JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Modeling Structural Flexibility of Proteins with Go-Models

Ping Jiang* and Ulrich H. E. Hansmann*

Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019-5251, United States

ABSTRACT: Structure-based models are an efficient tool for folding studies of proteins since by construction their energy landscape is only minimally frustrated. However, their intrinsic drawback is a lack of structural flexibility as usually only one target structure is employed to construct the potentials. Hence, a Go-model may not capture differences in mutation-induced protein dynamics, if—as in the case of the disease-related A629P mutant of the Menkes protein ATP7A—the structural differences between mutant and wild type are small. In this work, we introduced three implementations of Go-models that take into account the flexibility of proteins in the NMR ensemble. Comparing the wild type and the mutant A629P of the 75-residue large sixth domain Menkes protein, we find that these new Go-potentials lead to broader distributions than Go-models relying on a single member of the NMR ensemble. This allows us to detect the transient unfolding of a loosely formed $\beta1\beta4$ sheet in the mutant protein. Our results are consistent with previous simulations using a physical force field and an explicit solvent and suggest a mechanism by which these mutations cause Menkes disease. In addition, the improved Go-models suggest differences in the folding pathway between the wild type and mutant, an observation that was not accessible to simulations of this 75-residue protein with a physical all-atom force field and explicit solvent.

INTRODUCTION

Proteins are the workhorses in cells carrying out or regulating many of their biological processes. For a large class of proteins, their biological functions depend on them folding into a specific three-dimensional structure, and consequently, misfolding is associated often with diseases. Examples are Alzheimer's, Parkinson's, and type II diabetes.[1] Despite the experimental difficulties, X-ray crystallography or NMR led over the past decade to an impressive collection of high-resolution structures in the Protein Data Bank (PDB). However, knowledge of structure is not always sufficient for understanding protein function. Take as an example the copper-transporting ATPase encoded by the ATP7A gene on the X chromosome.[2] The single mutation A629P (replacing an alanine by a proline) can cause Menkes disease, a copper deficiency disease which leads in most cases to death in early childhood. This mutation is located in the sixth domain (all of which are copper-binding) of the cytosolic N-terminus of the 1500-residue transmembrane protein. The structure of this 75-residue large domain has been resolved by NMR for both the wild type (WT) and mutant (MT). An overlay of both structures is shown in Figure 1a. Shown are the central configurations in the corresponding NMR ensembles, i.e., the ones which have the lowest average root-mean-square deviation (RMSD) to all other configurations of their ensemble. Both structures resemble each other closely, and the RMSD between them is comparable to that between configurations in either the wild type or mutant ensemble. However, despite the close similarity of the two structures, there is a difference in function. The associated subtle conformational changes and dynamical processes that lead to the mutant of Menkes disease are difficult to resolve in experiments.

Computational studies can complement experiments in probing such processes,[3,4] but their application is often restricted by sampling difficulties[5] resulting from the rugged energy landscape of proteins. Two approaches are used to overcome these numerical difficulties. One is to use enhanced sampling techniques such as generalized-ensemble and biased-potential methods.[6] The other one relies on simplified models such as coarse-grained and structure-based Go models.[7] Both approaches aim at obtaining sufficient sampling in the conformational space within feasible time: efficient sampling algorithms "flatten" the energy landscape by reducing energy barriers, while simplified models lead by design to an energy landscape with a reduced number of energy valleys.

In our previous studies of the wild type and mutant of Menkes protein, we used the first approach relying on all-atom "physics-based" force fields and explicit solvent molecules. We found that for the mutant, the C-terminal β strand, where the mutation is located, has on average a reduced number of hydrogen bonds with the N-terminal strand than found in the wild type. The resulting more open structure increases the solvent-accessible area in the mutant as compared to the wild type.[8] This suggests a mechanism for the mutation-induced pathology: the frequent, partial solvent exposure of hydrophobic residues on the N-terminus makes the mutant prone to degradation, which in turn yields a low effective concentration of the copper transporting protein. In the present paper, we want to investigate our system with more cost-effective minimal models such as the structure-based Go model[9] developed by the Onuchic group.[10] This model with its by design minimal frustrated energy landscape has been shown to reproduce thermodynamic and kinetic properties of protein folding.[7,11−14] Our interest is 2-fold: first, we want to derive a formulation for Go-models that can capture the dynamic differences between the wild type and mutant arising despite their large similarity, and we want to confirm our previous results with a different energy function that allows for larger statistics. In this way, we hope to verify the correctness of our previous results.

Received: January 25, 2012
Published: May 9, 2012

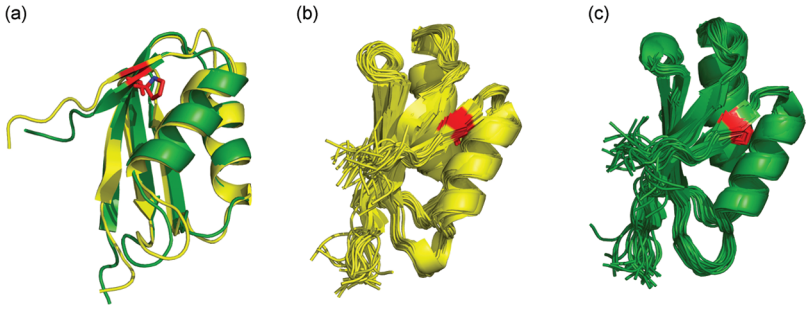

Figure 1. (a) Overlay of wild type (yellow) and mutant (green) of the N-terminal 75-amino-acid segment of Menkes protein, with the residue ALA69 and the mutant PRO69 highlighted by red sticks. The shown structures are the centroids of the respective NMR ensembles: the 27th model of the wild type and the fifth model of the mutant. The root-mean-square deviation (RMSD; over all heavy atoms) between the two structures is 2.3 Å. (b) Overlay of all 30 structures of the wild type NMR ensemble, with residue ALA69 marked in red. The pairwise, heavy-atom RMSD is between 1.6 and 3.2 Å. (c) Overlay of the 30 structures of the mutant NMR ensemble, with residue PRO69 marked in red. The pairwise, heavy-atom RMSD is again between 1.6 and 3.2 Å.

The first point is of general importance: a structure-based Go-model needs as a prerequisite the native structure. However, NMR experiments lead to an ensemble of structures that are compatible with the measured NOEs and chemical shifts, and it is not obvious how to setup a Go-model if there are more than one possible target structures available. Often, the differences between the structures are minor and will not change noticeably the outcome of computational protein studies. However, in the case of Menkes protein, the NMR data for both the wild type and mutant led to two ensembles each with 30 member structures, where the differences between ensembles (wild type vs mutant) and within an ensemble are comparable. This makes it difficult to choose representative structures. This is because the variance between structures within an ensemble contains information on the dynamics of the protein that gets lost by selecting a single representative for constructing the Go-model. As it is likely that the differences in dynamics lead to the loss of function in the mutant, the Go-model needs to be constructed carefully to capture these differences. For this reason, we compare in the present study a "naive" Go-model that simply relies on the NMR structure with the highest confidence score (model 1 of the NMR ensemble), with three implementations that take explicitly the ensemble character into account. The details will be described in the Methods section. We first compare the differences between the various Go-model implementations, before in the second part discussing the different transition paths as observed in simulations of an optimized Go-model of the wild type and mutant. Our results are compared with the previous ones of ref 8.

METHODS

Go-Model Implementations. Our simulations rely on the structure-based model SMOG (Structure-based MOdels in Gromacs), developed by the Onuchic group.[10,12,15] SMOG generates a structure-based energy function for proteins based on the native contacts (NC), i.e., the contacts found in the target structure. In our case, we consider all contacts between heavy atoms. The potential energy function is defined by

$$
\begin{aligned}
V = & \sum_{\text{bonds}} \varepsilon_r (r - r_0)^2 + \sum_{\text{angles}} \varepsilon_\theta (\theta - \theta_0)^2 \\
& + \sum_{\text{dihedral(impr/planar)}} \varepsilon_\chi (\chi - \chi_0)^2 \\
& + \sum_{\text{dihedral(backbone)}} \varepsilon_{BB} F_D(\phi) + \sum_{\text{dihedral(sidechain)}} \varepsilon_{SC} F_D(\phi) \\
& + \sum_{\text{contacts(long)}} \varepsilon_C \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r} \right)^{6} \right] \\
& + \sum_{\text{non-contacts(long)}} \varepsilon_{nC} \left[\left(\frac{\sigma_{nC}}{r} \right)^{12} \right]
\end{aligned}
\tag{1}
$$

where

$$
F_D(\phi) = [1 - \cos(\phi - \phi_0)] + \frac{1}{2}[1 - \cos(3(\phi - \phi_0))] \tag{2}
$$

Bonded interactions including bond lengths, angles, and improper dihedral angles (the first three terms (line 1 and line 2) in eq 1) are described by harmonic potentials with the equilibrium values derived from the target structure. In a similar way are the torsion energies for the flexible dihedrals described by eq 2 and the fourth and fifth terms (line 3) of eq 1. Long-range (in sequence space) contact atom pairs i and j ($i > j + 3$) are described by a Lennard-Jones potential (sixth term (line 4) in eq 1), whereas all of the other nonlocal interactions are assumed to be repulsive (last term in eq 1). The interaction strength parameters for each term are given in ref 15. The topology and coordinate files of the protein were generated by the *SMOG@ctbp* server. These files along with a setting file containing run-control parameters are sufficient for a simulation with this model in the GROMACS distribution. In our case, replica exchange molecular dynamics (REMD) simulations are performed using the GROMACS 4.5.1 package.[16] The time step is 0.0005. Fourteen parallel replicas are used for both wild type and mutant simulations and each Go-model variant. The width of the temperature range is between 5 and 10, and temperature distribution has been refined after one or more trial runs. Note that our energy function is nonphysical, and therefore time and temperatures are given in reduced quantities.

Nuclear magnetic resonance (NMR) can provide extensive structural information about the native state of proteins. Due to

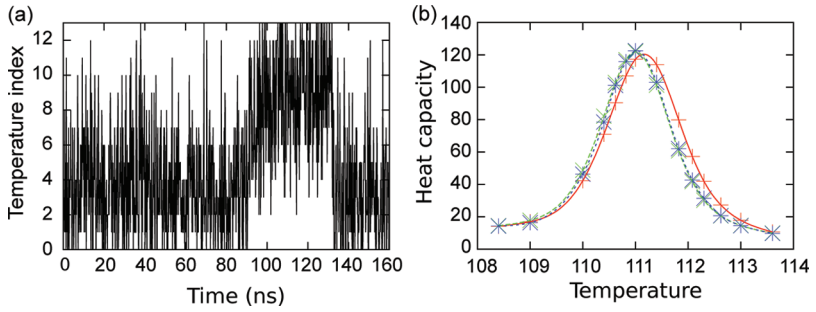

Figure 2. (a) A single replica traveling through temperature space. (b) Heat capacity as a function of temperature. Note that all quantities are given in reduced units. The three heat capacity curves differ in the time over which they were evaluated. The first one (red) is evaluated over the whole length of the simulation. In the second one (green), the first third of the simulation is omitted, and in the last one (blue), only the last third of the simulation is considered. The last two curves (green and blue) overlap, indicating that the system has reached equilibrium.

the limitations of these technologies and the intrinsically flexible nature of proteins, the native state of protein is usually given as an ensemble of structures. For instance, the wild type and mutant of Menkes protein, deposited in the Protein Data Bank (PDB) under identifiers 1YJV and 1YJT, are given as ensembles of 30 structures. In order to build a structure-based energy function, a target structure must be provided. As for the wild type, the first model is the one with highest confidence level; we choose as our baseline a Go-model based on this structure. A better description of the ensemble is given by the centroid structure. Hence, in our second Go-model, we use as input the centroids of wild type (the 27th NMR model) or mutant (the fifth NMR model) ensemble. Since the ensemble of structures is not necessarily centered around a single structure, we have constructed two more Go-potentials by taking into account native contacts from additional structures. Our third Go-potential is derived from the second model by removing the repulsive interactions of all native contacts found in any of the other 29 structures of the respective NMR ensembles. In the fourth version, we go a step further and add additional Lennard-Jones potentials to atom pairs corresponding to these native contacts found in the other 29 structures of the respective ensembles. This generates a Hamiltonian with no bias toward any specific structure of the ensemble. By introducing multiple Lennard-Jones potentials for atom pairs that appear in more than one NMR structure (e.g., 107 atom pairs are shared by all models of the wild type NMR ensemble), the equilibrium distance of an atom pair shifts toward the longest distance appearing in any of the structures, a known problem when constructing multifunnel Go-models.[12,17,18] In order to evaluate the impact of this issue in our case, we have considered these 107 common atom pairs shared in all 30 wild type NMR models. The distances between atoms in these pairs are on average 3.5 ± 0.2 Å, with the longest and shortest distance differing by less than 1 Å. Hence, the largest possible shift of equilibrium distance would be from 3 to 4 Å, which would change little the atom pairs' nature as native contacts. In summary, for the wild type, four models are generated and denoted by WT_1 (first model as reference), WTr (centroid as reference), WTnr (no repulsive for native contacts found in noncentroid structures), and WTa (Lennard-Jones term for all native contacts in all structures of the NMR ensemble). Similarly, the corresponding models for the mutant are denoted by MTr, MTnr, and MTa.

■ ANALYSIS METHODS

Because the Go model is not a physical model, the temperature does not necessarily correlate with the experimental temperature and depends on the target structure. For this reason, we compare results for the various Go-model realizations of the mutant and wild type at their respective folding temperatures T_f, defined for a given system as the temperature where unfolded and folded states have equal free energy. This temperature is also the one where the specific heat defined by

$$C_V = \frac{d\langle E\rangle}{dT} = (\langle E^2\rangle - \langle E\rangle^2)/RT^2 \tag{3}$$

is maximal. Here, $\langle E\rangle$ is the average of the total energy E. R is the gas constant 8.314472 J·K^{-1}·mol^{-1}. Due to the lack of physical meaning of temperature here, the unit of C_V is denoted as a.u.

We used two measures to describe the similarity between structures. The first is the root-mean-square-deviation (RMSD) between two structures. The other one is the dissimilarity score ∇:

$$\nabla = (1 - Q_c/Q_i)(1 - Q_c/q_j) \tag{4}$$

where Q_c is the number of shared native contacts between structure j of a given simulation and reference structure i. Q_i is the number of native contacts in the reference structure i. q_j is the number of native contacts in structure j of our simulation. For every j, the closest reference structure i is determined by the smallest ∇. A cutoff ($\nabla < 0.4$) was used to exclude unfolded configurations.

Convergence of Replica-Exchange Molecular Dynamics Simulations. While Go-models lead by construction to energy landscapes with minimal frustration, convergence can still be slow for large proteins. Replica exchange techniques,[19,20] first introduced to protein studies in ref 21, can help to overcome these sampling difficulties. As we are interested in the folding path of the Menkes protein, we distribute the temperatures in replica-exchange molecular dynamics simulations around the folding temperature (T_f). For this purpose, we start with a broad temperature distribution to search for the critical temperature where the specific heat is maximal. The trial simulations last for each system for at least 100 ns. Afterwards, the temperature distributions are optimized[22] and each system restarted with the initial structures derived from a similar temperature and continued for 110−250 ns. A given replica, as for instance shown in Figure 2a (here for the mutant), travels through all temperatures several times within the 160 ns of

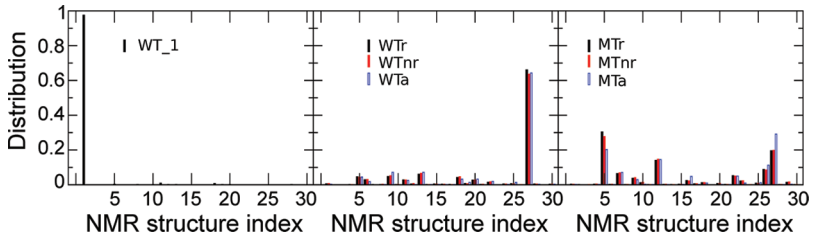

Figure 3. Distribution of folded configurations in our simulations. Configurations are clustered into 30 clusters according to their similarity to the 30 NMR models in the corresponding ensemble. The measure of similarity is the dissimilarity score $\bigtriangledown$ of eq 4. The frequencies for each of the 30 clusters are shown for the various Go-models for the wild type (left and middle) and the mutant (right).

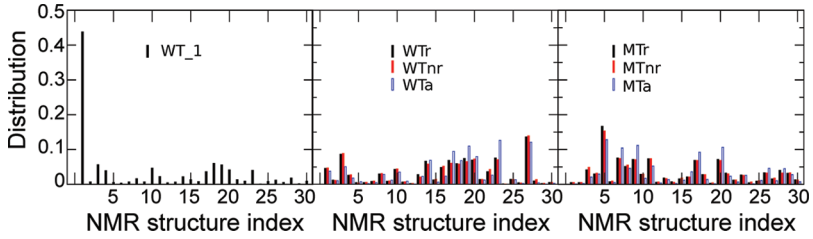

Figure 4. Distribution of folded configurations in our simulations. Configurations are clustered into 30 clusters according to their similarity to the 30 NMR models in the corresponding ensemble. The measure of similarity is the RMSD. The frequencies for each of the 30 clusters are shown for the various Go-models for the wild type (left and middle) and the mutant (right). A structure is considered folded if its RMSD to any of the NMR structures is less than 5 Å.

simulation time. All 14 replicas together fold and unfold 18 times per 100 ns. This number of events indicates that we have sufficient statistics for our analysis, and the simulation reached equilibrium. This can be seen also from Figure 2b where we show the specific heat as evaluated for three different time blocks. The curve has a single peak around the critical temperature (approximately 111 K) where our proteins undergo conformational reorganization. Note that the position of the peak does not change after the initial preparation time, indicating that equilibrium has been achieved. The results of the other six systems are not shown here, but they all exhibit qualitatively the same well-converged behaviors.

RESULTS AND DISCUSSION

Distributions of Conformations for Different Go-Models. The purpose of employing a Go model is to fold proteins in reasonable time. We consider a protein as folded if either its heavy-atom RMSD to one of the 30 structures in the NMR ensemble is smaller than 5 Å or if the dissimilarity score of eq 4 $\bigtriangledown$ < 0.4. The ensemble of folded configurations is clustered into 30 groups by determining for a given configuration the closest structure from the NMR ensemble. The so obtained distributions are plotted in Figures 3 and 4. Note that we consider for our analysis only configurations where the replica has previously at least once been unfolded. This is to avoid bias from the initial folded start configuration.

In Figure 3, the distribution of cluster frequencies demonstrates a pronounced prevalence of the target structure in WT_1 simulations (left figure), where the first structure from the wild type ensemble is chosen for constructing the Go-model. The cumulative statistics of the 29 other clusters is about 2.6%, leaving the first structure to dominate overwhelmingly the ensemble of configurations. When the centroid is used to define the Go-model (WTr), the cumulative occurrence of other structures increases to 34% (see the figure in the middle of Figure 3). Albeit the occurrence of the target structure is still high (66%), it is comparable to the WTa system where native contacts from all 30 NMRs contribute equally. Hence, the observed low frequency of model 1 (the one with highest confidence level in the NMR ensemble) in Figure 3 is likely not due to the construction of our Go-model. We rather believe that the centroid describes better the protein in cases where the confidence level of model 1 is not significantly higher than that of the other models. In the mutant simulations (Figure 3 right graph), we find for the corresponding case MTr, where the centroid of the mutant NMR ensemble defines the Go-model, that competing structures appear with higher occurrences than for the wild type. The cumulative occurrence of noncentroid structures reaches up to 70% for the WTnr model and is even higher (80%) in WTa (where attractive contacts from all 30 structures of the wild type NMR ensemble are used for constructing the model). These additionally observed structures are not necessarily the ones closest to the centroid, such as the 27th structure which has a higher occurrence than the centroid in MTa.

A similar picture is seen when the structural similarities between conformations are evaluated using the RMSD, see Figure 4. For WT_1, the fact that the first structure is the overwhelmingly dominant one is not changed although its population is below 50%. For the ensemble-based Go-models (middle and right graphs), the distributions become broader when the structures are clustered using RMSD than when using the dissimilarity score of eq 4 in Figure 3. The over-representation of the centroid structure in the middle panel of Figure 3 compared to the one in Figure 4 may indicate that a contact-based similarity measure is not accurate enough to distinguish the folds of this protein, and the RMSD may provide a more accurate picture. However, both Figure 3 and Figure 4 lead to the qualitativley same picture with the centroid appearing with higher frequency than the first model, and a broad distribution of other noncentroid structures occurring with similar frequency.

Comparing both figures, we conclude that the use of a central structure as a reference (WTr) gives a broader distribution of

structures than choosing the first structure (WT_1), albeit that structure has the highest confidence level. Setting the energy to zero for native contacts appearing in other than the centroid (WTnr and MTnr) leads for both the mutant and wild type to broader distributions than observed in the case where these contacts are treated as repulsive (WTr and MTr). Further broadening of the distribution of sample configurations is obtained by treating these additional contacts as attractive (WTa and MTa). However, the resulting frustration from the competing interactions leads to a reduced number of folding−unfolding events. The resulting reduction in sampling makes it more efficient to use the variants WTnr or MTnr in simulations of these proteins. It is likely that this result can be generalized to other proteins; i.e., the most efficient way of incorporating the full information of the NMR ensemble into a Go-model is the use of centroid structure for Go-model construction by turning off at the same time the repulsive interactions between atoms that form native contacts in other structures of the NMR ensemble. We have chosen this approach in our following comparison of the wild type and mutant.

Competitive and Cooperative Interactions during Folding. The wild type and mutant of the 75-residue domain are built out of the same six secondary structure elements denoted by $\alpha1$, $\alpha2$ for two helices, and $\beta1-\beta4$ for four β strands. The numbering is from the N to C terminus. As shown in Figure 1, the mutant and wild type share the same spatial arrangement of secondary structure elements leading to a RMSD between the two structures which is comparable to the fluctuation within the associated NMR ensembles. We have argued in previous work[8] that in the mutant the C-terminal strand has an increased flexibility, allowing transient unfolding of the protein which partially exposes hydrophobic residues. This makes the protein prone to degradation, leading to a low effective concentration of the copper transporting protein. A simple structure-based Go-model cannot describe these differences given the minuscule structural differences. We conjecture that a Go-model which takes information on the whole ensemble (and therefore on the structural fluctuations) into account could resolve the differences in dynamics between the wild type and mutant. Hence, following our discussion in the previous section, we have simulated the wild type and mutant of this protein using the Go-models WTnr and MTnr.

In Figure 5a, we show for both the wild type and mutant at their folding temperature the distribution of configurations as function of the end-to-end distance. As the N-terminal and C-terminal segments form a β-sheet, the end-to-end distance is related to the stability of the sheet. While the effect is small, the figure shows that the mutant has a lower propensity for states with small end-to-end distances (i.e., fully formed β-sheet) than the wild type, and a higher propensity for configurations with large end-to-end distances (where the β-sheet is dissolved). Figure 5b supports this picture. The mutant has a larger number of configurations, where the N- and C-terminal segments do not form a sheet, than the wild type, while configurations with such a sheet exist more frequently for the wild type than for the mutant. These results are similar to what we observed earlier in simulations of the protein with the Amber force field and explicit solvent.[8] Hence, our simulations with a Go-model that takes into account information on the whole NMR ensemble instead of only a single structure, support the conclusions of this earlier work using a different model. At the same time, they demonstrate the possibility of Go-models to probe such dynamic effects when the model is correctly formulated.

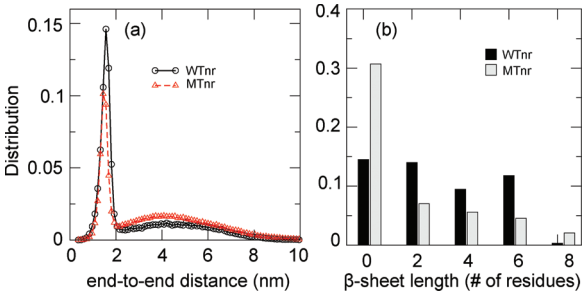

Figure 5. Distribution of end-to-end distances (a) and the length of the $\beta1\beta4$ sheet (b). Because of the high flexibility of N and C termini, we calculate the distance between the starting residue on strand $\beta1$ and on strand $\beta4$ instead of the first and last residues of the protein. The secondary structure content of the $\beta1$ and $\beta4$ strands was assessed by the DSSP package.[23]

Replica exchange molecular dynamics cannot generate a continuous trajectory of protein dynamics at a constant temperature of interest. Therefore, for tracking the folding process, it is crucial to find other folding coordinates than simulation time. A two-dimensional free energy landscape can be built with the total fractional number n_t of native contacts as one coordinate and the ones that signal interactions between certain secondary structure elements as its second coordinate n_s. Figure 6 shows as an example the free energy landscapes of the mutant and wild type where n_s marks the fractional number of native contacts formed between $\beta1$ and $\beta4$ strands. There is a high energy barrier around the transition states when n_t is between 0.2 and 0.4, indicating that the protein is a two-state folder. Note that the free energy barrier is lower for the mutant than for the wild type, allowing for easier transition between folded and unfolded states. This is in agreement with our above observation and supports again our results in ref 8. Note also that when the wild type and mutant are in the ensemble of unfolded states, $n_t < 0.2$, there is a slightly lower probability for states with $n_s > 0.3$ in the mutant than observed for the wild type. This indicates that in the mutant the proline on strand $\beta4$ interrupts the end-to-end interactions and reduces the chances of early contacts within the $\beta1\beta4$ sheet.

In order to understand better its dynamics, we describe the folding process in the following by the quantity n_s (the number of native contacts between a given pair of secondary structure elements) as a function of the number of total contacts n_t. For a given value of n_t, the most likely value of n_s is where the free energy $G(n_s)$ is minimal (for this value of n_t). Hence, in Figure 6, and similar two-dimensional free energy landscapes for other pairs of secondary structure elements, the manifold of such points defines a set of functions $f_s(n_t)$ that traces the most likely folding path (characterized by the interaction between pairs of secondary structure elements) as a function of our order parameter n_t. These are shown in Figure 7 for both the wild type and mutant. Note that the curves are smoothened for better visibility. Representative structures along these defined folding paths are shown in Figure 8.

Both the interactions within secondary structure elements and the ones between such secondary elements increase along the folding process, but α and β secondary structure elements have divergent growing behaviors. The growth of β sheets is a sigmoidal curve characterized by a short lag phase (the overall $n_t < 0.1-0.3$) followed by an exponential increase (Figure 7a,b).

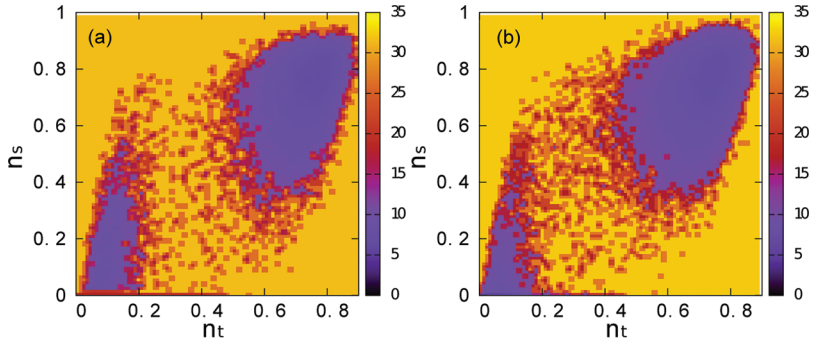

Figure 6. Free energy landscapes showing the growth of the $\beta1\beta4$ sheet with overall folding of the protein at the transition temperature. We denote with n_s the fractional number of native contacts formed between $\beta1$ and $\beta4$ strands, and with n_t the total fractional number of all native contacts. Data are taken from wild type simulations relying on the WTnr model (a) and MTnr for the mutant (b).

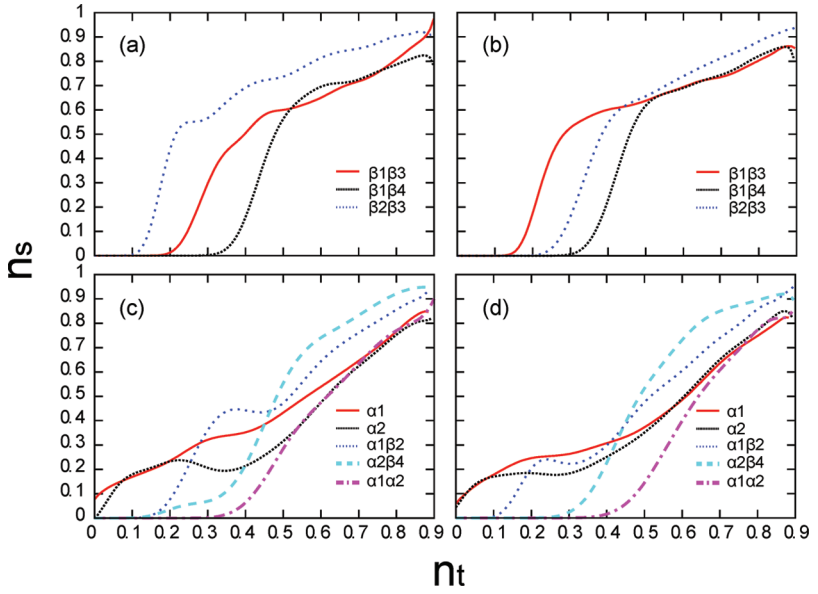

Figure 7. Folding path of the wild type and mutant using the interactions between secondary structure elements as marker n_s. The curves connect the points where for a given number of native contacts n_t the free energy as function of n_s is minimal. Data are taken from wild type simulations relying on the WTnr model (a, c) and MTnr for the mutant (b, d).

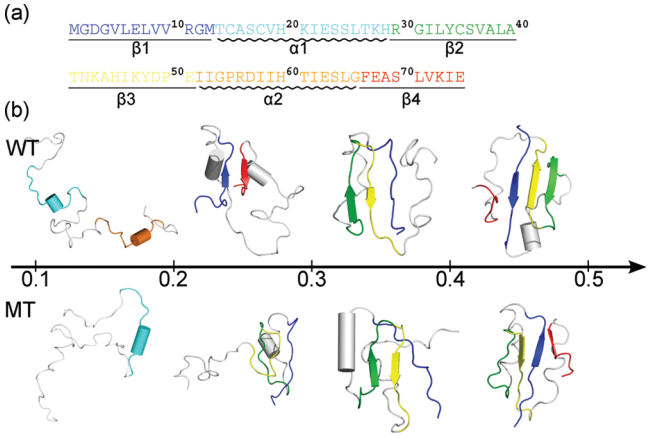

Figure 8. (a) The primary sequence of the mutant. Colors differentiate secondary structure elements. (b) Representative snapshots along the reaction coordinate, the normalized number of native contacts n_t. The parts that undergo significant conformational changes at a given value of n_t are colored. The coloring scheme is consistent with a. The structures are in cartoon representation, where small cylinders indicate the adoption of α-helices.

The high energy barrier observed during β-sheet formation is attributed to the existence of the lag phase. On the other hand, α-helices grow at a steady rate, and they emerge from the start, with the overall number of native contacts still small (Figure 7c,d), and start forming quickly for $n_t = 0.1-0.2$ (Figure 8). It is plausible to assume that the formation of the α-helix which separates two interacting β-strands has to take place before the formation of the β-sheet. Note also that the α-helices are only loosely formed before the start of β-sheet formation as indicated by their low n_s (Figure 7c,d). When the fast growth of β-structures begins, when n_t is between 0.2 and 0.4, α-helices stabilize and undergo only slowly conformational reorganizations without significantly growing. Once the β-sheet takes shape ($n_t > 0.5$), all intersecondary-structure interactions including those between $\alpha1$ and $\beta2$, between $\alpha2$ and $\beta4$, and between $\alpha1$ and $\alpha2$ begin to form (Figure 7c,d). Obviously, the emergence of these higher-level interactions requires the formation of the individual α and/or β components.

Earlier studies have shown that the folding of a three-strand β-sheet protein is a cooperative process where the β-hairpin stabilizes the addition of the third β-strand.[24−27] Similarly, we find that for the wild type of the Menkes protein, the stepwise

dx.doi.org/10.1021/ct3000469 | J. Chem. Theory Comput. 2012, 8, 2127−2133

formation of individual β-strands is a well-ordered process where the end of the formation of the $\beta2\beta3$ sheet marks the start of the growth of the $\beta1\beta3$ sheet. In turn, the $\beta1\beta4$ only starts growing after the $\beta1\beta3$ sheet is formed. In the mutant, the order in which these β-ladders (β-sheet with only two β-strands) are formed is switched. Formation of the $\beta1\beta3$-ladder precedes here that of the $\beta2\beta3$-ladder. We speculate that in the early stage of folding, when distant segments collapse due to the hydrophobicity-driven interactions, a competition exists between strands $\beta3$ and $\beta4$ for forming contacts with strand $\beta1$. Figure 6a and b indicate that in the ensemble of unfolded states there is for the wild type a larger probability for the formation of native contacts between strand $\beta1$ and $\beta4$ than seen for the mutant. These transient contacts in the wild type may decrease the propensity of the $\beta1$ strand to form contacts with the $\beta3$ strand and cause in this way the preference for the formation of the $\beta2\beta3$ ladder. On the other hand, the lack of $\beta1\beta4$ contacts allows in the mutant the $\beta1$ strand to form early contacts with strand $\beta3$. In turn, it delays the formation of the $\beta2\beta3$ ladder. Note that in both the wild type and mutant the $\beta1\beta4$ ladder forms last. This is common for β-sheets with terminal segments as strands, and due to the increased flexibility of such terminal strands.

CONCLUSIONS

We have demonstrated that SMOG is an efficient tool to study the folding of the 75-amino-acid Menkes protein, related to the outbreak of the copper-deficiency Menkes disease. The structural variations between the wild type and mutant are small, indicating that the functional deviations are caused by differences in dynamics and flexibility. Related to the flexibility of the protein is the variance between configurations in the NMR ensemble. For a numerical study of the wild type and mutant, relying on Go-models, it is therefore important to include the ensemble information in the construction of the model. In the current study, we have evaluated three implementations of Go potentials that differ in how ensemble information is included in the construction of the model. All of them yield a broader spectrum and better reoccurrence of NMR structures than a Go-model based on a single structure (the NMR configuration with highest confidence level). Although the implementation that uses native contacts from all structures in the NMR ensemble led to the best results, numerically more effective was to use the centroid for constructing the NMR model, and not adding a repulsive term for native contacts appearing in other NMR structures. The addition of explicit attractive terms led to frustration and reduced the frequency of transition. However, we speculate that in NMR ensembles with multiple distinct clusters of configurations, the centroid structure of each cluster or even all of the NMR structures may be required to construct an adequate Go potential.

Using this modified Go model, we compared the folding path of both the wild type and mutant. We reproduced previous results[8] relying on a different force field and demonstrated that in the mutant the $\beta4$ strand is more loosely connected to the $\beta1$ strand than in the wild type. The resulting transient unfolding leads to partial exposure of hydrophobic residues that makes the mutant prone to degradation. In turn, this leads to the low effective concentration of the copper transporting protein that is responsible for the pathology of Menkes disease. We further show that the differences in the binding affinities between the two terminal strands alter the folding mechanism for the mutant where the order that secondary structure elements form contacts between each other differs from that in the wild type. In principle, these differences should be accessible in experiments.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pingj@ou.edu; uhansmann@ou.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported, in part, by research grant GM62838 of the National Institutes of Health (U.S.A.).

REFERENCES

(1) Selkoe, D. *Nature* **2003**, *426*, 900−904.
(2) Vulpe, C.; Levinson, B.; Whitney, S.; Packman, S.; Gitschier, J. *Nat. Genet.* **1993**, *3*, 7−13.
(3) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751−758.
(4) Shakhnovich, E. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 11823−11824.
(5) Hansmann, U. H. E. In *Multiscale Approaches to Protein Modeling*; Kolinski, A., Ed.; Springer: New York, 2011; pp 209−230.
(6) Zimmermann, O.; Hansmann, U. H. E. *Biochim. Biophys. Acta* **2008**, *1784*, 252−258.
(7) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.
(8) Kouza, M.; Gowtham, S.; Seel, M.; Hansmann, U. H. E. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11390−11397.
(9) Ueda, Y.; Taketomi, H.; Gō, N. *Biopolymers* **1978**, *17*, 1531−1548.
(10) Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Nucleic Acids Res.* **2010**, *38*, 657−661.
(11) Turjanski, A. G.; Gutkind, J. S.; Best, R. B.; Hummer, G. *PLoS Comput. Biol.* **2008**, *4*, e1000060.
(12) Lammert, H.; Schug, A.; Onuchic, J. N. *Proteins* **2009**, *77*, 881−891.
(13) Zuo, G.; Wang, J.; Wang, W. *Proteins* **2006**, *63*, 165−173.
(14) Shea, J.; Onuchic, J.; Brooks, C. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 12512−12517.
(15) Whitford, P. C.; Noel, J. K.; Gosavi, S.; Schug, A.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Proteins* **2009**, *75*, 430−441.
(16) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(17) Best, R. B.; Chen, Y.-G.; Hummer, G. *Structure* **2005**, *13*, 1755−1763.
(18) Okazaki, K.-i.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11844−11849.
(19) Geyer, C. J.; Thompson, E. A. *J. Am. Stat. Assoc.* **1995**, *90*, 909−920.
(20) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604−1608.
(21) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.
(22) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.
(23) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(24) Dill, K. A.; Fiebig, K. M.; Chan, H. S. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 1942−1946.
(25) Griffiths-Jones, S.; Searle, M. *J. Am. Chem. Soc.* **2000**, *122*, 8350−8356.
(26) Sharman, G.; Searle, M. *J. Am. Chem. Soc.* **1998**, *120*, 5291−5300.
(27) Mohanty, S.; Hansmann, U. H. E. *Biophys. J.* **2006**, *91*, 3573−3578.