based on a cutoff distance of 10 Å having a density of 3.17%, but it is more uniform and produces smoother transition pathway. (*c*) New connections are shown. (*d*) Broken connections are displayed.

where $\mathbf{x}_i^1$ is the position of the ith residue out of the set $\{\mathbf{x}_i^1\}$. Likewise for the next incremental conformation $\{\mathbf{x}_i^2\}$,

$$\mathbf{x}_i^2 = \mathbf{x}_i^1 + \boldsymbol{\delta}_i \tag{24}$$

where $\boldsymbol{\delta}$ is the solution of Eq. 21 when $\alpha = 0.02$ in the next incremental step. The remaining conformations are then obtained in this iterative way. We store them in pdb format and generate 3D pictures with Rasmol. These static pictures are used sequentially to create movies.

Our interpolation method concerns itself not with the absolute spatial positions of atoms, but instead with distances between interacting pairs. For this reason, sometimes the solved conformations starting from one conformation do not converge to the actual spatial position and orientation of the other conformation, even though the shape is sequentially interpolated quite well. We resolve this problem simply by doing a rigid-body superposition at each timestep.

SIMULATION RESULTS

To test our interpolation method, we choose several protein conformation pairs: lac repressor headpiece (1LCC and 1LCD), lactate dehydrogenase (1LDM and 6LDH), citrate synthase (4CTS and 1CTS), and lactoferrin (1LFG and 1LFH). Movies of the conformational transitions in these systems can be found at http://custer.me.jhu.edu/proteins/movies.html. Table 1 shows that the size of the protein and the density of the linking matrix are major determinants of computational time.

For small proteins it appears that a cutoff distance of 10 Å is sufficient to form a linking matrix for generating a

TABLE 1 Relationships between linking matrix density and computational efficiency for sample proteins

Transition	No. of Res.	Cutoff Type	Density*	Flops†	Time‡ (sec)
1LCC → 1LCD	51	10Å	30.8%	5.41×10^6	0.44
1LDM → 6LDH	329	10Å	5.8%	3.72×10^8	52.12
4CTS → 1CTS	437	10Å	4.3%	6.13×10^8	96.17
1LFG → 1LFH	691	10Å	2.9%	2.11×10^9	257.22
1LFG → 1LFH	691	15Å	8.6%	4.36×10^9	1386.80
1LFG → 1LFH	691	No. of Contacts = 20	3.2%	2.57×10^9	275.12

*Density is the percentage of nonzero elements in the linking matrix.
†Flops is the total number of floating point operations used in a Matlab implementation of the algorithm.
‡Time is elapsed time for calculating a single intermediate conformation on a 1.5 GHz Pentium with 512 MB memory.

FIGURE 3 Simulation of intermediate conformations between 1LDM and 6LDH of lactate dehydrogenase structures using the cutoff distance of 10 Å. A large conformational change is observed at the upper left, where a surface loop opens at the active site.

feasible pathway, but the linking matrix of lactoferrin with the same cutoff distance creates unrealistic intermediate conformations between 1LFG and 1LFH. In such a case, poorly connected parts move discontinuously in Cartesian space during the transition. One can adopt larger cutoff distances to overcome this problem. However, the larger cutoff distances increase computation times sharply. Simulation with a cutoff distance of 15 Å for lactoferrin takes about five times as long as with a cutoff distance of 10 Å. Alternatively, our uniform and sparse linking matrix, generated with a contact number cutoff of 20, enables computing feasible pathways in relatively short times. This method not only generates feasible pathways more reliably but also runs faster.

Figs. 3 and 4 illustrate the conformational transition between the corresponding pair of "open" and "closed" crystallographic structures of lactate dehydrogenase and citrate synthase, respectively. Intermediate conformations obtained using the distance interpolation method developed in this paper give rise to feasible and continuous pathways.

Our interpolation method reliably generates conformational transitions without steric obstructions for large protein pairs. Fig. 5 shows the simulation results for the conformational transition of lactoferrin, which consists of 691 residues. This simulation illustrates the movement from the closed (diferric) form (1LFG) to the open (apo) form (1LFH). Virtual bond angles and torsion angles are calculated for the intermediate conformations. Secondary structures of the protein move approximately as rigid bodies during the transition. Their virtual bond and torsion angles change little between the two end conformations. However, the dominant angle changes often occur near loops, which connect two secondary structures. In Fig. 6, dark parts of the structure represent the residues with torsion angles having the largest changes during the transition. Most of them, except Val-250, are located in loop regions. Val-250 and Thr-90 play an important role in the transition, acting as a hinge between the two subdomains at the bottom of the structure. Fig. 7 *a* shows the minimum distance between the closest pairs of α-carbons. Our interpolation method creates intermediate conformations without the severe steric clashes that occur with the simple method of interpolating over internal coordinates as shown in Fig. 1. RMS value measures the position error between corresponding α-carbons of the two conformations. Fig. 7 *b* displays RMS values of all intermediate conformations with respect to the initial conformation $\{\mathbf{x}_i\}$. They increase linearly and monotonically throughout the transition. Fig. 7 *c* shows the value of the cost function in Eq. 2. Fundamentally, the cost function is a geometric measure of how far the conformation is from the prespecified simplest path between conformations. By definition, the cost at the endpoints will be zero if the method has successfully generated a feasible pathway, and as the cost is a nonnegative quantity, it will always have a maximum located between the initial and final conformations. The density of linking matrices for all intermediate conformations is shown in Fig. 7 *d* when using a distance cutoff. The intermediate conformations are more flexible than the two end conformations in the context of an elastic protein model with a distance cutoff, whereas density is constant when using a number cutoff.

We provide as a goal for the model a set of linearly interpolated distances between every pair of connected residues in Eq. 3. This is chosen because it is the simplest way, but this does not mean that the path that is actually generated by our method corresponds to linear interpolation of distances because of the cooperativity of the coupled set of springs. Conversely, if various nonlinear trajectories in inter-residue distances are specified, the cooperativity of the system can wash out deviations from the collective behavior. To illustrate this point, we examine an example in which the set of desired distances $l_{i,j}$ are no longer linearly interpolated, but still have the same initial and end values as

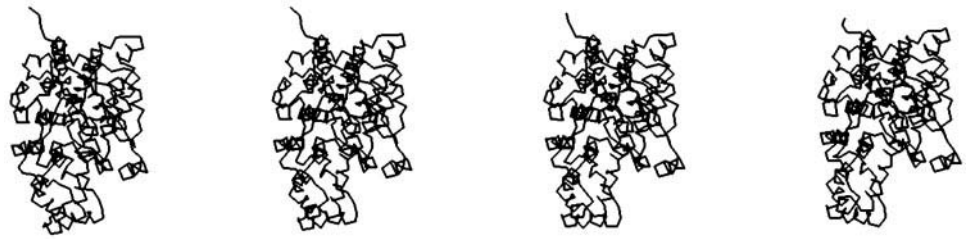

FIGURE 4 Simulation of intermediate conformations between 4CTS and 1CTS of citrate synthase structures using the cutoff distance of 10 Å. It consists of two domains with the active site between them. The small domain swings away from the large one to uncover the active site.

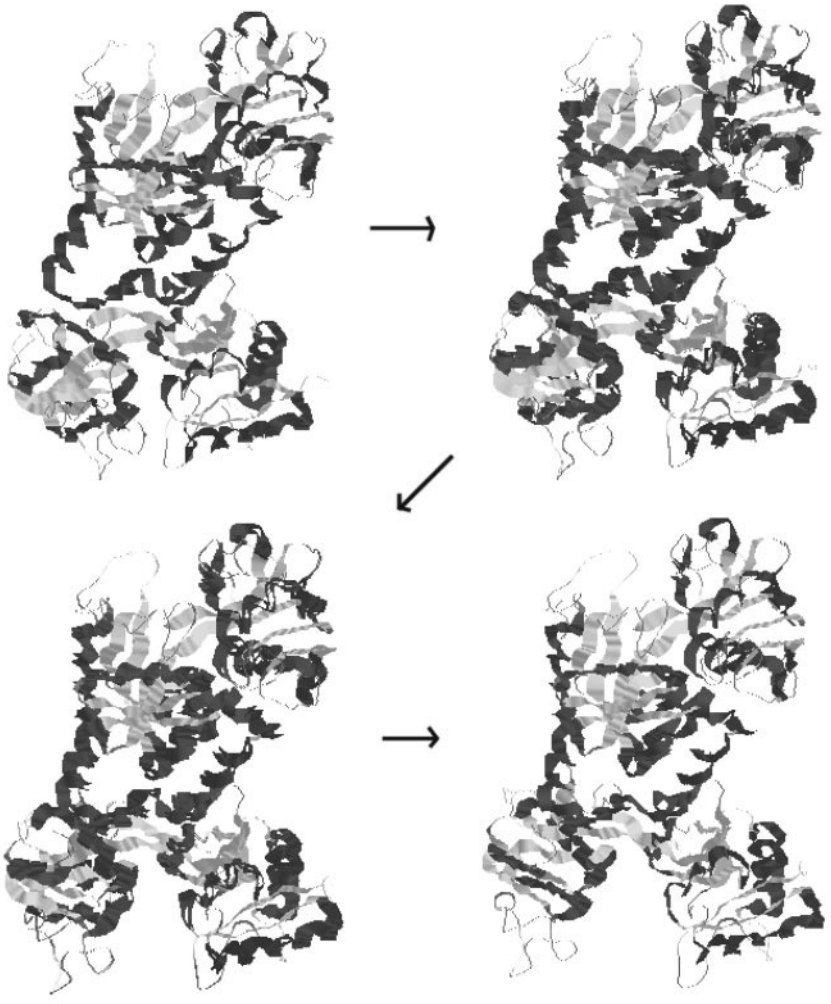

FIGURE 5 Simulation of intermediate conformations between lactoferrin forms 1LFG ("closed") and 1LFH ("open") using the contact number cutoff of 20. Here 99 intermediate conformations are obtained incrementally using our interpolation method, and 2 of these intermediate conformations (α = 0.33, 0.66) are shown. This shows the movement of lactoferrin from the "closed" form to the "open" form.

before. We generate sequences of coordinates using the new $l_{i,j}$ functions (some of which are nonlinear functions), and apply our method. The shape deviations in the resulting conformations during the animation are compared to those obtained when using linear distance interpolation by RMS position error (Fig. 8). We apply a quadratic $l_{i,j}$ to the two residue pairs having the largest distance changes during the transition with all the other $l_{i,j}$ values driven linearly, as in Eq. 3. However, the two pairs that are specified to behave in a nonlinear way end up being forced by their surrounding to behave linearly, as shown in Fig. 9. The constraints from the surroundings do not allow them to trace the "desired" input

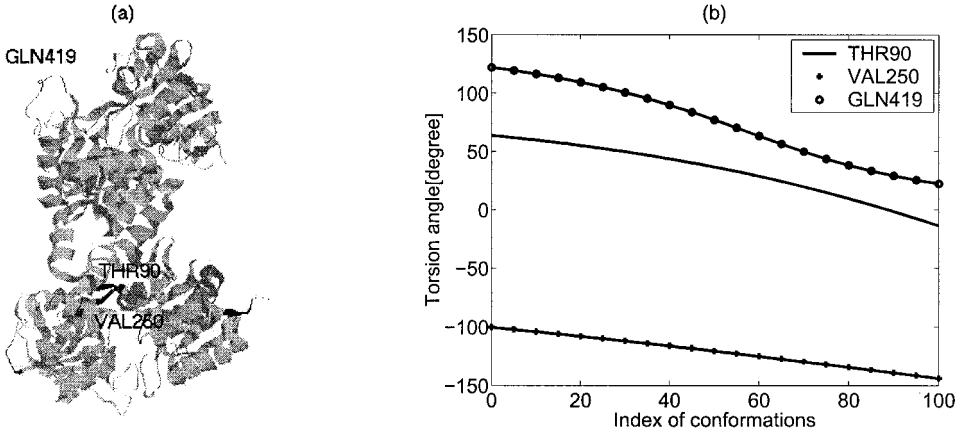

FIGURE 6 Virtual torsion angle variation during transition of lactoferrin. Bold stripes represent the residues for which torsion angles significantly vary. Especially, Thr-90 and Val-250 residues act like hinges to open two subdomains at the bottom, as shown in Fig. 5. Large angle changes between the two end conformations appear primarily in loop structures.

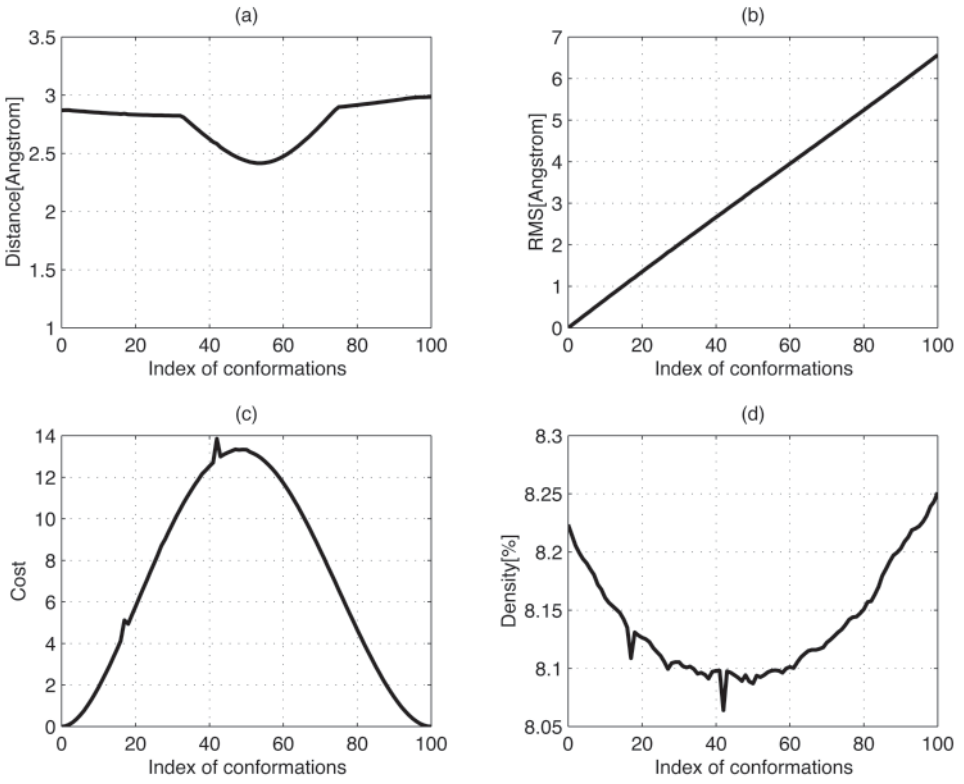

FIGURE 7 Statistics of the conformational transition in lactoferrin. (*a*) The minimum distance between all possible pairs of α-carbons in intermediate conformations shows that our method observes steric constraints. (*b*) RMS measures the difference of position between corresponding α-carbons relative to the starting form. (*c*) The value of the cost function in Eq. 1 is shown. (*d*) Linking matrices of intermediate conformation are sparser than those of the two end conformations when using a distance cutoff.

paths. Furthermore, all of the RMS values we calculated in Fig. 8 are smaller than the resolution of the experimentally determined crystal structures themselves.

From the above discussion we conclude that for all practical purposes our method generates feasible pathways that are somewhat insensitive to the user-specified choice of $l_{i,j}$. That is, the global behavior of this coarse-grained model overrides particular user-specified parameters when those parameters deviate substantially from the collective behavior. Hence, as long as the overall trend is for inter-residue distance trajectories to follow a monotonic temporal path, our linear inputs are a reasonable choice. However, if one is not interested in the simplest feasible pathway, our method can still be useful. For instance, if one is interested in generating ensembles of paths rather than a single feasible path, our method can be run many times (serially or in parallel) to generate truly different pathways by varying the $l_{ij}(t)$ relative to each other. Hence, the speed of our method has the potential to assist in the statistical mechanical analysis of protein conformational transitions, though that is not our emphasis here.

Incorporating partial conformational data

In this section we explain how our distance interpolation method can be used to incorporate incomplete conforma-

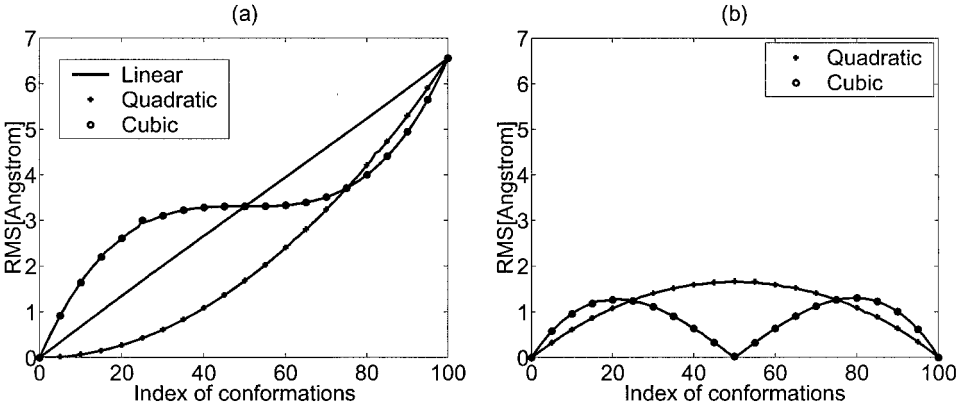

FIGURE 8 The comparison of simulation results using several different inputs in lactoferrin. (*a*) RMS values of the intermediate conformations generated by linear, quadratic, and cubic distance trajectory inputs with respect to the initial conformation $\{\mathbf{x}_i\}$ are shown, respectively. (*b*) The shape deviations of the intermediate conformations calculated using nonlinear distance trajectories as inputs compared to those obtained when using linear interpolation is shown. The magnitudes of the variations are smaller than the resolution of the crystal structure of lactoferrin.

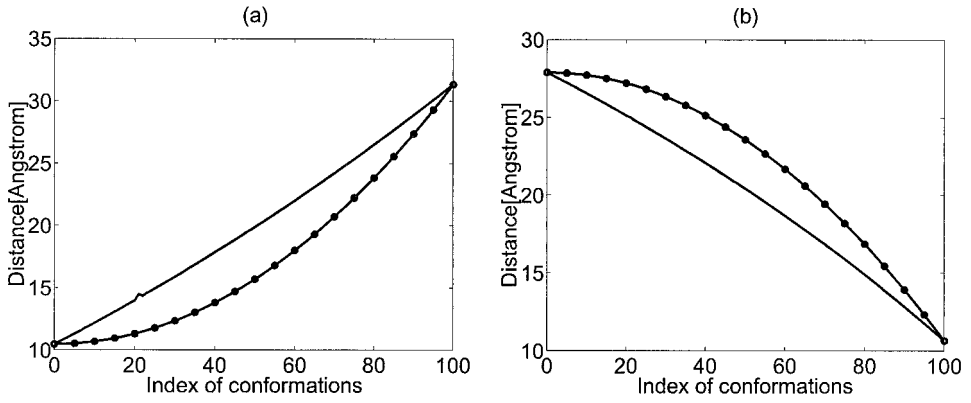

FIGURE 9 Simulation result in lactoferrin with mixed input such that most residue pairs are designated to follow the linear interpolation of distance, except two particular pairs that are driven by the quadratic input (*solid-dot line*). (*a*) The distance between Asn-13 and Gln-186 (*solid line*) appears to vary linearly, not following the input form. (*b*) Likewise, the distance between Asn-234 and Val-606 decreases linearly. This indicates that surrounding pairs force them to follow the global behavior with observation of sterics.

tional information obtained from experiments into computer simulations of protein motions. In instances when two crystallographic structures are not available, this application of our model will be of value. A number of experimental techniques are available for determining partial conformational data. One set of techniques is centered around the use of fluorescent energy transfer to capture a limited number of inter-residue distances in different conformations (Wu and Brand, 1994). Other techniques have been developed to mechanically manipulate large molecules directly and measure the history of the applied force (Bustamante et al., 2000). NMR can be used for determining time-resolved conformational data (Balbach et al., 1995; Dyson and Wright, 1996).

These experimental methods for determining conformational data generally do not provide as complete information as crystallography, but in some instances this is balanced by their ability to provide time-resolved information. Equation 2, which forms the basis for our method, applies in the context of interpolating between two crystal structures, and it also can serve as a tool for visualizing global protein motions that are consistent with time-resolved distance trajectories between two or more residues. In the context when a small subset of the $l_{i,j}$ values can be determined experimentally as functions of time, these values can be directly substituted into Eq. 2. Then our formulation proceeds as before.

We now demonstrate how this incorporation of partial conformational data into our model can be done. Consider again lactoferrin, and assume that only the open crystallographic state is given. In this case the linking matrix for this state alone (and not the union of two linking matrices) is used for $k_{i,j}$. Again, a value of one means two residues are in contact and a zero means that they are not. To test how

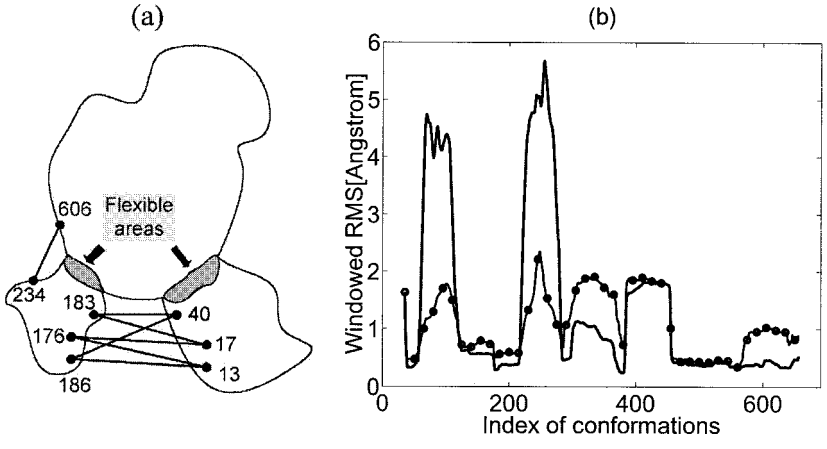

FIGURE 10 A schematic diagram of lactoferrin structure and the "windowed RMS" values of the simulation results from a single complete set of crystallographic data and limited amounts of time-resolved data. (*a*) The open lactoferrin structure is simplified as three rigid body pieces and it is assumed that seven new inter-residue distances are measured for a second conformation. The $l_{i,j}(t)$ values that describe the new conformation are enforced by direct substitution into Eq. 2 with a large value of that particular $k_{i,j}$ of 100, and the same $k_{i,j}$ value is applied to the connections within each rigid body component. (*b*) The RMS difference between the end conformation generated in this way, $\{\mathbf{x}_i^{100}\}$, and the targeted conformation, $\{\boldsymbol{\chi}_i\}$, is only 2.5 Å. The windowed RMS plots consecutively capture 70 residues per window. The solid line stands for the windowed RMS value of $\{\mathbf{x}_i\}$ relative to $\{\boldsymbol{\chi}_i\}$, while the solid-dot line indicates the windowed RMS value between $\{\mathbf{x}_i^{100}\}$ and $\{\boldsymbol{\chi}_i\}$. Our distance interpolation method with limited secondary conformational data captures the global behavior of lactoferrin's conformational transition well.

well our method would work if a second incomplete set of inter-residue distances were determined experimentally, we sample a small subset of $l_{i,j}$ values from the closed crystallized form of lactoferrin. We choose this subset of the $l_{i,j}$ values to contain at least one set of six residue-residue pairs. We have divided the whole structure into three essentially rigid pieces and assumed seven experimentally determined $l_{i,j}(t)$ values to be linear trajectories from the open to closed values in Fig. 10 *a*. Because most inter-residue distances are not specified for the second conformation, we use the simple update rule for those pairs that are not specified in the second conformation as follows

$$l_{i,j}(t + \Delta t) = \begin{cases} \|\mathbf{x}_i(t) - \mathbf{x}_j(t)\| & \text{if } \|\mathbf{x}_i(t) - \mathbf{x}_j(t)\| > l_{\min} \\ l_{\min} & \text{if } \|\mathbf{x}_i(t) - \mathbf{x}_j(t)\| \le l_{\min}, \end{cases} \tag{25}$$

where we set $l_{\min} = 3.8$ Å.

This allows any strain between intermediate and initial conformations to relax unless it results in steric clashes. The RMS difference between the end conformation generated in this way with the targeted crystallographic data is small, as shown in Fig. 10 *b*, indicating that the protein's cooperativity is captured well using our incremental distance interpolation method, and that much can be inferred about global protein motions from a single complete set of crystallographic data and limited amounts of partial data that describe a second conformation. Hence, our framework may be used as a visualization tool for experimentalists to superimpose partial conformational data onto crystal structures to examine the structural/kinematic implications of measured inter-residue distances.

CONCLUSIONS

We have developed a computationally efficient method for the realistic simulation of proteins exhibiting transitions between two crystallized conformations. Our method is also flexible and general enough to incorporate partially observed, time-dependent conformational data from experiments. It is based on a coarse-grained elastic network model. Using cutoffs in the number of nearest neighbors generates a linking matrix that is both sparse and uniformly dense, hence permitting efficient computations. Our distance interpolation method is the fastest method available for generating conformational transitions while still preserving steric constraints. This is because it involves only one inversion of a very sparse $3n \times 3n$ matrix for each frame in the animation, where n is the number of amino acid residues. Unlike dynamics-based methods in which the size of the timestep used is limited by the stiffest part of the structure, our method is purely geometric, and so the number of required animation frames is dictated only by the difference in shape between the two conformations. Typically, we choose 99 intermediate frames, whereas dynamics-based approaches must use several orders of magnitude more timesteps to achieve conformational transitions. Our method also has the benefit that it is not sensitive to parameter choices, solvation, or quantum mechanical effects, which is not the case with the most physically detailed models. While physical reality is the ultimate goal of all computational modeling methods, we believe that having a method effective for proteins of several hundred residues with an accessible PC in a few hours may be preferable to more "realistic" techniques requiring months of high-performance computer time, and are not guaranteed to converge due to round-off errors, instability of numerical integration, or even a lack of full knowledge about the true nature of the chemical potentials involved in proteins.

Simulation results illustrate that the distance interpolation method presented here reliably generates sequences of feasible intermediate conformations of proteins without steric clashes. Animations produced using this method are posted at http://custer.me.jhu.edu/proteins/movies.html. The distance interpolation method represents an improvement over simplified linear position interpolations in terms of the realism of intermediate forms, and over all-atom computational methods such as MD and NMA, in terms of computational efficiency.

REFERENCES

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

Bahar, I., A. R. Atilgan, and B. Erman. 1997. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Fold. Des.* 2:173–181.

Bahar, I., B. Erman, R. L. Jernigan, A. R. Atilgan, and D. G. Covell. 1999. Examination of collective motions in HIV-1 reverse transcriptase. Examination of flexibility and enzyme function. *J. Mol. Biol.* 285: 1023–1037.

Bahar, I., and R. L. Jernigan. 1998. Vibrational dynamics of transfer RNAs: comparison of the free and synthetase bound forms. *J. Mol. Biol.* 281:871–885.

Balbach, J., V. Forge, N. A. J. Vannuland, S. L. Winder, P. J. Hore, and C. M. Dobson. 1995. Following protein-folding in real-time using NMR-spectroscopy. *Nat. Struct. Biol.* 2:865–870.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

Booth, A. G. 2001. Visualizing protein conformational changes on a personal computer-alpha carbon pseudo bonding as a constraint for interpolation in internal coordinates space. *J. Mol. Graph. Model.* 19: 481–486.

Brooks, B., and M. Karplus. 1983. Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic typsin inhibitor. *Proc. Natl. Acad. Sci. USA.* 80:6571–6575.

Brooks, B., and M. Karplus. 1985. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA.* 82:4995–4999.

Bustamante, C., J. C. Macosko, and G. J. L. Wuite. 2000. Grabbing the cat by the tail: manipulating molecules one by one. *Nat. Rev. Mol. Cell. Biol.* 1:130–136.

Crippen, G. M., and T. F. Havel. 1988. Distance geometry and molecular conformation. John Wiley and Sons, New York.

Czerminski, R., and R. Elber. 1990. Reaction path study of conformational transitions in flexible systems: applications to peptides. *J. Phys. Phys.* 92:5580–5601.

Dellago, C., P. G. Bolhuis, F. S. Csajka, and D. Chandler. 1997. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* 108:1964–1977.

Dyson, H. J., and P. E. Wright. 1996. Insights into protein folding from NMR. *Annu. Rev. Phys. Chem.* 47:369–395.

Elber, R., and M. Karplus. 1987. A method for determining reaction paths in large molecules: application to myoglobins. *Chem. Phys. Lett.* 139: 375–380.

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

Huo, S., and J. E. Straub. 1997. The MaxFlux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature. *J. Chem. Phys.* 107: 5000–5006.

Jaaskelainen, S., C. S. Verma, and R. E. Hubbard. 1998. Conformational change in the activation of lipase: an analysis in terms of low-frequency normal modes. *Protein Sci.* 7:1359–1367.

Jónsson, H., G. Mills, and K. W. Jacobsen. 1998. Nudged elastic band method for finding minimum energy paths of transitions. *In* Classical and Quantum Dynamics in Condensed Phase Simulations. B. J. Berne, G. Ciccotti, and D. F. Coker, editors. World Scientific, Singapore. 385–404.

Kleywegt, G. J., and T. A. Jones. 1995. Where freedom is given, liberties are taken. *Structure.* 3:535–540.

Kleywegt, G. J., and T. A. Jones. 1996. Phi/Psi-chology: Ramachandran revisited. *Structure.* 4:1395–1400.

Olender, R., and R. Elber. 1996. Calculation of classical trajectories with a very large time step: formalism and numerical examples. *J. Chem. Phys.* 105:9299–9315.

Subbiah, S. 1996. Protein Motions. R. G. Landes Company, Austin, Texas.

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

Tirion, M. M., and D. Ben-Avraham. 1993. Normal mode analysis of G-actin. *J. Mol. Biol.* 230:186–195.

Tirion, M. M., and D. Ben-Avraham. 1998. Normal modes analyses of macromolecules. *Physica A.* 249:415–423.

Vonrhein, C., G. J. Schlauderer, and G. E. Schulz. 1995. Movie of the structural changes during a catalytic cycle of nucleoside monophosphate kinases. *Structure.* 3:483–490.

Wu, P., and L. Brand. 1994. Resonance energy transfer: methods and applications. *Anal. Biochem.* 218:1–13.

Xu, Z., A. L. Horwich, and P. B. Sigler. 1997. The crystal structure of the asymmetric GroEL-GroES-$(ADP)_7$ chaperonin complex. *Nature.* 388: 741–750.

Xu, Z., and P. B. Sigler. 1998. GroEL/GroES: structure and function of a two-stroke folding machine. *J. Struct. Biol.* 124:129–141.