Folding and Misfolding of the Collagen Triple Helix: Markov Analysis of Molecular Dynamics Simulations

Sanghyun Park,* Teri E. Klein,† and Vijay S. Pande*‡
*Department of Chemistry, †Department of Genetics, and ‡Department of Structural Biology, Stanford University, Stanford, California

ABSTRACT Folding and misfolding of the collagen triple helix are studied through molecular dynamics simulations of two collagenlike peptides, $[(POG)_{10}]_3$ and $[(POG)_4POA(POG)_5]_3$, which are models for wild-type and mutant collagen, respectively. To extract long time dynamics from short trajectories, we employ Markov state models. By analyzing thermodynamic and kinetic quantities calculated from the Markov state models, we examine folding mechanisms of the collagen triple helix and consequences of glycine mutations. We find that the C-to-N zipping of the collagen triple helix must be initiated by a nucleation event consisting of formation of three stable hydrogen bonds, and that zipping through a glycine mutation site requires a renucleation event which also consists of formation of three stable hydrogen bonds. Our results also suggest that slow kinetics, rather than free energy differences, is mainly responsible for the stability of the collagen triple helix.

INTRODUCTION

The collagen family is a group of structural proteins that make up tissues such as bone, skin, and cartilage (1,2). The defining feature of the collagen family is the triple helical structure composed of three chains wrapped around each other. Each chain is made of repeating Gly-Xaa-Yaa triplets, where the X and Y positions are often occupied by proline and hydroxyproline, respectively. The presence of glycine in every third position is crucial as glycine is the only amino acid that is small enough to fit in the narrow core of the triple helix. Point mutations of these glycines lead to misfolding of the triple helix, which in turn cause various diseases such as *osteogenesis imperfecta* (brittle bone disease) (3).

The biosynthesis of collagen is a coordinated process involving multiple stages (1,2). Here we focus on the stage where three chains fold into a triple helix, which is where the glycine mutations have the most impact (4). It is well established that most collagens fold by a zipping mechanism from the C-terminus toward the N-terminus, initiated by nucleation of the triple helix structure near the C-terminus (1,2,5). However, detailed mechanisms of folding and misfolding remain largely unknown.

The repetitive nature of collagen allows one to study short collagenlike peptides instead of long native collagen molecules (e.g., human collagen I contains >1000 residues per chain). Although there are many specific questions that need to be addressed with native sequences, much can be learned from the study of various collagenlike peptides. In fact, most progress to date, experimental or computational, has been made through the study of short collagenlike peptides. In this article, we examine the folding process of two collagenlike peptides, $[(POG)_{10}]_3$ and $[(POG)_4POA(POG)_5]_3$, hereafter referred to as POG and POG-A, respectively, using molecular dynamics (MD) simulations. (O denotes hydroxyproline.) Fig. 1 shows schematic diagrams of these two peptides, POG and POG-A, which model wild-type and mutant collagen, respectively. By analyzing the folding processes of the two peptides, we study the details of the folding mechanism and identify the effect of glycine mutations.

Collagen folding is a very slow process. Even the folding of small collagenlike peptides is too slow to be simulated in its entirety within currently feasible times for atomistic simulations. Therefore, we have adopted the approach of Markov state models (MSMs) (6,7). The idea of MSMs is to decompose the entire configuration space into a set of discrete states such that transitions between the states can be simulated within a practical timescale and that dynamics of the entire folding process can be reconstructed in terms of those transitions. In the following, we describe how we built MSMs from MD simulations of POG and POG-A. The self-consistency (Markovity, in particular) of the models obtained is checked to ensure that they properly capture long time dynamics. By analyzing these MSMs, we study the thermodynamics and kinetics of folding and misfolding of the two peptides, paying particular attention to the effect of the glycine mutation.

CONSTRUCTION OF MARKOV STATE MODELS

State assignment

The main force that holds three chains together in the triple helical form comes from a network of backbone hydrogen bonds, in which the nitrogen atom of Gly forms a hydrogen bond with the carbonyl group of Xaa located in a neighboring chain. There are 29 such hydrogen bonds in POG or POG-A. Due to fraying at the ends, the first two and the last two hydrogen bonds are unstable, and therefore we do not consider them in the state

Submitted March 5, 2007, and accepted for publication June 29, 2007.

Address reprint requests to Teri E. Klein, E-mail: teri.klein@stanford.edu; or Vijay S. Pande, E-mail: pande@stanford.edu.

Editor: Angel E. Garcia.

0006-3495/07/12/4108/08 $2.00

doi: 10.1529/biophysj.107.108100

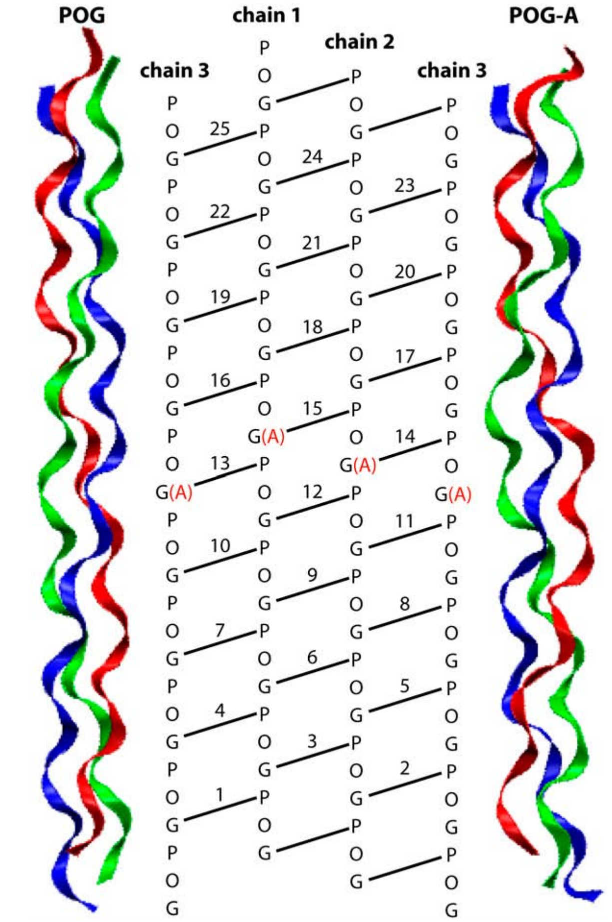

FIGURE 1 Collagenlike peptides, POG = $[(POG)_{10}]_3$ and POG-A = $[(POG)_4POA(POG)_5]_3$. The schematic diagram in the middle shows the network of backbone hydrogen bonds that stabilizes the triple helix. Chain 3 is shown in duplicate. Hydrogen bonds that are included in the state assignment are labeled from 1 to 25, which also indicates the order of hydrogen-bond formation according to the C-to-N zipping mechanism. The molecular figures (made with VMD) represent the backbone conformation of POG (*left*) and POG-A (*right*). Three chains are colored differently: chain 1, red; chain 2, green; and chain 3, blue. In POG-A, a slight bulge is noticeable around the Gly→Ala mutation sites.

assignment. The remaining 25 hydrogen bonds are labeled 1–25 in the C-to-N direction (Fig. 1).

To examine the stability of these hydrogen bonds, we performed 10-ns MD simulations (after 1 ns of initial equilibration) of POG and POG-A at 300 K starting with respective folded structures. Initial coordinates of POG were generated by the Gencollagen program (8) and those of POG-A were taken from a crystal structure (9). Termini were capped with acetyl and N-methyl groups. Initial structures of the two peptides are shown in Fig. 1. Hydrogen-bond distances, defined to be nitrogen-oxygen distances, measured from these simulations are shown in Fig. 2. In POG, all the 25 hydrogen bonds are more or less equally stable, with median distances at ~3.1 Å. In POG-A, the Gly→Ala mutations destabilize hydrogen bonds near the mutation sites.

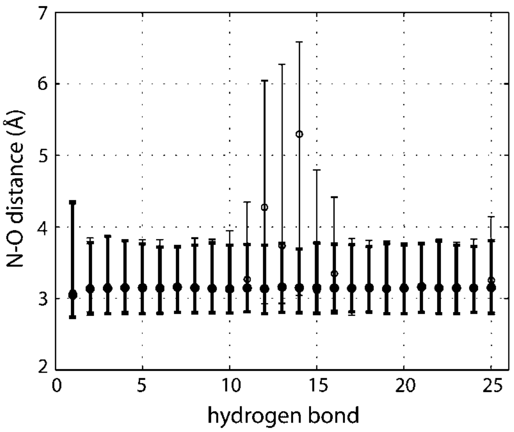

FIGURE 2 Hydrogen-bond distances of the folded structures of POG (*thick line*) and POG-A (*thin line*). Hydrogen bonds are labeled as in Fig. 1. Hydrogen-bond distances are defined to be nitrogen-oxygen distances. Shown are 95% intervals around medians (*circles*), obtained from 10-ns simulations starting with a fully folded structure of each peptide.

A folding event of a short collagenlike peptide may proceed in either N-to-C or C-to-N direction. Our focus here, however, is only the C-to-N mechanism, and therefore we assign states as follows. A hydrogen bond is considered formed when the nitrogen-oxygen distance is <4 Å. (Angles are also relevant for hydrogen bonds. But, for simplicity, we only use the nitrogen-oxygen distance as the criterion for hydrogen-bond formation.) State n ($n = 1,\ldots,25$) represents the stage of folding in which n consecutive hydrogen bonds have been formed by the C-to-N zipping mechanism. (One could contemplate including state 0, in which no hydrogen bonds are formed. The 0→1 transition, then, would have to include the association process that brings three chains together.) For example, state 25 indicates a complete helix with all the hydrogen bonds formed, and state 13 a half helix.

However, it is nontrivial to apply this state assignment consistently to both POG and POG-A. The difficulty lies in that hydrogen bonds near the mutation sites in POG-A are unstable (Fig. 2). For example, consider a conformation of POG-A in which hydrogen bonds 1–11 are formed, 12–14 are broken, and 15–25 are formed. This is a typical conformation observed in MD simulations starting with a fully folded structure of POG-A, and thus should be assigned to state 25, corresponding to the fully folded state. Merely counting the number of hydrogen bonds will lead to the wrong assignment. Therefore, we assign states by identifying the front line of folding (the front-line rule), where the front line is defined to be the location of the two consecutive hydrogen bonds formed closest to the N-terminus, except of course for state 1 for which we cannot but define the front line in terms of a single hydrogen bond. For example, if hydrogen bonds 21 and 22 are the two consecutive ones formed closest to the N-terminus, state 22 is assigned. (Identifying the front line as the single hydrogen bond formed closest to the N-terminus turned out to be too susceptible to fluctuations.)

MD simulations—restrained unfolding and refolding

To generate conformations that belong to various states, we performed unfolding simulations with 25 different restraining schemes (Fig. 3). In each unfolding simulation, the backbone atoms (N, Cα, and C) below a certain boundary were restrained to the initial triple helix structure by harmonic potential with spring constant 10 kcal/mol/Å^2, and the system was simulated for 1 ns at a high temperature ($T = 500$ K) and a high dielectric constant ($\varepsilon = 100$). The combination of a high temperature and a high dielectric constant breaks hydrogen bonds and scrambles three chains, but the restrained parts remain in the triple helix conformation. In this way, conformations for various states are generated. (The n^{th} restraining scheme generates conformations for state n.) For each restraining scheme, 700 unfolding simulations were performed with different initial velocities sampled from the Boltzmann distribution. Accordingly, we obtained a total of $25 \times 700 = 17{,}500$ conformations for each peptide. Typical conformations of a few states are shown in Fig. 4.

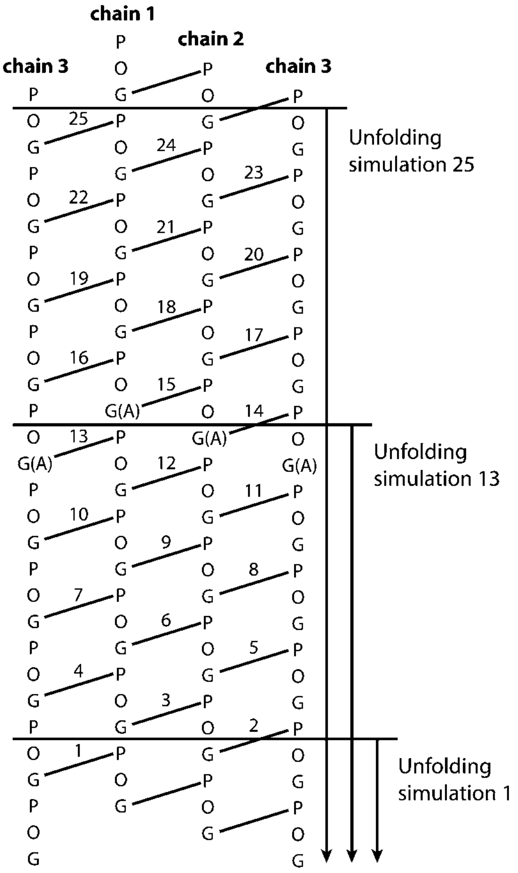

FIGURE 3 Restraining schemes for unfolding simulations. The arrows indicate the parts that were restrained during unfolding simulations. We performed unfolding simulations with 25 different restraining schemes; three of them are shown in this figure.

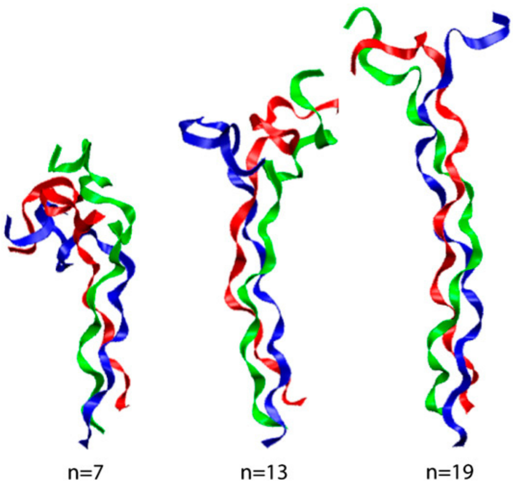

FIGURE 4 Typical conformations of POG for states 7, 13, and 19. Made with VMD (26).

Using as initial coordinates the 17,500 conformations obtained from the above restrained unfolding simulations, we performed refolding simulations for 100 ns at the room temperature ($T = 300$ K) and the standard internal ($\varepsilon = 1$) and external ($\varepsilon = 78.5$) dielectric constants without any restraints. Initial velocities were sampled from the Boltzmann distribution. Total simulation time for refolding would be $17{,}500 \times 100$ ns $= 1750$ μs for each peptide, but due to the heterogeneous nature of distributed computing, not all the trajectories reached 100 ns. The actual data analyzed here amounts to 933 μs for POG and 1229 μs for POG-A. As described in Estimating Transition Probabilities, the Markov analysis allows us to use trajectories of different lengths.

We performed all MD simulations using the AMBER 8 molecular simulation package (10) with the AMBER 99 force field (11). For hydroxyproline, we used the parameters developed in the literature (12). A modified generalized Born method (13) (with the cutoff distance of 1.6 nm for the calculation of Born radii and nonbonded interactions) was used as an implicit water model, and temperature was controlled using Langevin dynamics (using the leapfrog integrator) with the viscosity, i.e., collision frequency, of 1.0/ps. This choice of low viscosity (compared to the viscosity of water, ~90/ps) was necessary to speed up transitions between states; even with this low viscosity, many trajectories stay in the same state over the course of 100 ns. We compensate for the low viscosity when we estimate kinetic quantities. Thermodynamic quantities do not depend on the choice of viscosity. All bonds that involve hydrogen atoms were constrained with the SHAKE algorithm (14), which allowed a time step of 2 fs. Restrained unfolding and refolding simulations were performed on the Folding@Home distributed computing network.

Estimating transition probabilities

The final step in MSM construction is to estimate transition probabilities from the refolding trajectories. First, MD trajectories are turned into sequences of states, known as

Markov chains. Along each trajectory, observations are made to identify which state the system resides in, with the lag time τ_{lag} between observations. In this work, we have used 10 different lag times, $\tau_{\mathrm{lag}} = 1,2, \ldots ,10$ ns. More details on the generation of Markov chains can be found in the Appendix.

Then, from the entire set of Markov chains, transition counts $N(j \rightarrow i)$, i.e., the total number of occurrences of the $j \rightarrow i$ transition, are obtained. And transition probabilities $\theta(i|j)$, i.e., the probability of observing the system in state i at the next step (τ_{lag} later) given that it is presently in state j, are estimated from the transition counts. A transition probability matrix $\theta(i|j)$ constitutes an MSM. We use Bayesian inference for the estimation of transition probabilities from transition counts. Bayesian inference yields distributions of θ from which we can obtain not only point estimates but uncertainties of estimates as well. The error bars shown in the following figures were obtained in this way. More details on Bayesian inference can be found in the Appendix and in the literature (15,16).

One advantage of the Markov analysis is that trajectories of different lengths can be used together. Once we decide how to assign states and choose a lag time, the only information that is needed in constructing an MSM is transition counts. Therefore, trajectories of any lengths can be used as long as they are long enough to yield nonzero transition counts.

ANALYSIS OF MARKOV STATE MODELS

Self-consistency check—verifying Markovity

The validity of an MSM hinges on the assumption that transition probabilities do not depend on the past history of visited states,

$$\theta(i|j) = \theta(i|j \leftarrow k \leftarrow l \leftarrow \ldots) \quad \text{for all } i, j, k, l, \ldots, \tag{1}$$

which is known as Markovity, hence the name Markov state model. For verification of Markovity, various methods have been suggested such as the eigenvalue test (7) and the entropy test (17). However, it turns out that, due to the complexity of our problem, the amount of data we have is not sufficient to get definitive answers from these rigorous tests (uncertainties of eigenvalues and entropies are too large). Therefore, we turned to a heuristic approach.

Markovity is defined with respect to a certain choice of state assignment and lag time. Accordingly, an MSM can be improved (in the sense of Markovity) by refining states or choosing a longer lag time. (Another element of MSM is the order. An MSM can be improved by incorporating higher order transition probabilities, e.g., second order $\theta(i|j \leftarrow k)$, third order $\theta(i|j \leftarrow k \leftarrow l)$, and so on.) Suppose there is a quantity Q that we want to estimate through the Markov analysis. As we make further improvements on our MSM, the model will become more Markovian and the estimate of Q will converge. Therefore, instead of attempting to establish Markovity through rigorous tests, which is very difficult for systems of high complexity, we can simply monitor the estimate of Q as we make improvements. If the estimate of Q has converged, we consider our MSM adequate for the estimation of Q. Notice that an MSM adequate for Q may be inadequate for another quantity Q'.

Below we examine thermodynamic and kinetic quantities using MSMs. For each quantity, we perform a self-consistency check by monitoring how much the corresponding estimate changes over 10 different lag times, $\tau_{\mathrm{lag}} = 1,2, \ldots ,10$ ns, and then discuss implications of our results on the folding mechanism of POG and POG-A, thereby identifying the effects of the Gly → Ala mutations.

Thermodynamics—free energy profile

From an MSM, namely from a transition probability matrix $\theta(i|j)$, a free energy profile is calculated as

$$G(n) = -k_{\mathrm{B}} T \log \phi(n) + C, \tag{2}$$

where $\phi(n)$ is the stationary probability obtained by solving

$$\textstyle\sum_{\mathrm{j}} \theta(i|j)\phi(j) = \phi(i). \tag{3}$$

For the arbitrary constant C, we choose $C = 0$ in all our calculations. Recall that Bayesian inference yields a distribution of the transition matrix θ. From this distribution, we sample 100 transition matrices and calculate $G(n)$ from each. Thus, we obtain a set of 100 different free energy profiles. From this set, we take the median as a point estimate and the 95% symmetric interval about the median as an error bar.

Fig. 5 shows the free energy profiles for POG and POG-A with respect to 10 different lag times, $\tau_{\mathrm{lag}} = 1,2, \ldots ,10$ ns. Overall, no significant change is noticed over the different lag times, and the error bars are fairly small (mostly <0.5 kcal/mol) except for $n = 1$ and 2. (The large error bars at $n = 1$ and 2 are due to the lack of transition counts involving those states. Many simulations starting in state 2 quickly go to state 1 or 3. And many of those starting in state 1 quickly lose the last hydrogen bond and move out of our state space.) It seems that the estimates of free energy have converged and that our MSMs pass the self-consistency check as far as free energy is concerned.

The free energy profile for POG (Fig. 5 *a*) displays three distinct regions. There is a barrier located at $n = 2$, which indicates the difficulty of folding when there are not enough hydrogen bonds already formed. With two hydrogen bonds formed (state 2), the triple helix is unstable and can easily go back to state 1. It is only after forming three hydrogen bonds that zipping can proceed. Remembering that there are two more hydrogen bonds at the C-terminal end that were not included in our state assignment because they are unstable, we state that the folding of collagen requires a nucleation of three stable hydrogen bonds. For states 3–21, the free energy decreases more or less linearly by the amount of 2 kcal/mol (0.1 kcal/mol per step), which we identify as zipping. The rest (states 21–25)

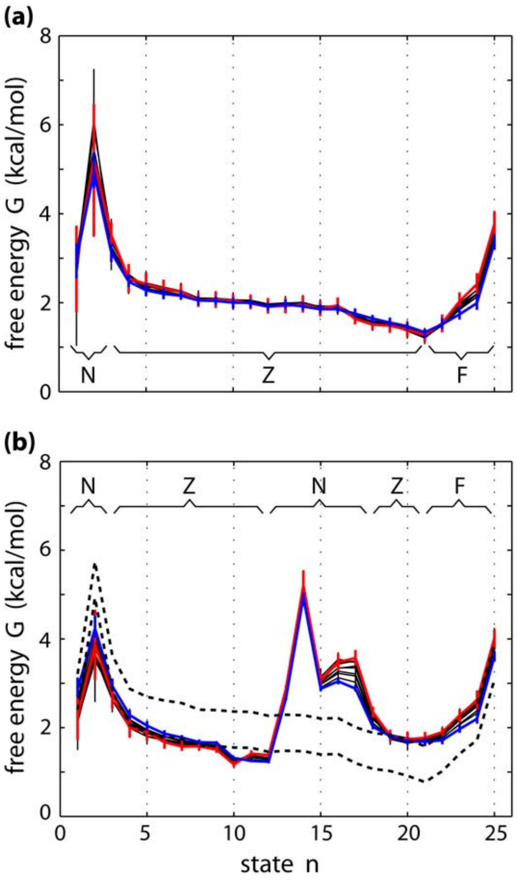

FIGURE 5 Free energy profile $G(n)$. (*a*) POG. (*b*) POG-A. In each panel, there are 10 graphs of $G(n)$ obtained with 10 different lag times, $\tau_{\mathrm{lag}} = 1,2, \ldots ,10$ ns. The graphs for the shortest (1 ns) and the longest (10 ns) lag times are shown in blue and red, respectively. And the other eight graphs are shown in black. The error bars indicate 95% Bayesian intervals around medians. We characterize free energy regions in terms of nucleation (*N*), zipping (*Z*), and fraying (*F*). In panel *b*, for comparison, $G(n)$ for POG—the average of the 10 graphs in panel *a*—is shown as two dashed lines; the upper line is placed such that it matches the POG-A graph on the right, and the lower line is placed such that the match happens on the left.

is uphill, indicating the difficulty of folding the last several hydrogen bonds due to the fraying effect at the N-terminal end. In summary, we characterize regions of the free energy profile for POG as nucleation (state 1–3), zipping (state 3–21), and fraying (state 21–25).

Many things happen when folding propagates by the C-to-N zipping mechanism: hydrogen bonds are formed, water molecules are rearranged (this is only implicitly modeled in our simulations; see MD Simulations—Restrained Unfolding and Refolding), three chains are locked into a triple helical conformation, and so on. All these contribute to the free energy decrease of 0.1 kcal/mol per step during the zipping process. This value of free energy decrease, however, suggests rather marginal stability of the collagen triple helix against thermal fluctuations. (Notice that $k_{\mathrm{B}}T = 0.6$ kcal/mol at 300 K.) We will revisit the issue of stability in the next section.

We now examine the effect of the Gly → Ala mutations on the free energy profile. As can be seen in Fig. 5 *b*, the free energy profiles of POG and POG-A can be made to overlap over states 1–12 by shifting them relative to each other. With a different shift, they overlap over states 18–25. (Recall that a free energy profile can be vertically shifted by an arbitrary amount by changing the constant C. Also note that we compare two free energy profiles, $G(n) = -k_{\mathrm{B}}T \log \phi(n) + C$ for the POG peptide and $G'(n) = -k_{\mathrm{B}}T \log \phi'(n) + C'$ for the POG-A peptide. The constants C and C' could be fixed by specifying a standard state, but that is not necessary for our purpose. The comparison we make is based on the equivalency of states—state n represents the same stage of folding for either POG or POG-A. We compare, for example, $G(7) - G(6)$ and $G'(7) - G'(6)$, for which the values of C and C' are irrelevant. The choice of $C = C' = 0$ is entirely arbitrary and has no implication for our comparison of the two free energy profiles.) This indicates that the effect of the mutations on the free energy is localized on states 13–17. The mutations affect not only the hydrogen bonds 13–15 that are directly connected with the mutant residues (Fig. 1) but also two more hydrogen bonds toward the N-terminus. The free energy profile of POG-A in this region appears to be a barrier, and we interpret it as renucleation. This renucleation requires formation of three stable hydrogen bonds (16–18) just as the initial nucleation at the C-terminal end. Summarizing, we characterize regions of the free energy profile for POG-A as nucleation (states 1–3), zipping (states 3–12), renucleation (states 12–18), zipping (states 18–21), and fraying (states 21–25).

One might expect renucleation to be easier than nucleation, since in renucleation three chains are already held together. In this work, however, we do not address the association process that brings three chains together, as indicated by the exclusion of state 0. That is, the initial nucleation is also considered to take place after three chains have been already associated. Therefore, it is not surprising that our results indicate renucleation is not easier than nucleation (Fig. 5).

Perhaps, it is no coincidence that triple-helix nucleation requires formation of three consecutive hydrogen bonds. As illustrated in Fig. 6 *a*, three hydrogen bonds are needed to hold three chains together, which suggests that three is necessary. And our results show that three is also sufficient, at least in the case of the collagenlike peptides considered here. In a deletion experiment of Bulleid et al. (18), it was found that a minimum of two imino-rich tripeptide units at the C-terminal end are required for nucleation to occur. The rigidity of imino acids stabilizes the collagen triple helix since the ϕ-angles of proline and hydroxyproline are optimal for forming the triple helix structure (12,19,20). As Fig. 6 *b* suggests, one imino-rich tripeptide unit encloses two hydrogen bonds, but the third hydrogen bond is not totally surrounded by the imino-rich environment. Two such units (Fig. 6 *c*), however, enclose five hydrogen bonds. It appears that the result of this deletion experiment is consistent with our conclusion that

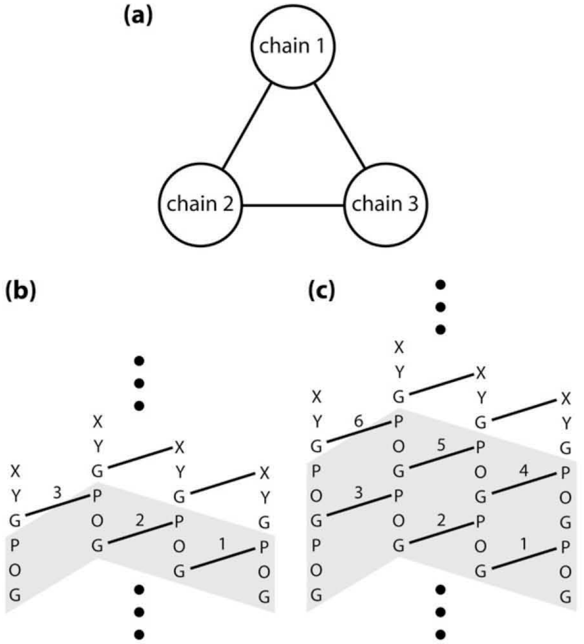

FIGURE 6 Nucleation of the triple helix. (*a*) Nucleation consists of formation of three consecutive hydrogen bonds that hold the three chains together. (*b*) One imino-rich tripeptide unit encloses two hydrogen bonds. (*c*) Two consecutive imino-rich tripeptide units enclose five hydrogen bonds.

triple-helix nucleation consists of formation of three consecutive stable hydrogen bonds.

Kinetics—mean dwell time and mean folding time

To understand the kinetics of the folding of POG and POG-A, we examine two quantities: the mean dwell time $\tau_{\mathrm{dwell}}(n)$ and the mean folding time τ_{fold}. (Note that construction of an MSM erases dynamics at timescales that are shorter than the lag time. Therefore, an MSM can give reliable results for kinetics only if the timescale of interest is much longer than the lag time, which seems to be the case here. Thermodynamic properties are safe from this issue.) The mean dwell time $\tau_{\mathrm{dwell}}(n)$ is the average amount of time the system spends in state n before making a transition to any other state, and indicates the kinetic stability of each state. It is calculated from a transition matrix θ by

$$\tau_{\mathrm{dwell}}(n) = \frac{\tau_{\mathrm{lag}}}{1-\theta(n|n)}. \tag{4}$$

We define the mean folding time to be the mean first passage time from state 3 to state 21,

$$\tau_{\mathrm{fold}} = \tau_{\mathrm{MFPT}}(3 \rightarrow 21), \tag{5}$$

with a reflecting boundary condition imposed at state 3, i.e., $\theta(i|j) = 0$ for $i < 3$ and $j \geq 3$. (Note that we define τ_{fold} to be $\tau_{\mathrm{MFPT}}(3 \rightarrow 21)$, instead of $\tau_{\mathrm{MFPT}}(1 \rightarrow 25)$, to make mutational effects more noticeable. The change in $\tau_{\mathrm{MFPT}}(1 \rightarrow 25)$ due to the Gly → Ala mutations is almost unnoticeable since $\tau_{\mathrm{MFPT}}(1 \rightarrow 25)$ is dominated by nucleation (states 1–3) and fraying (states 21–25).) The mean first passage time $\tau_{\mathrm{MFPT}}(i \rightarrow j)$ is calculated by solving

$$\tau_{\mathrm{MFPT}}(i \rightarrow j) = \begin{cases} \tau_{\mathrm{lag}} + \sum_{\mathrm{k}} \theta(k|i)\tau_{\mathrm{MFPT}}(k \rightarrow j) & \text{for} \quad i \neq j \\ 0 & \text{for} \quad i = j \end{cases}. \tag{6}$$

While the mean dwell time represents local kinetics at each state, the mean folding time represents global kinetics. As in the previous section, we sample 100 transition matrices from the distribution produced by Bayesian inference, and calculate these kinetic quantities from each. We take the median as a point estimate and the 95% interval as an error bar.

Figs. 7 and 8 show the estimates of $\tau_{\mathrm{dwell}}(n)$ and τ_{fold}. To compensate for the low viscosity (1.0/ps) used in our MD simulations, we multiplied each estimate by a factor of 10 as suggested by Zagrovic and Pande (21); the results shown in Figs. 7 and 8 include this correction. While estimates of free energy have more or less converged, estimates of these kinetic quantities have not fully converged with respect to the lag time, although they do show trends of convergence. Accordingly, we only draw conclusions that are invariant over different lag times.

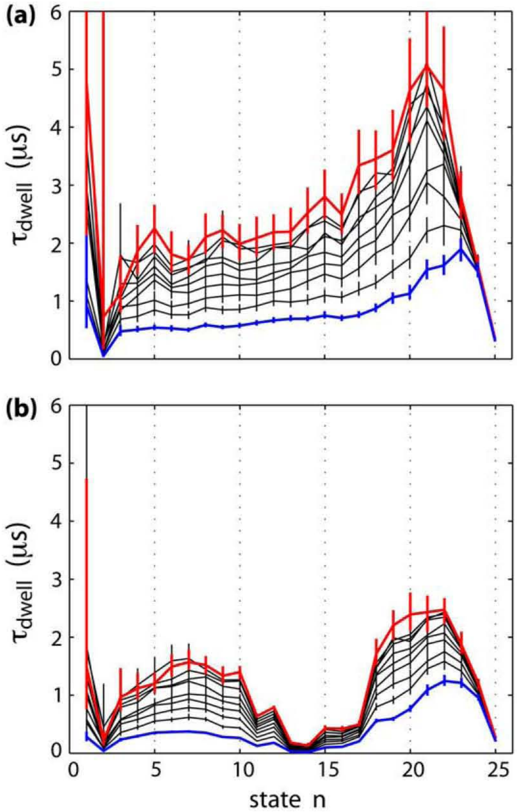

FIGURE 7 Mean dwell time $\tau_{\mathrm{dwell}}(n)$. (*a*) POG. (*b*) POG-A. In each panel, there are 10 graphs of $\tau_{\mathrm{dwell}}(n)$ obtained with 10 different lag times, $\tau_{\mathrm{lag}} =$ 1,2, . . . ,10 ns. The graphs for the shortest (1 ns) and the longest (10 ns) lag times are shown in blue and red, respectively. And the other eight graphs are shown in black. The error bars indicate 95% Bayesian intervals around medians.

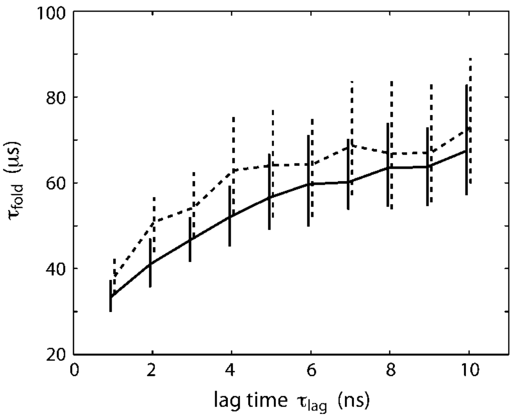

FIGURE 8 Mean folding time $\tau_{\rm fold} \equiv \tau_{\rm MFPT}(3 \rightarrow 21)$ calculated with 10 different lag times, $\tau_{\rm lag} = 1,2,\ldots,10$ ns. (*Solid line*, POG; *dashed line*, POG-A.)

The mean dwell time shown in Fig. 7 indicates the kinetic stability of each state. Comparing with Fig. 5, low free energy regions roughly correspond to long dwell-time regions, and high free energy regions to short dwell time regions. Overall, many states have long dwell times on the order of microseconds. In Thermodynamics—Free Energy Profile, we discussed the stability of the collagen triple helix based on the free energy profile. We get a more complete view of the stability by considering the free energy profile and the mean dwell time together. Free energy decreases only by 0.1 kcal/mol per step during the zipping process, which indicates only marginal stability of hydrogen bonds. In fact, surprisingly low melting temperatures of collagen also points to the same conclusion (22). The mean dwell time, on the other hand, suggests that the system stays in the same state for microseconds. It appears that the stability of the collagen triple helix comes from slow kinetics rather than free energy differences. That is, for collagen, kinetic stability seems to be the dominant factor. A related conclusion was drawn by Miles and Bailey (23).

As can be seen in Fig. 8, the presence of the Gly→Ala mutations increases the mean folding time, although the large error bars make it difficult to estimate the exact amount of increase. In POG-A, due to its imino-rich nature, renucleation occurs immediately after the mutation site. In native collagen containing a mutation, on the other hand, the next available renucleation domain may be located further down from the mutation site. The delay in folding of mutant collagen sequences from patients of *Osteogenesis imperfecta* was found to be on the order of minutes (24). It is natural to expect a correlation between the delay in folding and the distance between the mutation site and the next available renucleation domain, although the precise relationship is unknown (25).

CONCLUSIONS

We have studied the thermodynamics and the kinetics of folding of collagenlike peptides, POG and POG-A, using Markov analyses of MD simulations. We find that the C-to-N zipping of the collagen triple helix must be initiated by a nucleation event consisting of formation of three stable hydrogen bonds, and that zipping through a glycine mutation site requires a renucleation event which also consists of formation of three stable hydrogen bonds. Our results also suggest that slow kinetics, rather than free energy differences, is mainly responsible for the stability of the collagen triple helix. The next step is to explore the folding process of native collagen sequences and examine the effects of mutations associated with genetic diseases, for which the methodology and the outcome of this study will provide useful guidance.

The MSM analysis of MD simulations allows one to extract long time dynamics from short trajectories, and can bridge the timescale gap between simulations and experiments. It is a general tool that can be applied to various problems. Especially, the methodology presented here will find its most natural applications in problems for which a one-dimensional state space can be effectively defined, e.g., DNA zipping/unzipping, formation of coiled coil structures, formation of fibrous structures, etc.

APPENDIX

Generation of Markov chains

State assignment (or state decomposition) is often the most delicate matter in MSM construction. For a MSM to be successful, each state must correspond to a metastable region in the configuration space; otherwise, rapid fluctuations at state boundaries will plague the analysis. In general, however, it is difficult to identify metastable regions a priori. Our state assignment is based on formation of hydrogen bonds, which may not necessarily correspond to metastable regions in configurational space. Therefore, to smooth out rapid fluctuations at state boundaries, we employed the following scheme.

In refolding simulations, coordinates were saved every 100 ps, from which we obtain sequences of states by applying the state assignment rule described in State Assignment. We then divide each sequence into nanosecond blocks; the first block contains the first 10 states, and so on. From each block of 10 states, we pick one state by majority rule. (In case of a tie, we simply exclude that block from the analysis. This did not affect the results considerably since ties were rare.) Consequently, we obtain sequences of states with the time interval of 1 ns, from which we generate Markov chains with respect to various lag times. For example, a Markov chain with respect to $\tau_{\rm lag} = 2$ ns is generated by taking every other state from the 1-ns sequence. The process of picking one state out of 10 by majority rule smoothes out rapid fluctuations at state boundaries.

Bayesian inference of transition probabilities and the choice of the prior distribution

Bayesian inference for the estimation of transition probabilities $\theta(i|j)$ from transition counts $N(j \rightarrow i)$ operates according to Bayes' theorem:

$$P(\theta|N) = \frac{P(N|\theta)P(\theta)}{\int d\theta P(N|\theta)P(\theta)}. \qquad (7)$$

Here $P(\theta)$ is the prior distribution that represents our knowledge of θ before seeing the data N, $P(N|\theta)$ is the likelihood of observing the data N given the parameter θ, and $P(\theta|N)$ is the posterior distribution that represents our estimate of θ based on the prior distribution and the data. The likelihood is given as a multinomial distribution:

$$P(N|\theta)=\prod_{\mathrm{j}}\left\{\frac{[\sum_{\mathrm{i}}N(j\rightarrow i)]!}{\prod_{\mathrm{i}}N(j\rightarrow i)!}\prod_{\mathrm{i}}\theta(i|j)^{\mathrm{N(j\rightarrow i)}}\right\}. \qquad (8)$$

For the prior distribution, we choose a Dirichlet distribution,

$$P(\theta)=\prod_{\mathrm{j}}\left\{\frac{\Gamma(\sum_{\mathrm{i}}\alpha_{\mathrm{ij}})}{\prod_{\mathrm{i}}\Gamma(\alpha_{\mathrm{ij}})}\left[\prod_{\mathrm{i}}\theta(i|j)^{\alpha_{\mathrm{ij}}-1}\right]\delta(\textstyle\sum_{\mathrm{i}}\theta(i|j)-1)\right\}, \qquad (9)$$

where α_{ij} are the parameters that specify a particular Dirichlet distribution. The advantage of using a Dirichlet distribution is that the resulting posterior distribution is also a Dirichlet distribution; namely, as one can verify from the above three equations,

$$P(\theta|N)=\prod_{\mathrm{j}}\left\{\frac{\Gamma(\sum_{\mathrm{i}}\alpha'_{\mathrm{ij}})}{\prod_{\mathrm{i}}\Gamma(\alpha'_{\mathrm{ij}})}\left[\prod_{\mathrm{i}}\theta(i|j)^{\alpha'_{\mathrm{ij}}-1}\right]\delta(\textstyle\sum_{\mathrm{i}}\theta(i|j)-1)\right\}, \qquad (10)$$

where $\alpha'_{\mathrm{ij}}=\alpha_{\mathrm{ij}}+N(j\rightarrow i)$.

We still need to choose α_{ij} to specify a prior distribution. Unless one has a significant amount of prior information on θ, typical choices are broad distributions such as the uniform distribution ($\alpha_{\mathrm{ij}}=1$) and Jeffreys' invariant distribution ($\alpha_{\mathrm{ij}}=1/2$). These distributions allow a nonzero transition probability $\theta(i|j)$ even if the $j\rightarrow i$ transition was never observed [$N(j\rightarrow i)=0$], which is reasonable in most cases since having no transition count in a finite amount of data never guarantees that the transition probability is actually zero. In our case, however, some states (e.g., state 1 and 25) are so far apart from each other that transition probabilities between them should be virtually zero with respect to the range of lag times we consider (1~10 ns). Therefore, we choose $\alpha_{\mathrm{ij}}=0$. This distribution is classified as improper because it is not normalizable. It can be interpreted, however, as a limit of proper distributions: $\alpha_{\mathrm{ij}}\rightarrow 0^{+}$. Under this prior distribution, if $j\rightarrow i$ transition was never observed, the posterior distribution assigns $\theta(i|j)$ to be strictly zero. This choice of prior distribution can be considered as a way of incorporating into the prior distribution our prior information about the system, i.e. that some states are far apart from each other.

We had valuable discussions with Randy Radmer, Dale Bodian, John Chodera, and Dan Ensign. Young Min Rhee and Guha Jayachandran helped us with the use of Folding@Home.

This work was made possible by the participants of Folding@Home who kindly donated their precious CPU time. This work was funded by the National Institutes of Health grants No. AR051582 (T.E.K., principal investigator) and No. GM062868 (V.S.P., principal investigator).

REFERENCES

1. Kielty, C. M., and M. E. Grant. 2002. The collagen family: structure, assembly, and organization in the extracellular matrix. *In* Connective Tissue and Its Heritable Disorders: Molecular, Genetic, and Medical Aspects. P. M. Royce and B. Steinmann, editors. Wiley-Liss, New York.
2. Prockop, D. J., and K. I. Kivirikko. 1995. Collagens: molecular biology, diseases, and potentials for therapy. *Annu. Rev. Biochem.* 64:403–434.
3. Byers, P. H., and W. G. Cole. 2002. *Osteogenesis imperfecta. In* Connective Tissue and Its Heritable Disorders: Molecular, Genetic, and Medical Aspects. P. M. Royce and B. Steinmann, editors. Wiley-Liss, New York.
4. Baum, J., and B. Brodsky. 1999. Folding of peptide models of collagen and misfolding in disease. *Curr. Opin. Struct. Biol.* 9:122–128.
5. Bachinger, H. P., P. Bruckner, R. Timpl, D. J. Prockop, and J. Engel. 1980. Folding mechanism of the triple helix in type-III collagen and type-III pN-collagen: role of disulfide bridges and peptide bond isomerization. *Eur. J. Biochem.* 106:619–632.
6. Singhal, N., C. D. Snow, and V. S. Pande. 2004. Using path sampling to build better Markovian state models: predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.* 121: 415–425.
7. Swope, W. C., J. W. Pitera, and F. Suits. 2004. Describing protein folding kinetics by molecular dynamics simulations. I. Theory. *J. Phys. Chem. B.* 108:6571–6581.
8. Huang, C. C., G. S. Couch, E. F. Pettersen, T. E. Ferrin, A. E. Howard, and T. E. Klein. 1998. The object technology framework: an object-oriented interface to molecular data and its application to collagen. Pacific Symposium on Biocomputing. 98:349–361.
9. Bella, J., M. Eaton, B. Brodsky, and H. M. Berman. 1994. Crystal and molecular structure of a collagenlike peptide at 1.9 Å resolution. *Science.* 266:75–81.
10. Case, D. A., T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, B. Wang, D. A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. W. Caldwell, W. S. Ross, and P. A. Kollman. (2004). AMBER 8. University of California, San Francisco, CA.
11. Wang, J., P. Cieplak, and P. A. Kollman. 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* 21:1049–1074.
12. Park, S., R. J. Radmer, T. E. Klein, and V. S. Pande. 2005. A new set of molecular mechanics parameters for hydroxyproline and its use in molecular dynamics simulations of collagenlike peptides. *J. Comput. Chem.* 26:1612–1616.
13. Onufriev, A., D. Bashford, and D. A. Case. 2004. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins Struct. Funct. Bioinform.* 55:383–394.
14. Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* 23:327–341.
15. Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. Bayesian Data Analysis, 2nd Ed. CRC Press, Boca Raton, FL.
16. Jaynes, E. T. 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK.
17. Park, S., and V. S. Pande. 2006. Validation of Markov state models using Shannon's entropy. *J. Chem. Phys.* 124:054118.
18. Bulleid, N. J., J. A. Dalley, and J. F. Lees. 1997. The C-propeptide domain of procollagen can be replaced with a transmembrane domain without affecting trimer formation or collagen triple helix folding during biosynthesis. *EMBO J.* 16:6694–6701.
19. Josse, J., and W. F. Harrington. 1964. Role of pyrrolidine residues in structure stabilization of collagen. *J. Mol. Biol.* 9:269.
20. Vitagliano, L., R. Berisio, L. Mazzarella, and A. Zagari. 2001. Structural bases of collagen stabilization induced by proline hydroxylation. *Biopolymers.* 58:459–464.
21. Zagrovic, B., and V. Pande. 2003. Solvent viscosity dependence of the folding rate of a small protein: distributed computing study. *J. Comput. Chem.* 24:1432–1436.
22. Leikina, E., M. V. Mertts, N. Kuznetsova, and S. Leikin. 2002. Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci. USA.* 99:1314–1318.
23. Miles, C. A., and A. J. Bailey. 2004. Studies of the collagenlike peptide (Pro-Pro-Gly)$_{10}$ confirm that the shape and position of the type I collagen denaturation endotherm is governed by the rate of helix unfolding. *J. Mol. Biol.* 337:917–931.
24. Raghunath, M., P. Bruckner, and B. Steinmann. 1994. Delayed triple helix formation of mutant collagen from patients with *Osteogenesis imperfecta. J. Mol. Biol.* 236:940–949.
25. Hyde, T. J., M. A. Bryan, B. Brodsky, and J. Baum. 2006. Sequence dependence of renucleation after a Gly mutation in model collagen peptides. *J. Biol. Chem.* 281:36937–36943.
26. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD—Visual Molecular Dynamics. *J. Mol. Graph.* 14:33–38.