Assessing protein loop flexibility by hierarchical Monte Carlo sampling

Jerome Nilmeier ,[1†] Lan Hua,[1†] Evangelos A. Coutsias,[2] Matthew P. Jacobson[1*]

[1] *Department of Pharmaceutical Chemistry, University of California in San Francisco, San Francisco, California 94158-2517*

[2] *Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico 87131*

March 8, 2011

Abstract

Loop flexibility is often crucial to protein biological function in solution. We report a new Monte Carlo method for generating conformational ensembles for protein loops and cyclic peptides. The approach incorporates the triaxial loop closure method which addresses the inverse kinematic problem for generating backbone move sets that do not break the loop. Sidechains are sampled together with the backbone in a hierarchical way, making it possible to make large moves that cross energy barriers. As an initial application, we apply the method to the flexible loop in triosephosphate isomerase that caps the active site, and demonstrate that the resulting loop ensembles agree well with key observations from previous structural studies. We also demonstrate, with 3 other test cases, the ability to distinguish relatively flexible and rigid loops within the same protein.

*Correspondence to Matthew P. Jacobson, Department of Pharmaceutical Chemistry, University of California in San Francisco, MC 2240, California 94158-2517, Tel: (415) 514-9811, Email: Matt.Jacobson@ucsf.edu.

† Jerome Nilmeier and Lan Hua contributed equally to this work.

1 Introduction

A great deal of effort has been directed towards the development of computational methods for predicting the conformations of protein loops, which is a critical task in comparative protein modeling and in computational protein design [1–4]. The success of these methods has been evaluated primarily by comparing the results of the loop predictions with the loop conformations observed in crystal structures. That is, the focus is predicting the structure of the loop—a specific conformation—rather than the ensemble of conformations populated at biologically relevant conditions. Although these loop prediction methods can be used to identify multiple low-energy conformations, it is challenging to determine populations of the conformations, i.e., to relate energies of individual conformations to free energies of micro or macrostates in the ensemble, although significant progress in this regard has been made by Meirovitch and coworkers[5–7].

The flexibility of loops, i.e., the ability to adopt multiple conformations at relevant temperatures, is often critical to biological function, by playing an important role in molecular recognition. For example, the active site loop of the triosephosphate isomerase (TIM barrel) changes its conformation from an open to a closed state after binding of ligands [8,9]. In kinases, two critical loops near the active site are flexible, with important implications for drug discovery: the glycine-rich loop (also called the P-loop) and the activation loop, including the DFG motif, which can adopt at least 2 major conformations in some kinases, referred to as 'out' and 'in'. For example, while c-Src generally adopts the DFG-in conformation, the unfavorable DFG-out conformation can be induced by binding small molecules [10]. Loop flexibility can also play an important role in antibody-antigen recognition. The H3 loop in the complementarity-determining region of antibodies, which has the most diversity in sequence and is the most critical loop for antigen affinity and specificity, frequently demonstrates evidence of conformational flexibility [11–13].

More broadly, there are many cases where loops adopt different conformations in different crystal structures, e.g., holo vs. apo, or even different crystal unit cells for the same protein [14]. Although the B-factors in crystal structures provide some information about conformational flexibility, each structure is best viewed as a snapshot from the equilibrium ensemble. NMR experiments can provide some direct information about conformational equilibria, but generally cannot provide complete information about the ensemble of interconverting structures.

Molecular dynamics (MD) has been widely used to study protein flexibility, including loop dynamics [15,16]. The main liability of MD is that the timescales for interconverting between loop conformations can be long relative to the femtosecond time steps used, such

as the millisecond timescale for the TIM capping loop to interconvert between the open and closed states [17]. Although such timescales may soon become accessible by MD simulation, they will remain extremely computationally expensive. Methods like replica exchange MD can be used to accelerate convergence but are likewise computationally expensive.

Here we describe a Monte Carlo method for generating ensembles of loop conformations and cyclic peptides. It is related to classes of loop prediction methods that use torsion-angle sampling of backbone and side chain degrees of freedom (DoF), which makes it possible to make large conformational moves that cross energy barriers. Specifically, it builds on loop prediction methods that exploit "inverse kinematics" methods for creating move sets that do not 'break' the loop [18–24]. The new contribution here is implementing these moves in a Monte Carlo scheme that also samples side chain DoF [25]. We apply the method to a number of proteins with flexible loops, including the well-known case of TIM. We also evaluate our ability to distinguish between (relatively) rigid and flexible loops within the same protein.

2 The Move Set: Torsional Perturbations via Inverse Kinematics

2.1 Torsions and Sterics

It is widely accepted that the essential dynamics of a protein backbone can be captured by moves involving only the torsions ϕ, ψ with the other internal variables (bond lengths, bond angles and ω torsions) being kept close to their canonical values, although not necessarily rigid [19,23,24].

Compared to the high energy associated with ω angle deformation, ϕ and ψ angles are relatively free to rotate but their range is restricted by steric interactions. Ramachandran regions in the (ϕ, ψ) coordinates for each peptide ensure intra-peptide steric avoidance, and additional restrictions are imposed by more distant clashes. Clashes involving backbone atoms (or atoms bonded to them) are completely determined from the backbone angles. On the other hand, atoms further along sidechains (from the γ position out) are not completely determined from the backbone, although their placement may be restricted by it. Significantly, sidechains may interact with other sidechains so that their placement must be accomplished as a whole. Given a backbone conformation, a separate search is required to determine sterically acceptable or otherwise energetically viable sidechain conformations. Reciprocally, backbone moves may be restricted by fixed sidechain geometry.

2.2 MC move and state variables

To design a Monte Carlo move for reversibly exploring the torsion space, we must therefore consider the state space as the set of all torsions, $\{t_i; \chi_j\}$ where the t_i are backbone torsions and χ_j are sidechain torsions, with the indices running respectively over all the backbone and sidechain *DoF*. A chain of $\{N, C\alpha, C\}$ triplets (a standard backbone) is one possibility but chains through e.g. cysteine bridges, or other macromolecules, such as nucleic acids, could also be considered. In the following, we will assume the standard case (protein backbone loops) exclusively. For the case of a loop of N residues bridging two fixed ends the essential backbone *DoF* would be $M = 2N - 6$. Here, 6 backbone *DoF* are involved in placing the end of the loop in a fixed rotation/translation relationship to the beginning. We call these *DoF*, labeled arbitrarily as $t_i, (i = 1, \ldots, 6)$ the compensators. The remaining M *DoF*, labeled as $t_i, (i = 7, \ldots, 2N)$ are the controls. This separation in controls and compensators is arbitrary and may change from one move to the next. We could assume that the end residues 0 and $N + 1$ act as hinges, i.e. the ϕ_0 and ψ_{N+1} torsions are fixed, but ψ_0, ϕ_{N+1} are free, adding two *DoF* to the backbone. The treatment is essentially the same, replacing M by $M + 2$ and redefining some indices. We will only discuss the first case (no hinge mobility). It will be assumed that there are K sidechain *DoF* in the set $\mathcal{S}$ of sidechains interacting with the loop; we may only wish to include in $\mathcal{S}$ those sidechains on the loop and hinges. The placement for those depends on the loop conformation. We may also include sidechains on residues in some sphere of influence about the loop. Or we may simply include all the sidechains in the protein. We make no distinction at this stage.

Then, to design a reversible MC move that involves only the loop backbone *DoF* as well as the selected group $\mathcal{S}$ of sidechains coupled to the loop we must establish the Metropolis criterion for acceptance of a move of the form

$$\{t_i, \chi_j\} \rightarrow \{t_i + \delta t_i \ ; \ \chi_j + \delta\chi_j\} \ , \ i \in [7, 2N] \ , \ j \in [1, K] \tag{1}$$

The shape space geometry accessible via our formulation characterizes our moves: assume that the L $(= 2N)$ torsions for a loop kinematic chain are divided into the $L - 6$ controls and 6 compensators. The method used here employs the ϕ, ψ pairs of 3 amino acids (the *pivots*). These can be chosen at arbitrary locations along the loop, breaking it into three subfragments for kinematic purposes. To each value of the $L - 6$ controls there correspond up to 16 distinct conformations satisfying the closure conditions, each characterized by a unique set of values of the compensators. As discussed in our earlier work [26], the 16 alternative solutions represent different orientations of the 3 subfragments between successive pivots in a reference frame attached to the 3 pivot $C\alpha$ atoms about the

3 axes joining each pair of pivots. Thus we refer to the method as Triaxial Loop Closure (**TLC**). The basic idea in the TLC method (discussed more in detail in the next section) is to construct a loop with arbitrary internal degrees of freedom, taking advantage of the fact that the inverse kinematic problem can be solved by determining appropriate values of six torsions. Thus any variation in the remaining DoF's - other torsions, including omegas, bond angles and even bond lengths - can be considered, if so desired. Here we treated only $\phi - \psi$ variations as these are the most "flexible" DoF's, but we could have included all other DoF's in the MC scheme in any combination desired. The conformational variability of the constitutive pieces for loop closure, i.e. the three subfragments, is of course an important factor for solving the closure problem. We see that this variability can be decomposed in two types: the end-to-end variability of the individual fragments, and the inherent variability of the loop closure problem, i.e. relative locations and orientations of the ends of the loop as well as the environment in the loop vicinity.

The first is a direct problem: compute the fragment (in practice we do not check that the fragment is indeed sterically feasible until the assembly is successful). The individual fragment assembly, being subject to no end constraints, is only limited by the Ramachandran and other steric restrictions. However, for purposes of assembling the three subfragments into a self consistent loop, each individual fragment of length L_i residues with $i = 1, 2, 3$, is encoded by 4 variables: the overall geometric length of the virtual bond joining first and last atoms, d_i; the angles θ_i, ξ_i made by the two end bonds to the virtual bond; and the torsion of the two end bonds about the virtual bond, δ_i. The variability of the closure problem is governed by these twelve parameters $(d_i, \theta_i, \xi_i, \delta_i)$, $i = 1, 2, 3$. The equations expressing closure depend on these parameters smoothly; small changes cause usually small changes in the number and disposition of solutions except that, for certain arrangements, solutions could spontaneously appear or disappear (pairs of polynomial roots may join and become complex, or the converse, see the discussion of the Inverse Kinematic problem below).

We now search the nearby conformation space by perturbing one of the control torsions. This will result in perturbing the overall structure of one of the chains, leading to a perturbed set of solutions. These changes may lead to overall large motions, see e.g. Ref.[27] for a discussion of the end conditions and their constraining of various inner *DoF*. However, a reasonable acceptance ratio for the method can be more or less guaranteed by varying the controls and restricting the stepsize. Below we discuss a two stage scheme, splitting the move into a pure backbone and a pure sidechain stage.

2.3 Solving the Inverse Kinematic problem

Many methods for finding solutions that satisfy the closure conditions have been proposed, both exact [18,22,26,28–32] and approximate [6,21,33–37]. Exact methods address the Inverse Kinematic problem by searching for the values of a certain torsion, say τ, in terms of which all other torsions can be determined. Go and Scheraga [18] pursued a direct solution in the original angle variables. This involves finding the zeros of a certain transcendental expression, a process that may require substantial computation to adequately resolve the entire domain. Subsequent works employ standard techniques from the robotics literature to convert to a more tractable polynomial form in the variable $u = \tan \tau/2$. All the real roots of this 16th degree polynomial can be found efficiently and stably by the use of the method of Sturm chains [38]. All other torsions can be recovered readily and therefore such methods are capable of finding all backbone solutions for any given combination of control torsion values. On the other hand, approximate methods typically use an iterative procedure to find a solution. As a result they are not guaranteed to find all solutions consistent with a given set of control values, and the same is true for the approach in Ref. [18], which is also followed in Refs. [20,23,24], although for this class of methods the issues are mainly related to the computational sensitivity of multiple roots.

In previous applications the *conrot* algorithm has been used [20]. It places the rotatable bonds on 6 consecutive bonds plus a driver. A generalization by Wu and Deem [22] uses one driver on either end. A weakness of the *conrot* approach is that a change on either side of the short compensator segment may make the closure problem unsolvable [24]. A generalization from robotics removes that restriction [29]. Our own method for solving the tripeptide closure problem, explained in detail in Ref. [26], has the advantage of mathematical simplicity, speed and robustness. It also allows for a straightforward generalization for longer chains of arbitrary geometry. Its simplicity comes from taking advantage of the natural pairing up of rotatable bonds in amino acids to reduce the closure problem to three rotations, and we refer to this as the TLC method [26]. Referring to Fig. 1(b) we note that each $C\alpha, C, N, C\alpha$ unit is identified by four variables: the overall geometric length of the virtual bond joining first and last atoms, d_i; the angles θ_i, ξ_i made by the two end bonds to the virtual bond; and the torsion of the two end bonds about the virtual bond, δ_i (actually, the formulation uses the angles α_i of the triangle formed with edges d_i). These definitions remain unchanged even if arbitrary structure exists between the two end pairs (Fig. 1(a)). We may produce multiple conformations for a long closed chain by partitioning into 3 subsegments and mapping each to a simple kinematic generalization of the tetrad $C\alpha, C, N, C\alpha$ (Fig. 1(a)(b)).

In brief, three $C\alpha$ atoms are selected (the pivots). The chain between any two of these, containing L atoms including the endpoints is determined to within a rotation/translation (i.e. in its own body frame) by its own internal coordinates: $L-3$ torsions, $L-2$ angles, $L-1$ lengths. With fixed (to any prescribed value) bond-lengths and bond-angles, each chain can be completely described by its $L-3$ internal torsions. Below, we will index the residues of the three pivots as 1, 2 and 3 and we will index their backbone atoms as N_i, $C\alpha_i$ and C_i, $i=1,2,3$ accordingly. Below we use the atom names interchangeably with their cartesian coordinates, e.g. N_1 can be thought of as equivalent to the vector $\mathbf{R}_1$, etc (see eq. 5).

As is explained in Ref. [26] and somewhat more at length in Ref. [39] (see also the supplementary material discussion in Ref. [40]), the three fragments, respectively between pivots 1-2, 2-3 and 3-1, form a triangle with edges $d_i, i=1,2,3$. The parameters necessary for setting up and solving the TLC equations can be extracted from knowledge of only the first two and last two atoms of each chain (Fig. 2). Once the three 4-atom fragments have been assembled into a triangle, the relative rotation of each fragment about the triangle must place the end-atoms relative to those on each neighboring fragment so that the angles ($N_iC\alpha_iC_i$, $i=1,2,3$ assume prescribed values (Fig. 1). In this way, loop closure is accomplished when an appropriate rotation for each piece has been found. It turns out that the problem overlays the solution of a 16th degree polynomial, so that to each real root there corresponds a possible backbone loop geometry (subject, of course to overall steric viability) to a total of, at most, 16 solutions possible for a given collection of state variables, the control $2N-6$ torsions.

2.4 Jacobian

Since fixing the end of the chain (the *Closure Conditions*) implies relationships among the torsions, we seek solution of these relationships such that specifying M torsions along the loop leads to complete determination of all $2N$ torsions and unambiguous Cartesian coordinates for all loop backbone atoms that are sterically self-consistent. In general, for any feasible value of the controls there may exist multiple sets of compensators that allow the loop to close. They are functions of the controls and their values solve the loop closure problem.

As a result, the element of volume in torsion space, initially uniform in these variables

$$d\mathcal{V} = d\phi_1 d\psi_1 \ \ldots \ d\phi_N d\psi_N d\chi_1 \ \ldots \ d\chi_K$$

will need to be modified by

$$dt_1 \cdots dt_6 = \frac{\partial(t_1,\ldots,t_6)}{\partial(\mathbf{R}_6,\mathbf{\Gamma}_6,t_6)} d\mathbf{R}_6 d\mathbf{\Gamma}_6 dt_6$$

leading to the well known expression (e.g. see formula 23 in [23]) for the inverse of the above Jacobian:

$$J_i = \det \frac{\partial\left(\mathbf{R}_6, \mathbf{\Gamma}_6, t_6\right)}{\partial \mathbf{t}} = \left| \begin{array}{cccccc} \dfrac{\partial \mathbf{R}_6}{\partial t_1} & \dfrac{\partial \mathbf{R}_6}{\partial t_2} & \dfrac{\partial \mathbf{R}_6}{\partial t_3} & \dfrac{\partial \mathbf{R}_6}{\partial t_4} & \dfrac{\partial \mathbf{R}_6}{\partial t_5} & \dfrac{\partial \mathbf{R}_6}{\partial t_6} \\ \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_1} \cdot \mathbf{e}_1 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_2} \cdot \mathbf{e}_1 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_3} \cdot \mathbf{e}_1 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_4} \cdot \mathbf{e}_1 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_5} \cdot \mathbf{e}_1 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_6} \cdot \mathbf{e}_1 \\ \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_1} \cdot \mathbf{e}_2 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_2} \cdot \mathbf{e}_2 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_3} \cdot \mathbf{e}_2 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_4} \cdot \mathbf{e}_2 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_5} \cdot \mathbf{e}_2 & \dfrac{\partial \mathbf{\Gamma}_6}{\partial t_6} \cdot \mathbf{e}_2 \\ \dfrac{\partial t_6}{\partial t_1} & \dfrac{\partial t_6}{\partial t_2} & \dfrac{\partial t_6}{\partial t_3} & \dfrac{\partial t_6}{\partial t_4} & \dfrac{\partial t_6}{\partial t_5} & \dfrac{\partial t_6}{\partial t_6} \end{array} \right| .$$

Since

$$\frac{\partial \mathbf{R}_k}{\partial t_j} = \mathbf{\Gamma}_j \times \mathbf{R}_{jk} \ , \ \frac{\partial \mathbf{\Gamma}_6}{\partial t_j} = \mathbf{\Gamma}_j \times \mathbf{\Gamma}_6 \ , \ \frac{\partial t_i}{\partial t_j} = \delta_{ij} \tag{2}$$

this Jacobian can assume the simpler, 5×5 form

$$\begin{aligned} J_i \ &:= \ \mathbf{J}(\mathbf{R}_6, \mathbf{\Gamma}_6, t_6; t_1, \cdots, t_6) \qquad (3) \\ &= \left| \begin{array}{ccccc} \mathbf{\Gamma}_1 \times \mathbf{R}_{16} & \mathbf{\Gamma}_2 \times \mathbf{R}_{26} & \mathbf{\Gamma}_3 \times \mathbf{R}_{36} & \mathbf{\Gamma}_4 \times \mathbf{R}_{46} & \mathbf{0} \\ (\mathbf{\Gamma}_1 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 & (\mathbf{\Gamma}_2 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 & (\mathbf{\Gamma}_3 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 & (\mathbf{\Gamma}_4 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 & (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 \\ (\mathbf{\Gamma}_1 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 & (\mathbf{\Gamma}_2 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 & (\mathbf{\Gamma}_3 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 & (\mathbf{\Gamma}_4 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 & (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 \end{array} \right| \end{aligned}$$

Here

$$\mathbf{R}_{ij} = \mathbf{R}_j - \mathbf{R}_i \ , \ \mathbf{\Gamma}_i = \frac{\mathbf{R}'_i - \mathbf{R}_i}{||\mathbf{R}'_i - \mathbf{R}_i||} \tag{4}$$

and $\mathbf{e}_i \ , \ i = 1, 2, 3$ are the usual unit vectors along axes x, y, z of an arbitrary reference frame (the *Lab frame*). The atoms associated with closure are

$$\mathbf{R}_{2k-1} = N_k \ , \ \mathbf{R}_{2k} = C\alpha_k (= \mathbf{R}'_{2k-1}) \ , \ \mathbf{R}'_{2k} = C_k \ ; \ k = 1, 2, 3 \tag{5}$$

We note that the term $\mathbf{\Gamma}_5 \times \mathbf{R}_{56} = 0$ and was omitted. In the general case, the three pivot residues are indexed by $1 \leq n_1 < n_2 < n_3 \leq N$, and this reindexing will be implied where appropriate.

It is well known [22] that the Jacobian in the form first proposed by Dodd et al.[20] is incomplete, and lacks frame invariance. In a rigorous derivation of the Jacobian from the configuration integral, Wu and Deem [22] show that the correct, frame invariant form is

$$J^{-1} = \frac{1}{\mathbf{\Gamma}_6 \cdot \mathbf{e}_3} J_i \tag{6}$$

However, since the acceptance criterion involves ratios of Jacobians computed at the same frame, the additional factors cancel and the relative probabilities remain unchanged.

Although the latter form (6) is indeed invariant if all vectors are changed by an arbitrary affine transformation, it has the undesirable feature that it involves a projection to an arbitrary frame. Consequently, the factor $\mathbf{\Gamma}_6 \cdot \mathbf{e}_3$ may accidentally vanish (in which case J_i will also vanish) necessitating a random reorientation of the frame to break the degeneracy. Thus it is desirable to eliminate this superfluous dependence and derive a form that depends only on intrinsic (body frame) coordinates, for which invariance is easily seen. This can be accomplished by carrying out an expansion of this determinant in complementary minors; indeed, the top 3 rows are expressed in terms of intrinsic coordinates, while the last two involve projections to the space frame. We thus expand the determinant as

$$J_i = \sum_{i=1}^{4} (-1)^i \left| \begin{array}{cc} (\mathbf{\Gamma}_i \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 & (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 \\ (\mathbf{\Gamma}_i \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 & (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 \end{array} \right| \left| \begin{array}{ccc} \mathbf{\Gamma}_j \times \mathbf{R}_{j6} & \mathbf{\Gamma}_k \times \mathbf{R}_{k6} & \mathbf{\Gamma}_l \times \mathbf{R}_{l6} \end{array} \right| \tag{7}$$

where the indices (i, j, k, l) are a cyclic permutation of $(1, 2, 3, 4)$.

Applying the well known identity (e.g., [41], Eq.(25), p.76)

$$\left| \begin{array}{cc} \mathbf{A} \cdot \mathbf{C} & \mathbf{B} \cdot \mathbf{C} \\ \mathbf{A} \cdot \mathbf{D} & \mathbf{B} \cdot \mathbf{D} \end{array} \right| = \mathbf{A} \cdot \mathbf{C} \;\; \mathbf{B} \cdot \mathbf{D} - \mathbf{B} \cdot \mathbf{C} \;\; \mathbf{A} \cdot \mathbf{D} = (\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) \tag{8}$$

to the first of the 2×2 minors in Eq.(7) we have:

$$\left| \begin{array}{cc} (\mathbf{\Gamma}_1 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 & (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_1 \\ (\mathbf{\Gamma}_1 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 & (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_2 \end{array} \right| = (\mathbf{\Gamma}_1 \times \mathbf{\Gamma}_6) \times (\mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \cdot \mathbf{e}_3 = (\mathbf{\Gamma}_1 \cdot \mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \, (\mathbf{\Gamma}_6 \cdot \mathbf{e}_3) \tag{9}$$

The remaining 2×2 minors result in analogous expressions. Substituting these into Eq.(7) we have

$$\frac{J_i}{\mathbf{\Gamma}_6 \cdot \mathbf{e}_3} = \sum_{i=1}^{4} (-1)^i \, (\mathbf{\Gamma}_i \cdot \mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \left| \begin{array}{ccc} \mathbf{\Gamma}_j \times \mathbf{R}_{j6} & \mathbf{\Gamma}_k \times \mathbf{R}_{k6} & \mathbf{\Gamma}_l \times \mathbf{R}_{l6} \end{array} \right| \tag{10}$$

(as above, the indices (i, j, k, l) are a cyclic permutation of $(1, 2, 3, 4)$), which can be recombined to give the expression for the inverse Jacobian

$$J^{-1} = \frac{1}{\mathbf{\Gamma}_6 \cdot \mathbf{e}_3} J_i = \left| \begin{array}{cccc} \mathbf{\Gamma}_1 \times \mathbf{R}_{26} & \mathbf{\Gamma}_2 \times \mathbf{R}_{26} & \mathbf{\Gamma}_3 \times \mathbf{R}_{46} & \mathbf{\Gamma}_4 \times \mathbf{R}_{46} \\ (\mathbf{\Gamma}_1 \cdot \mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) & (\mathbf{\Gamma}_2 \cdot \mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) & (\mathbf{\Gamma}_3 \cdot \mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) & (\mathbf{\Gamma}_4 \cdot \mathbf{\Gamma}_5 \times \mathbf{\Gamma}_6) \end{array} \right| \tag{11}$$

where we took advantage of the fact that $\mathbf{\Gamma}_i \times \mathbf{R}_{i6} = \mathbf{\Gamma}_i \times \mathbf{R}_{i+1,6}$ with $i = 1, 3$ due to the fact that the axes $\mathbf{\Gamma}_i, \mathbf{\Gamma}_{i+1}$, $i = 1$ or 3 are coterminal. Fig.(1(a)) shows all quantities that enter in the Jacobian.

This 4×4 determinant is the frame invariant form of the inverse Jacobian for the TLC method. It has the advantage that it is expressed entirely in terms of body coordinates, and thus it is free from degeneracies and can be evaluated without projecting to an ad hoc coordinate system. It is numerically equivalent to the Wu and Deem form (6), when the latter is defined. The Jacobian (11) can be easily expressed in terms of the intrinsic parameters $(d_i, \theta_i, \xi_i, \delta_i), i = 1, 2, 3$ entering in the TLC algorithm [42], a feature that it shares with reduced Jacobians derived by other authors [22,43]. However such expressions lack the simplicity and geometrical appeal of (11).

2.5 Backbone Perturbation Procedure

The loop closure algorithm described in the previous section, while perfectly general, is currently implemented as a strategy for perturbing only the backbone coordinates. The sidechain coordinates perturbation procedure, as well as the strategy for combining these perturbations in a way such that detailed balance is maintained, will be outlined in the two sections. An important design feature of this approach is that the backbone and sidechain perturbations are generated independently.

An important feature of both the backbone selection probability and the sidechain selection probability is that they are reversible, or

$$\alpha(\mathbf{t} \rightarrow \mathbf{t}') = \alpha(\mathbf{t}' \rightarrow \mathbf{t}) \tag{12}$$

where $\mathbf{t}' = \mathbf{t} + \delta\mathbf{t}$ is the trial move starting from the torsion state $\mathbf{t}$ and $\delta\mathbf{t}$ is the perturbation vector to the loop of interest. For the purposes of this work, we require the selection probability to be uniform to enforce equation 12. For this to be true, we need to establish the procedure which ensures that a uniform distribution of torsions over the entire loop can be generated.

The procedure for generating a trial move $\delta\mathbf{t}$ closely follows that of Ref. [20,22,23,29]. Since the algorithm currently solves for $2N - 6$ torsions, and we wish to have a procedure that is valid for loops of arbitrary length, we must select a subset of $2N - 6$ torsions. There is some flexibility in how this could be done, but the present implementation is as follows (see Fig. 2):

1) From the designated loop torsions, a single torsion angle i is selected uniformly and identified as a driver angle coordinate (the yellow circle in Fig. 2), as has been described in previous work [26].

2) For torsion t_i, a random variate δt_i is generated, with a maximum value of up to π.

3) A randomly constructed triaxial closure is generated by randomly selecting 3 α carbons as pivots from the loop (excluding the α carbon on which the driver angle resides), and assigning the ϕ/ψ angles as the torsions (the grey triangle in Fig. 2).

4) A set of torsions for the stationary solution $\mathbf{t}_k, k \in [1, K]$ is generated, resulting in up to $K = 16$ solutions. For this case, only the alternative sets of pivot coordinates are considered, with the driver angle held at t_i. For each solution, a Jacobian term $J(\mathbf{t}_k)$ is computed.

5) A set of torsions for the perturbed solution $\mathbf{t}_l, l \in [1, L]$ is similarly generated, with associated $J(\mathbf{t}_l)$ terms.

6)A trial solution $\mathbf{t}'$ is selected from the solutions $(\mathbf{t}_k, \mathbf{t}_l)$ with the following probability:

$$\alpha(\mathbf{t} \to \mathbf{t}') = \frac{J(\mathbf{t}')}{\sum_{k=1}^{K} J(\mathbf{t}_k) + \sum_{l=1}^{L} J(\mathbf{t}_l)} \tag{13}$$

To show that this procedure generates a uniform distribution, the phi/psi angles of an 11 residue polypeptide is sampled with no potential. Half of the time, the loop closure procedure is applied as described abover, and the other half of the time, only driver angle is perturbed uniformly, with the remaining cartesians updated accordingly (with no closure condition enforced). The second procedure is required so that the full space of dihedral angles are accessible. Every move is accepted, with no potential applied or steric exclusion. This procedure generates a uniform distribution of torsions, as is shown in Fig. 3. It shows a distribution of an 11-residue peptide sampled with the loop closure procedure described above. Only backbone *DoF* are sampled, and no forcefield is applied. The endpoints are constrained to fixed positions. This control closely follows previous work [20,23]. Fig. 3(a) shows the distribution of angles with no Jacobian selection term applied, and Fig. 3(b) shows the distribution with the reweighting term applied. The Jacobian term clearly improves the uniformity of the sampling here.

2.6 Sidechains

The efficient sampling of sidechains [22] is important since sidechain conformations often determine the biological function of proteins. In the current work, the sidechain χ angles are not taken from the rotamer library due to their nonuniform distribution. Instead, to generate the sidechain trial moves, a single sidechain is randomly selected and each χ angle

is perturbed by a value which is randomly and uniformly distributed in a defined domain $[-d/2, d/2]$ [25,44]. The polar hydrogens for the selected residue are sampled as well over the domain $[-\pi, \pi]$.

To improve the sampling efficiency, no energy is computed for the states with steric clashes, which are defined based on the distances between heavy atoms. Specifically, a steric clash is defined when pairs of heavy atoms are closer than 0.7 times the sum of their Lennard-Jones radii. Rapid identification of steric clashes (using neighbor lists) avoids computationally expensive energy evaluations, for conformations that will result in very high energies and negligible acceptance probabilities.

The most expensive term in energy evaluation is the solvation energy in which the time consuming step is the computation of Born radii. Since the Born radii and the long range energy terms generally vary slowly for relatively small, local conformational changes, less frequent evaluation of these terms will contribute more to the sampling performance. For this purpose the Multiple Time-Step Monte Carlo sampling (MTSMC) procedure [45] is incorporated in the present method, in a scheme based on that in Ref. [44]. The Born radii and the long range interactions are held fixed at the latent state of the original coordinates during the inner loop sampling, and only updated in every outer loop calculation. The final configuration from the inner loop is then taken to be a trial move and subjected to the MTSMC acceptance criterion (see Eq. 20 in Ref. [25])

2.7 The POSH Monte Carlo Method

Both the TLC method for determining the backbone moves of loop residues and the sidechain sampling via perturbation have been incorporated in the POSH (port out, starboard home) Monte Carlo method introduced in a previous work [25]. The application of this method on small peptide systems has shown reasonable agreement with experiments [25]. In the present work, we are interested in its performance in more complicated protein systems with flexible loops.

Briefly, the move sets in this approach consist of two steps: an initial trial ($1 \rightarrow 2$) move with large perturbation followed by a series of annealing moves consisting of smaller perturbation within the inner loop of length N_I ($2 \rightarrow 3$). The generalized Metropolis acceptance probability for this series of moves is given by:

$$acc(1 \rightarrow 3) = min\left(1, \frac{p_3 T_{41}}{p_1 T_{23}}\right) \tag{14}$$

where p_1 and p_3 are the probabilities of being in the original and final trial state, respectively. T_{41} and T_{23} are transition probabilities. T_{23} is the normal forward transition

probability, as would be given in the usual derivation of detailed balance, but T_{41} is a reverse transition probability that is constructed using an alternative reverse path through configuration space that is constructed by taking the final state (state 3) and subtracting the perturbation ($1 \rightarrow 2$) from state 3 to arrive at state 4. Further details are given in [25].

The trial moves are generated by perturbation which uniformly varies over some domain $[-d/2, d/2]$ with a different magnitude for the initial and annealing steps. In this work, for both types of trial moves, either backbone or sidechain is allowed to be perturbed with equal probability. For backbone perturbations, the ϕ or ψ dihedral angle can vary over the domain of $[-2\pi, 2\pi]$ for initial steps and $[-\pi/4, \pi/4]$ for annealing steps. For sidechain χ angles, the domain is $[-\pi, \pi]$ and $[-\pi/9, \pi/9]$, respectively, for the initial and inner step trial moves. The number of inner steps N_I is set to 20 which was reported as the upper bound of inner steps for generating precise distribution. For all protein systems studied in this work, a mixture of 50% posh and 50% standard MC sampling followed by MTSMC procedure is used due to its better performance as studied in the previous work [25].

3 Simulations

We applied the loop Monte Carlo method described above to several proteins with flexible loops. The first is the enzyme triosephosphate isomerase (TIM) which has been used as a model system for studying loop flexibility, primarily by NMR. This enzyme catalyzes the reversible isomerization of dihydroxy-acetone phosphate (DHAP) to D-glyceraldehyde 3-phosphate (GAP). The active site loop 6 (residues 167–176) undergoes conformational changes upon ligand binding, and is believed to be flexible in the absence of ligand binding, transitioning between 'open' and 'closed' states. To assess the capability of our method to capture the dynamical properties of this flexible loop, three sets of simulations were performed. The first one started from the apo yeast TIM (PDB ID 1YPI) with open loop conformation (we call this SIM1), and the second started from the 2-phosphoglycolate (PGA)-bound TIM (PDB ID 2YPI) with closed loop conformation (SIM2), and the third is the same as the second except that the ligand PGA was removed from the initial structure (SIM3).

The titratable residues in the starting structures were predicted according to the experimental conditions. Specifically, in all simulations, His95 was treated as neutral, and protonated on the Nϵ2. Glu165 is protonated in SIM2 in order to maintain the strong interaction with ligand PGA [9], but was unprotonated in the other simulations. Residues within 8 Å of the active site loop were included for the side chain sampling and the flexible

loop was extended to include residues 165–178 in the simulations for both the backbone and side chain sampling. The force field OPLS-AA [46,47] was used for the protein TIM and ligand PGA except that the partial charges for the phosphate group of PGA were adjusted based on the previous work by Wong et al. [48] The Surface Generalized Born (SGB) [49,50] model was used for implicit solvent with the treatment of nonpolar term [50]. To prevent the sampling from being trapped in local minima, all simulations were performed at the temperature of 600K. Each simulation has a length of $N_o = 2 \times 10^5$ up to 5×10^5 outer steps. Data analyses were performed over the equilibrium simulations (roughly after 10^5 outer steps) during which the potential energy is relatively stable.

The same protocol was also applied to other protein systems which have been studied by NMR experiments, specifically those with PDB ID 1H2O, 1XWE, and 1Q9P. By choosing NMR structures, we eliminate any concerns about crystal packing influencing the loop conformation or flexibility. These specific proteins were chosen because each has two loops consisting of 5–8 residues, one of which has multiple conformations with large variation among the various NMR models (flexible loop) and the other has a narrow range of loop conformations among the NMR models (rigid loop). Both the flexible and rigid loops were simulated using the same sampling protocol and the same parameter settings in order to compare with the experimental data since both loops within the same protein were measured in the same experimental conditions. The titratable residues in the starting structures were protonated at the experimental pH = 7.0 for 1H2O, 6.0 for 1XWE and 5.8 for 1Q9P. The flexible loops consist of residues 59–64 for 1H2O, 1609–1616 for 1XWE and 48–53 for 1Q9P; the residues in the rigid loops are 46–51 for 1H2O, 1536–1540 for 1XWE and 78–82 for 1Q9P.

4 Results and Discussion

As an initial illustration of the utility of our loop MC method for sampling conformation space of protein loops, we applied this method to the well-studied enzyme triosephosphate isomerase (TIM). The active site loop undergoes large-scale motions interconverting between open and closed conformations. This conformational transition occurs on the time scale of milliseconds [17], making it a challenge for molecular dynamics simulations in previous studies [51,52].

In the current work, multiple transitions between open and closed loop conformation of yeast TIM have been observed in the simulation of the apo protein. Figure 4 (a) and (c), which start from the open and closed state, respectively, show sampled loop conformations from the equilibrium ensemble, spanning both open and closed form. In the simulation with

the ligand PGA bound, the active site loop stays in the closed conformation, as can be seen in Fig. 4(b). These results agree qualitatively with NMR experiments which found that the loop samples open and closed conformations whether or not a ligand was bound, but that ligand binding shifted the equilibrium strongly towards the closed conformation [17,53]. Upon PGA binding, the carboxylate of the ligand protonates residue Glu165 making it hydrogen bonded with PGA instead of with Ser96 in the apo structure, such that the closed loop conformation is prefered with the presence of ligand.

It has been known that the active site loop of TIM moves largely as a rigid unit [51,54]. Figure 5 shows that the backbone dihedral angles of the flexible loop in the X-ray structure of apo TIM are very similar to those in the structure of ligand-bound TIM. The ensembles generated by the loop MC method largely agree with the experimental data in this regard. We calculated the backbone ϕ and ψ angles and averaged them over the equilibrium ensemble for each of the three simulations. For the holo simulations, the ensemble averaged ϕ and ψ angles agree well with those measured in the X-ray structures, as shown in Fig. 5 (a) and (b) (blue lines). Similar agreement was also found for the apo simulations started from both the open and closed conformation, except that residues 170–173 have relatively large deviation and fluctuation, which is actually consistent with the findings in previous simulation studies [17,52] (red and green lines in Fig. 5 (a) and (b)).

NMR spectroscopy can provide information for both structure and dynamics of proteins in physiologically relevant environment [55]. The chemical shift is NMR's most ubiquitous parameter, the variation of nuclear magnetic resonance frequencies of the same kind of nucleus due to variations in the electron distribution. To directly compare with the experimental data, ensemble averaged chemical shifts were calculated for each equilibrium ensemble by using SHIFTX [56] to calculate chemical shifts for the residues of the flexible loop in each conformation and then averaging over all the conformations in the ensemble. For the apo simulations, starting from either the open or closed structures, the ensemble-averaged chemical shifts were compared with NMR measurements of apo yeast TIM [57]. For the simulation of the ligand-bound, closed structure, NMR data measured for G3P-bound yeast TIM [57] were used. [The chemical shifts for the closed loop of the enzyme bounded with G3P and GPA are very similar (Yimin Xu, personal communication)]. A strong linear correlation was found between the ensemble-averaged and experimentally measured chemical shifts for C_α (Fig. 6 (a)) and C_β (Fig. 6 (b)) atoms with the correlation coefficient r of 0.98 or higher in all cases. For carbonyl C and amide N atoms of the flexible loop, although there are fewer experimental chemical shifts available, the calculated ensemble averages have small variations from experimental values (Fig. 6 (c) and (d)). We noticed that the experimental chemical shifts were measured at 300 K, while our simulations were

performed at 600 K. This is because at 300 K it is difficult to observe the conformational transitions between the open and the closed state since the implicit solvent model we used over stabilizes the salt-bridges. However higher temperature weakens this effect and reasonable ensembles are generated which agree with the NMR chemical shifts. The effect of constraining the omega angles, as well as the bond angles and lengths, in addition to the loop closure condition can also limit the sampling, such that a higher temperature sampling protocol is appropriate. Since the system is a monte carlo system, and the degrees of freedom are constrained such that the overall structure is preserved, a higher temperature sampling protocol can still provide physical insights. The efficiency gained by sampling a lower dimensional space, while still having a reasonable estimate of ensemble properties motivates the use of this set of approximations.

As a second initial application, we also applied our sampling method to other NMR protein targets which have loops with differing flexibility in order to evaluate our ability to distinguish the flexible and rigid loops within the same protein. The conformational ensembles from equilibrium simulations for both the flexible and rigid loops are shown in Fig. 7 for three proteins with PDB ID 1H2O (a), 1XWE (b), and 1Q9P (c) sampled at 600 K (left) and 300 K (right). These results clearly show that the loop residues which are flexible in the experimentally derived structures consistently are more floppy in the sampled ensemble at either temperature than the loop residues which are relatively rigid in the same NMR structures. To further quantify these results, root mean square fluctuation (RMSF) of the heavy atoms in both loops were calculated for the sampled and NMR models as shown in Table 1. We recognize that the set of NMR models for each protein cannot be viewed as a true ensemble, but the qualitative agreement is nonetheless encouraging. Thus our method can be an alternative way besides molecular dynamics simulations to capture the distinction between loop flexibilities within the same protein. Comparing the composition of floppy loop with that of rigid loop, we found that the latter has at least one proline residue in the middle of loops which restrict the range of conformations that loop can adopt. On the other hand, the floppy loops are more solvent exposed and have less interaction with their neighbors. For simulations of all studied protein systems, three NMR targets and TIM, the average acceptance ratio is about 14%.

Our current approach only variates phi-psi angles as they are most flexible, but actually we could include all other dof's in the MC scheme in any desired combination. We are working on a further version of the algorithm that will incorporate sampling which allows omega angles, as well as bond lengths and angles to fluctuate more freely, which may allow for lower temperature sampling of systems of this type. The preliminary results show that it gives a pretty good sampling with better statistics. Although in the present study we

have considered solvation effects implicitly only, including water molecules explicitly in the simulation is possible in principle. Since water molecules in the immediate vicinity of a loop would introduce additional steric clash possibilities for large moves, that would introduce additional restriction on move size. We plan to explore the implications in future work.

Acknowledgments

This work was supported in part by grants from NIH-NIGMS, GM081710 (to MPJ and EAC) and R01-GM090205 (EAC). MPJ is a consultant to Schrodinger LLC.

References

(1) Jones, D. *Curr Opin Struct Biol* **1997,** *7,* 377.

(2) Fiser, A.; Do, R.; Sali, A. *Protein Sci* **2000,** *9,* 1753.

(3) Al-Lazikani, B.; Jung, J.; Xiang, Z.; Honig, B. *Curr Opin Struct Biol* **2001,** *5,* 51.

(4) Jacobson, M. P.; Pincus, D.; Rapp, C.; Day, T.; Honig, B.; Shaw, D.; Friesner, R. *Proteins* **2004,** *55,* 351.

(5) Meirovich, H. *Chem Phys Lett* **1977,** *45,* 389.

(6) Baysal, C.; Meirovich, H. *J. Phys. Chem. A* **1997,** *101,* 2185.

(7) Mihailescu, M.; Meirovitch, H. *J. Phys. Chem. B* **2009,** *113,* 7950.

(8) Lolis, E.; Abler, T.; Davenport, R.; Rose, D.; Hartman, F.; Petsko, G. *Biochemistry* **1990,** *29,* 6609.

(9) Lolis, E.; Petsko, G. *Biochemistry* **1990,** *29,* 6619.

(10) Dar, A.; Lopez, M.; Shokat, K. *Chemistry and Biology* **2008,** *15,* 1015.

(11) Padlan, E. *Adv. Protein Chem.* **1996,** *49,* 57.

(12) Xu, J.; Davis, M. *Immunity* **2000,** *13,* 37.

(13) Wong, S.; Jacobson, M. P. *Proteins* **2010,** in press.

(14) Rapp, C.; Pollack, R. *Proteins* **2005,** *60,* 103.

(15) Wong, S.; Jacobson, M. P. *Proteins* **2008,** *71,* 153.

(16) Yi, M.; Tjong, H.; Zhou, H. *Proc. Natl. Acad. Sci. USA* **2008,** *105,* 8280.

(17) Massi, F.; Wang, C.; Palmer, A. G. *Biochemistry* **2006,** *45,* 10787.

(18) Go, N.; Scheraga, H. *Macromolecules* **1970,** *3,* 178.

(19) Bruccoleri, R. E.; Karplus, M. *Macromolecules* **1985,** *18,* 2767.

(20) Dodd, L. R.; Boone, T. D.; Theodorou, D. N. *Mol Phys* **1993,** *78,* 961.

(21) Deem, M.; Bader, J. *Mol.Phys.* **1996,** *87,* 1245.

(22) Wu, M. G.; Deem, M. W. *Mol. Phys.* **1999,** *97,* 559.

(23) Dinner, A. *J Comput Chem* **2000,** *21,* 1132.

(24) Ulmschneider, J. P.; Jorgensen, W. L. *J Chem Phys* **2003,** *118,* 4261.

(25) Nilmeier, J.; Jacobson, M. P. *J. Chem. Theory Comput.* **2009,** *5,* 1968.

(26) Coutsias, E. A.; Seok, C. L.; Jacobson, M. P.; Dill, K. A. *J Comput Chem* **2004,** *25,* 510.

(27) Hayward, S.; Kitao, A. *Biophysical Journal* **2010,** *98,* 1976.

(28) Wedemeyer, W. J.; Scheraga, H. A. *J Comp Chem* **1999,** *20,* 819.

(29) Wu, M. G.; Deem, M. W. *J. Chem. Phys.* **1999,** *111,* 6625.

(30) Cortes, J.; Simeon, T.; Remaud-Simeon, M.; Tran, V. *J Comput Chem* **2004,** *25,* 956.

(31) Noonan, K.; O'Brien, D.; Snoeyink, J. *Int J Robotics Res* **2005,** *24,* 971.

(32) Milgram, R.; Liu, G.; Latombe, J. *J Comput Chem* **2008,** *29,* 50.

(33) Favrin, G.; Irbäck, A.; Sjunnesson, F. *J Chem Phys* **2001,** *114,* 8154.

(34) Wang, L.-C. T.; Chen, C. C. *IEEE TRANS ROBOT. AUTOM.* **1991,** *7,* 489.

(35) Cahill, S.; Cahill, M.; Cahill, K. *J Comp Chem* **2003,** *24,* 1364.

(36) Canutescu, A.; Dunbrack, R. *Protein Sci* **2003,** *12,* 963.

(37) Lee, A.; Streinu, I.; Brock, O. *Phys Biol* **2005,** *2,* 108.

(38) Stoer, J.; Bulirsch, R. *Numerical Analysis;* Springer: Berlin, Second ed.; **1991**.

(39) Coutsias, E. A.; Seok, C.; Wester, M. J.; Dill, K. A. *Int J. Quant. Comp.* **2006,** *106,* 176.

(40) Mandell, D. J.; Coutsias, E. A.; Kortemme, T. *Nature Methods* **2009,** *6,* 551.

(41) Gibbs, J. W.; Wilson, E. B. *Vector Analysis;* Yale University Press: New Haven, First ed.; **1901**.

(42) Pollock, S. N.; Coutsias, E. A. "Numerical Analysis of Inverse Kinematic Algorithms", preprint, **2011**.

(43) Hoffman, D.; Knapp, E.-W. *European Biophysical Journal* **1996,** *24,* 387.

(44) Nilmeier, J.; Jacobson, M. P. *J Chem Theory Comput.* **2008,** *4,* 835.

(45) Hetenyi, B.; Bernacki, K.; Berne, B. *J. Chem. Phys.* **2002,** *117,* 8203.

(46) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B.* **2001,** *105,* 6474.

(47) Jorgensen, W.; Maxwell, D.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996,** *118,* 11225.

(48) Wong, S.; Bernacki, K.; Jacobson, M. P. *J. Phys. Chem. B* **2005,** *109,* 5249.

(49) Ghosh, A.; Rapp, C.; Friesner, R. *J Phys Chem B* **1998,** *102,* 10983.

(50) Gallicchio, E.; Zhang, L.; Levy, R. *J Comput Chem* **2002,** *23,* 517.

(51) Joseph, D.; Petsko, G.; Karplus, M. *Science* **1990,** *249,* 1425.

(52) Derreumaux, P.; Schlick, T. *Biophys. J.* **1998,** *74,* 72.

(53) Williams, J. C.; McDermott, A. E. *Biochemistry* **1995,** *34,* 8309.

(54) Davenport, R.; Bash, P.; Seaton, B.; Karplus, M.; Petsko, G.; Ringe, D. *Biochemistry* **1991,** *30,* 5821.

(55) Teng, Q. Protein Structure Determination from NMR Data. In *Structural Biology: Practical NMR Applications*, First ed.; Lee, W., Ed.; Springer: Berlin, **2005**.

(56) Neal, S.; Nip, A.; Zhang, H.; Wishart, D. *J Biomol NMR.* **2003,** *3,* 215.

(57) Xu, Y.; Lorieau, J.; McDermott, A. E. *J. Mol. Biol.* **2010,** *397,* 233.

Table 1: Root-mean-squared fluctuation (RMSF) (Å) of heavy atoms of both floppy and rigid loops in the equilibrium ensemble simulated by POSH MC method with initial structure of the first model of NMR structures. For comparison, the RMSF over all NMR models for each protein are also computed at both 600 K and 300 K.

Heavy atom RMSF	1H2O			1Q9P			1XWE		
	NMR	POSH		NMR	POSH		NMR	POSH	
		600 K	300 K		600 K	300 K		600 K	300 K
Flexible loop	2.75	1.64	0.75	3.51	1.27	1.06	4.83	2.50	1.10
Rigid loop	0.49	0.40	0.15	1.25	0.38	0.18	1.03	0.50	0.31

Figure Captions

Fig. 1: (A) The atoms and parameters defining triaxial loop closure (TLC). (B) The generalized 6R/3A kinematic chain.

Fig. 2: Construction of a tripeptide move. A node consists of a ϕ/ψ pair at each alpha carbon of the loop (with only backbone shown). The yellow filled circle is the alpha carbon whose dihedral angle servers as a driver angle (the wide black arrow). A randomly constructed triaxial closure is shown as the grey triangle in which each grey circle represents the randomly selected pivot.

Fig. 3: Distribution of ϕ/ψ angles without (A) and with (B) Jacobian weighting of selection for an 11-residue peptide. A Total of 4.5×10^5 trial moves were generated. No forcefield is used the selection probability, and all trial moves are accepted.

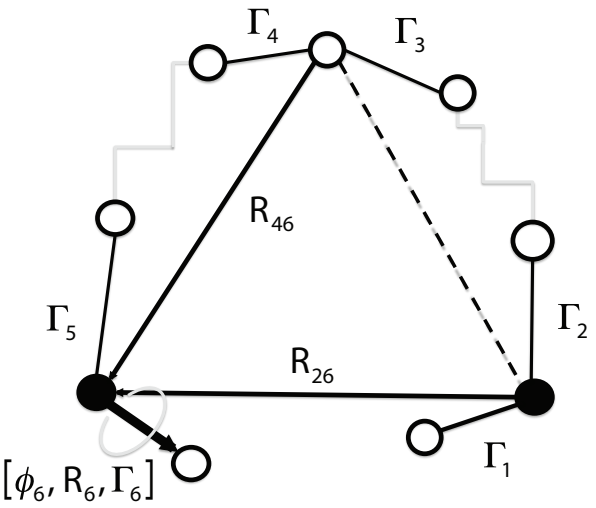
Fig. 4: The ensemble structures (red) for the flexible loop (residues 165–178) of yeast TIM were taken from the equilibrium simulation with initial structures of (A) apo (open) conformation, (B) bound (closed) conformation and (C) the closed conformation with the ligand PGA removed. The X-ray structure of apo yeast TIM (PDB 1YPI) is shown in yellow and bound state (PDB 2YPI) in cyan. The ligand PGA is depicted by spheres.

Fig. 5: The comparison of the calculated backbone dihedral angles, ϕ (A) and ψ (B), with those measured in the X-ray structures. The black solid line is for apo TIM (PDB 1YPI) and the dashed line for the ligand-bound TIM (PDB 2YPI). The calculated dihedral angles were averaged over the equilibrium ensemble simulated from the initial structure of apo (red), ligand-bound (blue), and closed form with the ligand PGA removed (green).

Fig. 6: Ensemble-averaged chemical shifts (ppm) versus the NMR experimental measurements for C_α (A), C_β (B), carbonyl C (C), and amide N (D) atoms of the flexible loop 6 of yeast TIM. SHIFTX [56] was used to calculate chemical shifts which were then averaged over an ensemble of 1000 structures from the equilibrated MC simulations. The starting PDB structures for the simulations are: 1YPI (black); 2YPI with the ligand PGA removed (red); and 2YPI with PGA bound (green). The experimental chemical shift data are those for apo yeast TIM in NMR experiment [57] (for comparison with the apo simulations), and for yeast TIM with ligand G3P [57] (for comparison with the holo simulation). Experimental chemical shifts are not available for some atoms and these are omitted.

Fig. 7: Ensembles of loop structures from equilibrium simulations using MC sampling for proteins with PDB ID: (A) 1H2O, (B) 1XWE and (C) 1Q9P sampled at T=600 K (left) and T=300 K (right). The sampled flexible loops ('floppy') which have large fluctuation in the NMR models are shown in red and the rigid loops with very small fluctuations are in blue. The structures in yellow are taken from MODEL 1 of the PDB file.

(A)


Γ_4
Γ_3
R_{46}
Γ_5
Γ_2
R_{26}
Γ_1
$[\phi_6, R_6, \Gamma_6]$

(B)

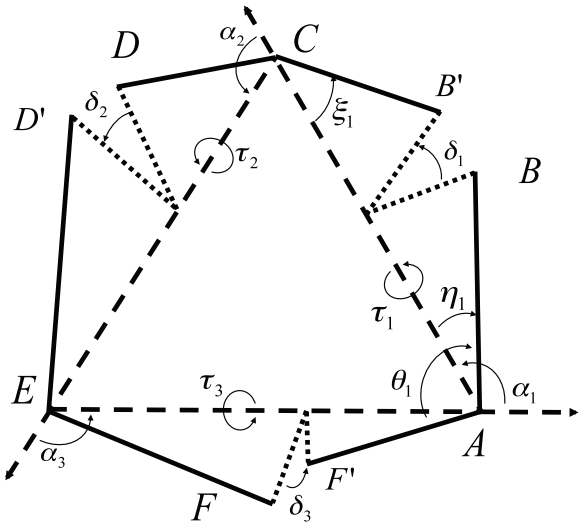

D
α_2
C
B'
D'
δ_2
ξ_1
τ_2
δ_1
B
τ_1
η_1
E
τ_3
θ_1
α_1
α_3
A
F'
F
δ_3

Figure 1:

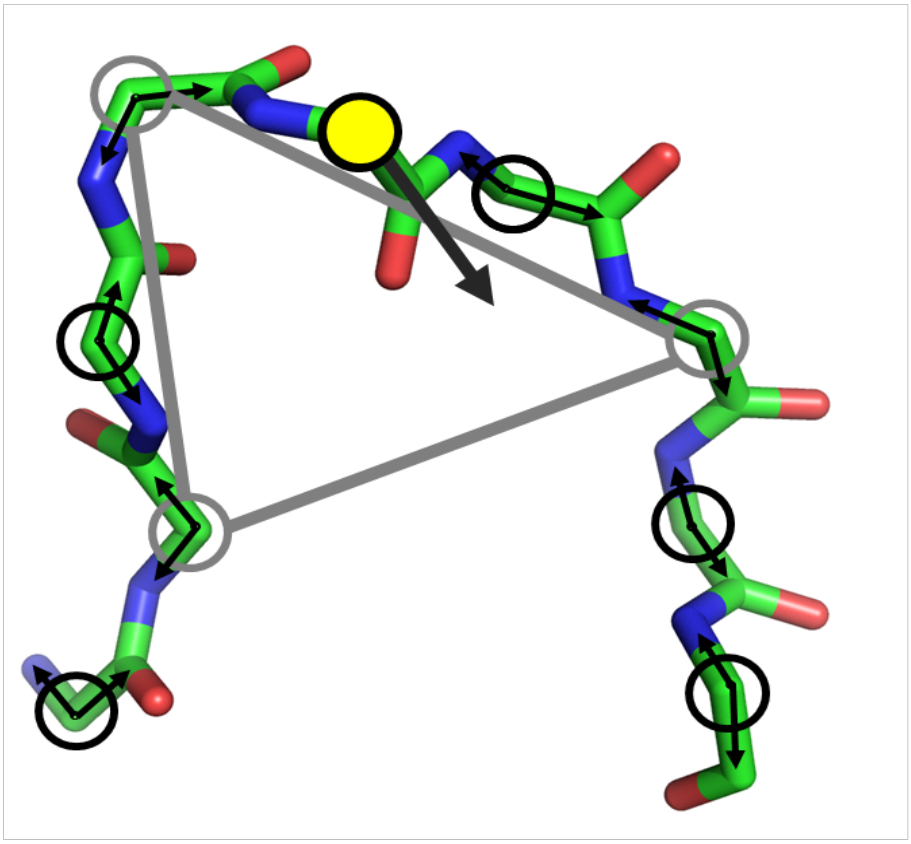


Figure 2:

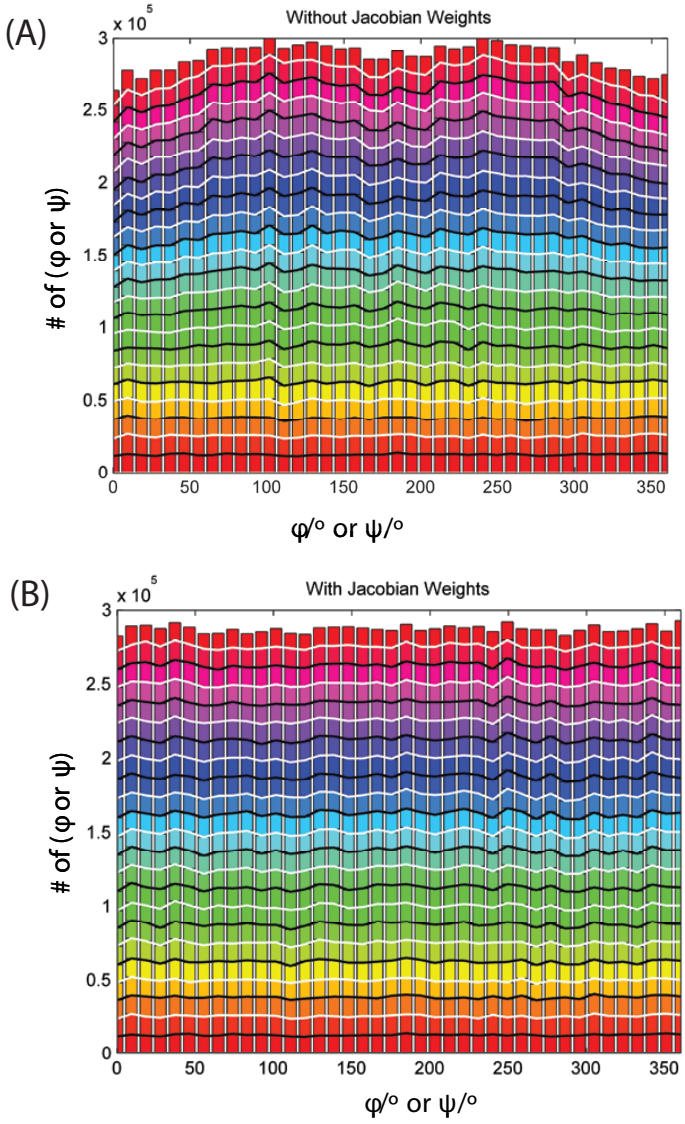

(A)
Without Jacobian Weights
x 10^5
of (φ or ψ)
φ/° or ψ/°
(B)
With Jacobian Weights
x 10^5
of (φ or ψ)
φ/° or ψ/°

Figure 3:

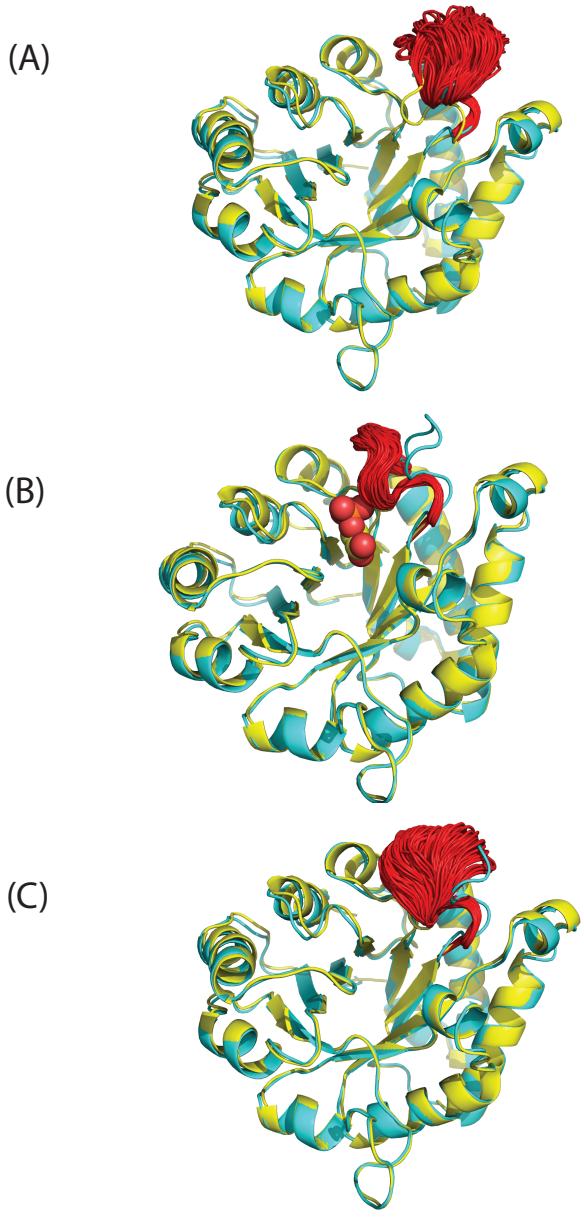


Figure 4:

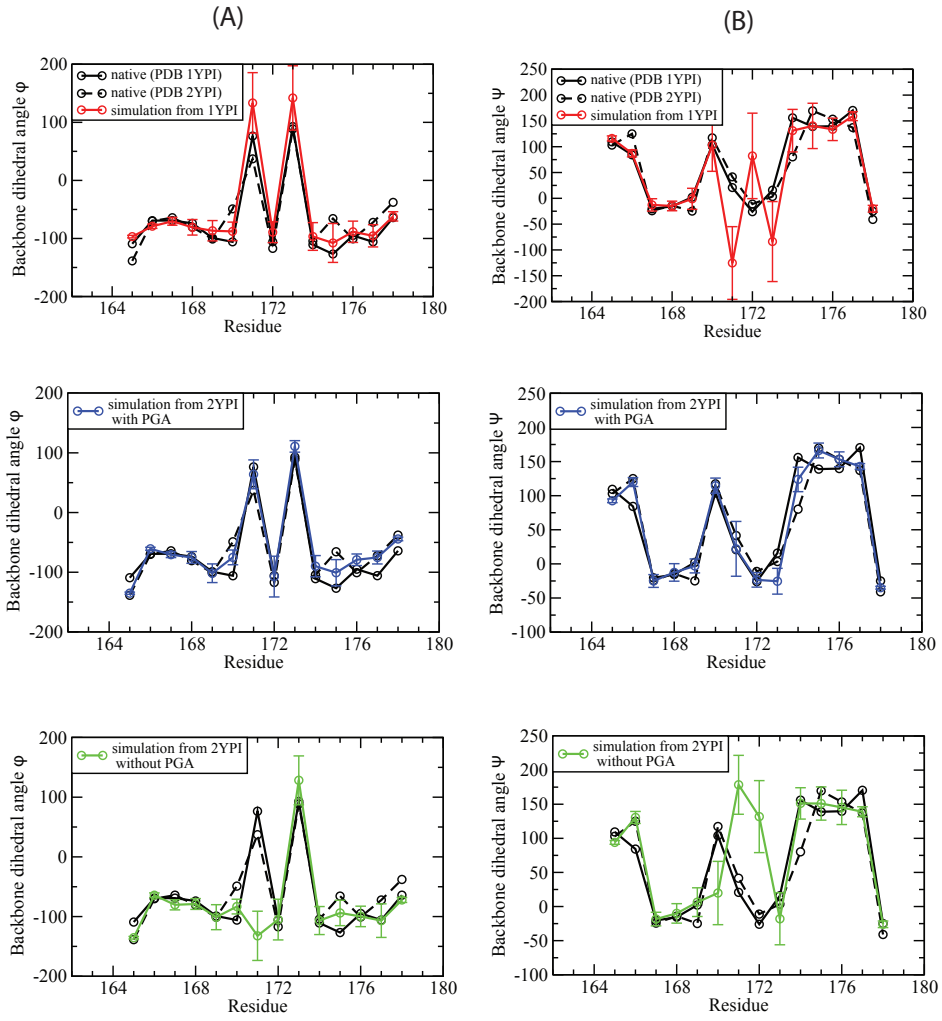

(A)
(B)
native (PDB 1YPI)
native (PDB 2YPI)
simulation from 1YPI
simulation from 2YPI with PGA
simulation from 2YPI without PGA
Backbone dihedral angle φ
Backbone dihedral angle ψ
Residue

Figure 5:

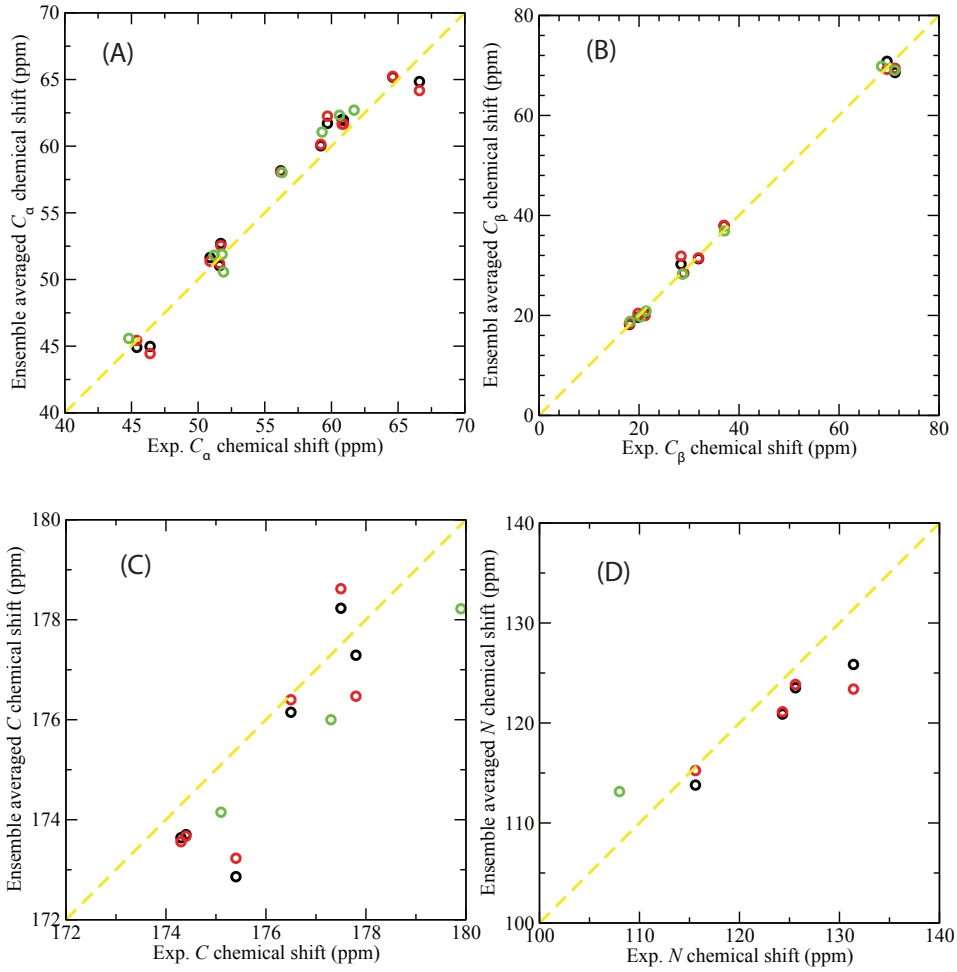

(A)
Ensemble averaged C_α chemical shift (ppm)
Exp. C_α chemical shift (ppm)
(B)
Ensembl averaged C_β chemical shift (ppm)
Exp. C_β chemical shift (ppm)
(C)
Ensemble averaged C chemical shift (ppm)
Exp. C chemical shift (ppm)
(D)
Ensemble averaged N chemical shift (ppm)
Exp. N chemical shift (ppm)

Figure 6:

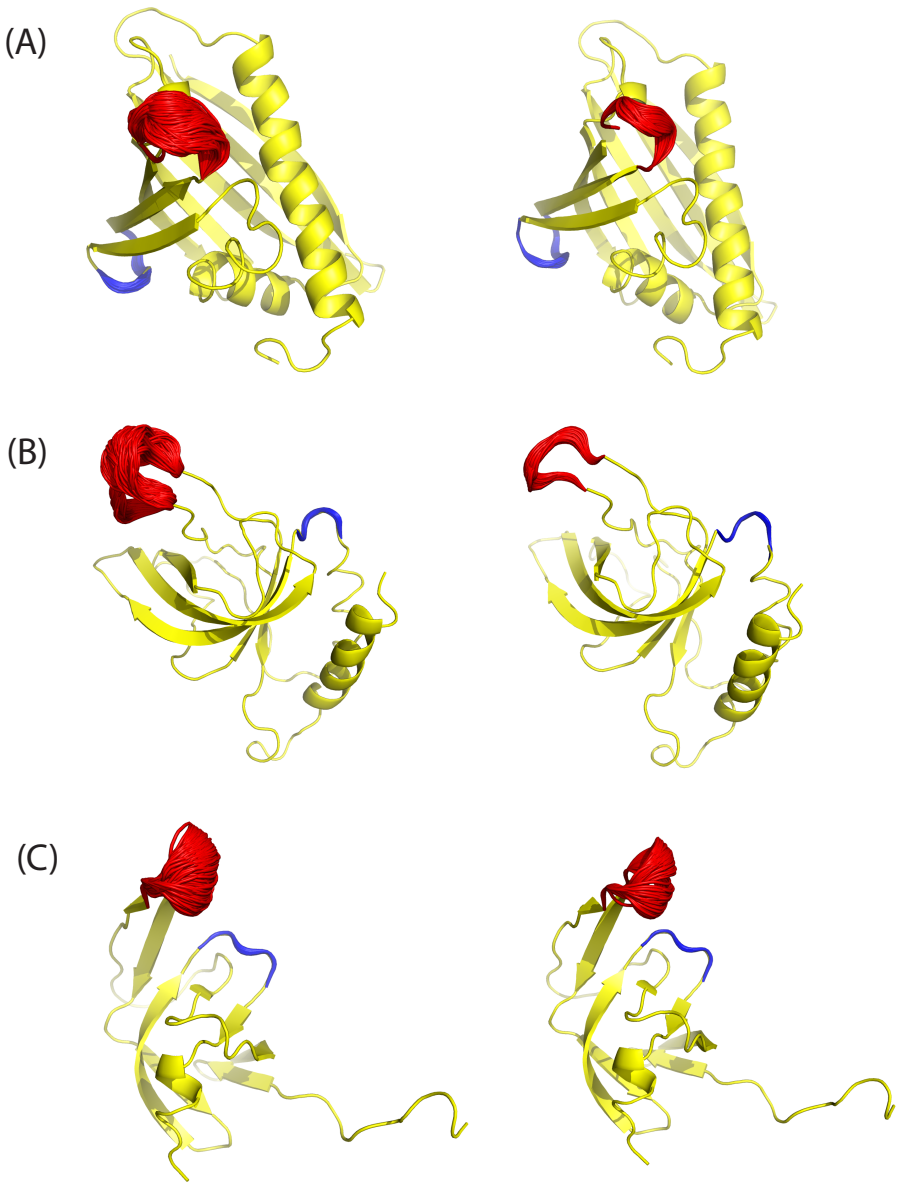


Figure 7: