THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

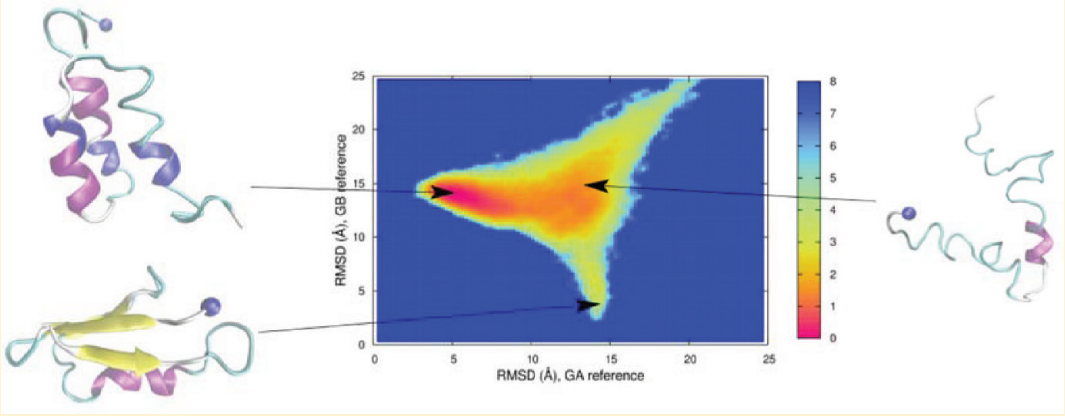
pubs.acs.org/JPCB

Folding Simulations of the A and B Domains of Protein G

Maksim Kouza*,† and Ulrich H. E. Hansmann*,†,‡

†Department of Physics, Michigan Technological University, Houghton, Michigan 49931, United States

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



ABSTRACT: We study wild type and mutants of the A and B domain of protein G using all-atom Go-models. Our data substantiate the usefulness of such simulation for probing the folding mechanism of proteins and demonstrate that multifunnel versions of such models also allow probing of more complicated funnel landscapes. In our case, such models reproduce the experimentally observed distributions of the GA98 and GB98 mutants which differ only by one residue but fold into different structures. They also reveal details on the folding mechanism in these two proteins.

INTRODUCTION

Anfinsen's seminal experiments[1] on denaturation and refolding of ribonuclease A showed that proteins can fold spontaneously into their active structure. This implies that the three-dimensional structure of a protein is encoded in its sequence of amino acids. Correspondingly, one finds often that homology between protein sequences implies similarity among their structures. This observation is the basis of many present structure prediction algorithms, but not without counter examples. Recently, Orban, Bryan, and co-worker performed a set of mutation experiments,[2,3] starting from the A and B domains of protein G (GA and GB), that led to proteins with sequence identities over 90% but distinct structures and functions. In the extreme case (GA98 and GB98), these proteins differed by only a single residue that acts as a switch between the two structures.[2] Such behavior is difficult to understand in a simple funnel picture[4,5] which assumes that the free energy landscape has a "funnel"-like shape when projected onto a suitable order parameter (often the number of native contacts), with multiple possible folding pathways leading to a unique native state.

The mutation experiments of Bryan and Orban suggest an extension of this picture to a double funnel model in which the mutations decide on the relative weight of the two funnels, where in the extreme case a single residue acts as a gatekeeper between the two basins of attraction. This is an intriguing concept because it would suggest that the sequence of amino acids in a protein may not only contain information on the native structures but also on other structures related to the evolutionary history or future of a protein (a protein may accumulate mutations that may not change its present structure but together with additional mutations could lead to new structures) or correspond to kinetically important intermediates.[6,7] In the case of GA98, the competing structure (resembling the B domain of protein G instead of the A domain) is also observed experimentally with a certain, but low, probability.[8] However, in general, the competing, dormant, structures appear with little probability and it is experimentally difficult to test whether a given protein can take potentially an alternate structure given different environmental conditions or further mutations.

These limitations do not exist in computer simulations, which allow, in principle, mapping the free energy landscape of a protein and exploring its basins of attraction or the connecting transition states. Unfortunately, computer simulations that probe the fundamental processes of folding, binding, and aggregation of proteins or their interaction within a cell are extremely difficult for realistic protein models: all-atom models lead to a rough energy landscape with a huge

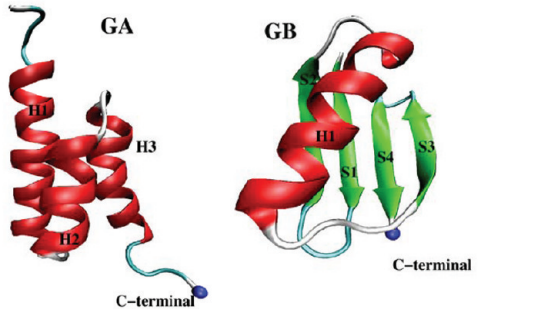
Special Issue: Harold A. Scheraga Festschrift

Received: November 1, 2011
Revised: December 31, 2011
Published: January 3, 2012

   dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645–6653

number of local minima separated by high barriers. The resulting increase in computational costs increases exponentially with the size of the protein. Although this problem can be alleviated with use of sophisticated simulation techniques such as parallel tempering,[9,10] multicanonical sampling,[11] or other generalized ensemble techniques[12] by Harold Scheraga[13,14] and others (for a recent review, see, for instance, ref 15), it limits the size of proteins that can be studied, even if coarse-grained models[16] are used. If the interest is not in predicting new structures, but understanding the folding mechanism of proteins with known structures, the problem can be lessened by using all-atom or simple C_α Go-models, as introduced Brooks,[17] Onuchic,[18−20] and others.[21−25] The underlying idea in all Go-models is to favor energetically the formation of contacts that are observed in the (known) native structure of the protein over that of other contacts; hence, these models are especially suitable for simulating proteins with a single folding funnel.

Our interest is in proteins with a more complex free energy landscape. For these cases, it was recently proposed[26] to extend Go-models such that the contacts as found in two distinct structures are energetically favored. In the present paper, we explore what we can learn from such Go-model simulations on folding of the A and B domains of protein G, whose native structures are shown in Figure 1. For this purpose, we study the folding mechanism, free energy landscape, and transition states of the wild types of these two proteins and compare our results with the GA98 and GB98 mutants that differ in only a single residue (see also Figure 1).



GA MEAVDANSLA QAKEAAIKEL KQYGIGDYYI KLINNAKTVE GVESLKNEIL KALPTE
GA98 TTYKLILNLK QAKEEAIKEL VDAGTAEKYF KLIANAKTVE GVWTYKDEIK TFTVTE
GB98 TTYKLILNLK QAKEEAIKEL VDAGTAEKYF KLIANAKTVE GVWTWKDEIK TFTVTE
GB MTYKLILNGK TLKGETTTEA VDAATAEKVF KQYANDNGVD GEQTYDDATK TFTVTE

Figure 1. Folded structure of the A domain and B domain of protein G as deposited in the Protein Data Bank under identifiers 2FS1 and 1PGB, respectively. The mutants GA_{98} and GB_{98} share the fold of the parent wild type. Sequence of wild types and mutants are listed in one-letter code.

Our interest is in both probing the differences between wild type and mutant and testing the reliability of our model. For the latter purpose, we simulate the A domain (B domain) with an all-atom Go-model that favors the contacts found in the native structure of the B domain (A domain). We find that our all-atom Go-model seems to distinguish between a sequence that "fits" a certain fold and one that does not. Simple Go-models that have only a single funnel are compared with the extended models that can also model more complicated landscapes. Although the simple all-atom Go-models allow probing of the folding mechanism of the wild type GA and GB proteins, their use is limited for the mutants GAS and GB. Only the modified versions correctly reproduce the experimental observation that GA98 samples both the GA and the GB fold, and GB98 samples only the GB fold, albeit both sequences differ only in a single residue. We propose a mechanism that explains the differences between the two mutants.

METHODS

Our simulations rely on the all-atom Go-model recently introduced by the Onuchic lab.[20] In this model, the energy of a protein configuration is given by

$$E = \sum_{\text{bonds}} \varepsilon_r (r - r_0)^2 + \sum_{\text{angles}} \varepsilon_\theta (\theta - \theta_0)^2 + \sum_{\text{impropers/planar}} \varepsilon_\chi (\chi - \chi_0)^2 + \sum_{\text{backbone}} \varepsilon_{BB} F_D(\varphi) + \sum_{\text{sidechains}} \varepsilon_{SC} F_D(\varphi) + \sum_{\text{non-contacts}} \varepsilon_{NC} \left(\frac{\sigma_{NC}}{r}\right)^{12} + \sum_{\text{contacts}} \varepsilon_C \left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - 2\left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \quad (1)$$

with $F_D(\phi) = [1 - \cos(\phi - \phi_0)] + 1/2[1 - \cos(3(\phi - \phi_0))]$. The harmonic term accounts for chain connectivity, the second term represents the bond angle potential, and the potential for the improper and planar degrees of freedom is described by the third term. Flexible dihedrals are given by fourth and fifth terms. A soft sphere repulsive potential (the sixth term in the equation) disfavors the formation of non-native contacts. Finally, the nonlocal interaction energy between atoms in a native contact is modeled by a 6−12 Lennard-Jones potential. A comprehensive listing of parameters can be found in ref 20.

Simulation Details. To model the wild type CFr protein and its mutant, we use the implementation of the above energy function by SMOG (**s**tructure-based **mo**dels in **G**romacs). This publicly available web server, located at http://smog.ucsd.edu,[27] accepts pdb files as input and returns topology and parameter files needed to set up the corresponding Go-model simulation in the Gromacs program package.[28] Given a reference configuration, SMOG generates topology and other input files that allow an immediate simulation of the protein if only a single funnel is assumed. In the cases that we assumed a more complex topology of the energy landscape, we edited the topology files by merging the information (i.e., contacts and dihedral maps) of two funnels. The few contacts that appeared in both funnels are treated separately by setting their energy to zero. In this way, we avoided additional roughness of the energy landscape from competing interactions resulting from these contacts.

We use Gromacs 4.5.3[28] with Langevin dynamics, starting with an extended protein placed in a cubic box with minimal distance of 1 nm from protein to the box walls. The time step is 0.0005, and we used 2×10^8 steps for production runs. For different systems, we use different temperatures in the 102−125 range, which is in all cases close to the folding temperature, that is, allowing the protein to fold and unfold. The corresponding values are given in the text. Note that time and temperatures are given in reduced units because our force

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645−6653

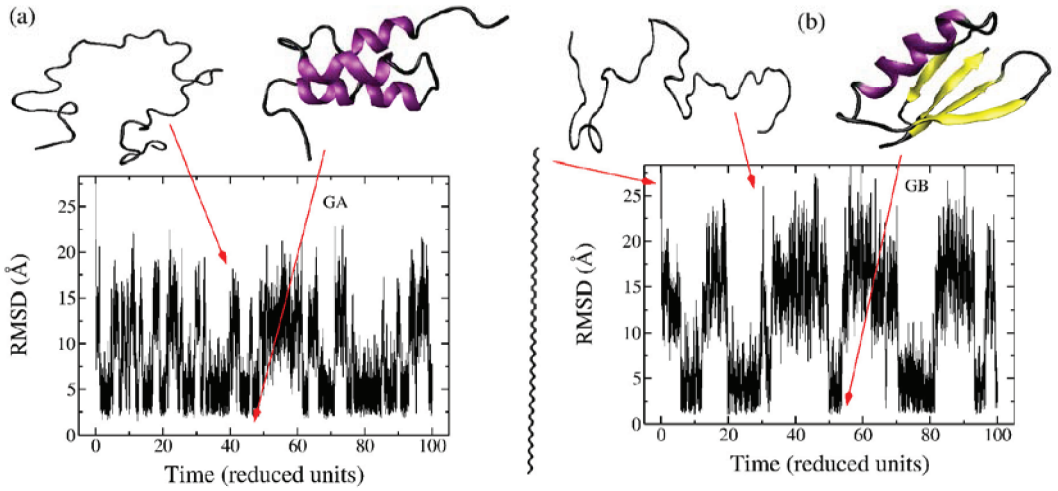


Figure 2. Root-mean-square deviation (rmsd) to the corresponding PDB structure as function of time for (a) the A domain (GA) and (b) the B domain (GB) of protein G. Data are from all-atom Go-model simulations at the respective folding temperatures $T_f^{GA} = 125$ and $T_f^{GB} = 110$.

field is not a physical energy function. Determination of critical temperature required several test runs for each protein. Temperatures where the population of folded and unfolded states are approximately equal where chosen for the main simulations runs of 2×10^8 MD steps, For analysis, we reweighed the so-obtained data to the folding temperatures, defined by us by the condition that the populations of folded and unfolded configurations are of equal depth.

RESULTS

We first present our results from Go-model simulations of the wild types of A and B domain. In Figure 2a), we show for the A domain the root-mean-square deviation (rmsd) to the PDB structure as a function of simulation time. The frequent changes between configurations with rmsd below 3 Å (i.e., folded configurations) and such with rmsd above 15 Å indicates that our simulation is with $T = 125$ (in reduced units) at a temperature that is close to the folding temperature. The same is true for our simulation of the B domain at $T = 110$, shown in Figure 2b). The corresponding plots of the free energy as a function of the number of native contacts at these two temperatures in Figure 3 display in both cases two basins of

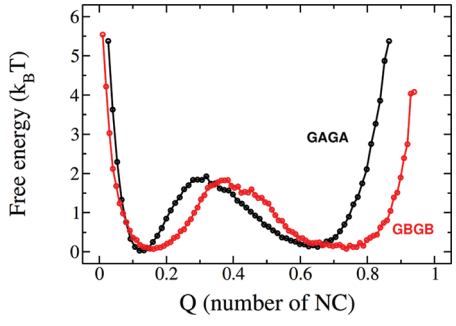


Figure 3. Free energy as a function of the number of native contacts. The data are from all-atom Go-model simulations at the respective folding temperatures $T_f^{GA} = 125$ and $T_f^{GB} = 110$.

attraction of almost equal depths, corresponding to the folded and denatured states. The two states are separated by a free energy barrier of $\approx 2k_BT$. Note again that the k_B here is not the usual Boltzmann constant but an arbitrary chosen constant that sets the energy (or temperature) scale.

An advantage of all-atom Go-model simulations over that relying on physical force fields is the ability to sample with high statistics the free energy landscape of large proteins, such as ours at the folding temperatures. This allows one to probe the folding mechanism of proteins. In Figure 4a, we show for GA the free energy landscape at the folding temperature as a function of the rmsd and the number of native contacts. The L-shaped landscape indicates that first, the rmsd decreases rapidly with an increasing number of contacts, but stays low and decreases little further once ≈100 contacts are formed. Clearly visible is a transition area separating the two regimes of folded and unfolded states. More details on the folding mechanism can be derived from Figure 4b, where the free energy is plotted as function of the number of native contacts and relative contact order (RCO) parameter,[29] defined by

$$\mathrm{RCO} = \frac{\sum_{ij} \Delta_{ij} |i - j|}{N \times L} \qquad (2)$$

where N is the total number of contacts and L is the total number of residues: $\Delta_{ij} = 1$ if residues form contact, and $\Delta_{ij} = 0$, otherwise. The two regions appear also in this plot, with the unfolded region characterized by a number of native contacts below 100 and a RCO of 0.06, and the folded region has an increased RCO of 0.08. This indicates that initially, local contacts are formed, and once a certain number is reached, long-range contacts are formed. This is a common scenario for helix bundles,[32,33] consistent with a framework model[30,31] that assumes as a first step formation of helices with their short-range contacts resulting from hydrogen bonding. In a second step, these helices arrange each other and form long-range contacts with such between the N- and C-terminal helices forming after such between the terminal and the central helix (data not shown).

These interhelical contacts lead to the observed increase in the RCO. This picture is supported by our analysis of transition states shown in Figure 6 and discussed later. This final folding step by arrangement of the three helices is connected with an increase in the side-chain contacts, as can be seen from Figure 4c. In this figure, the free energy is plotted as a function of the

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645–6653

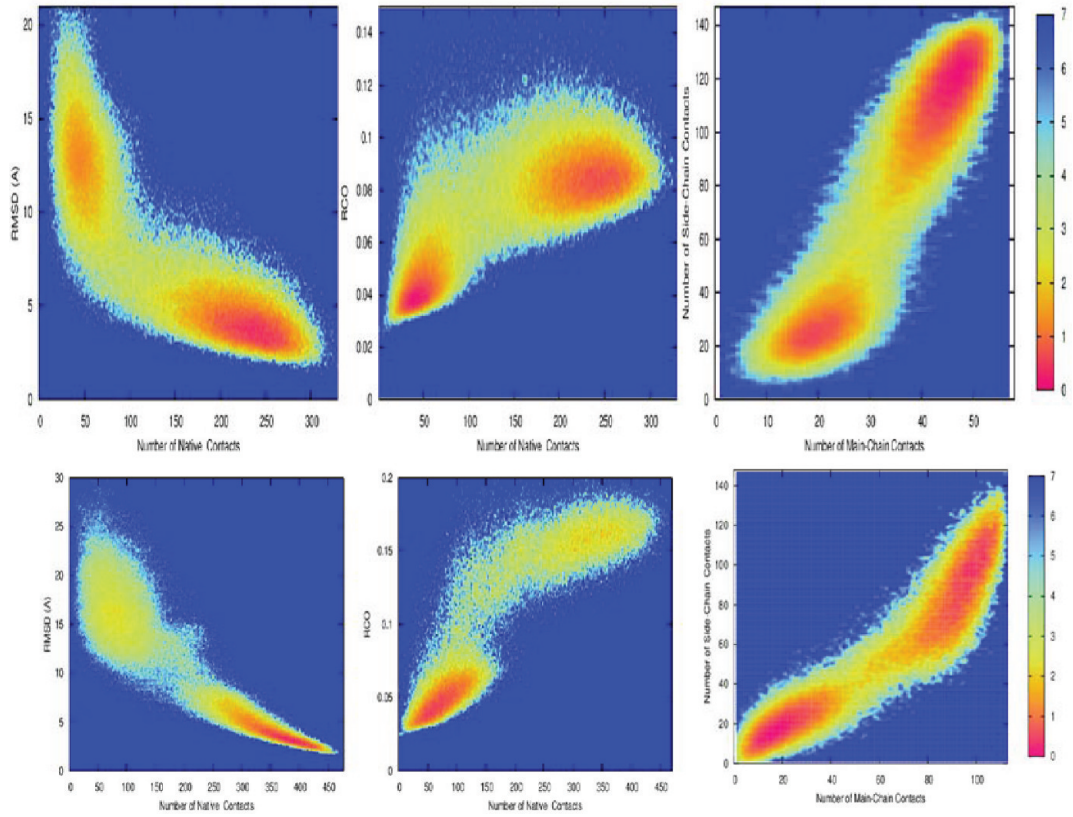


Figure 4. Two-dimensional free energy landscapes as a function of various quantities for the A domain (a–c) and B domain (d–f) of protein G. The data are from all-atom Go-model simulations at the respective folding temperatures $T_f^{GA} = 125$ and $T_f^{GB} = 110$.

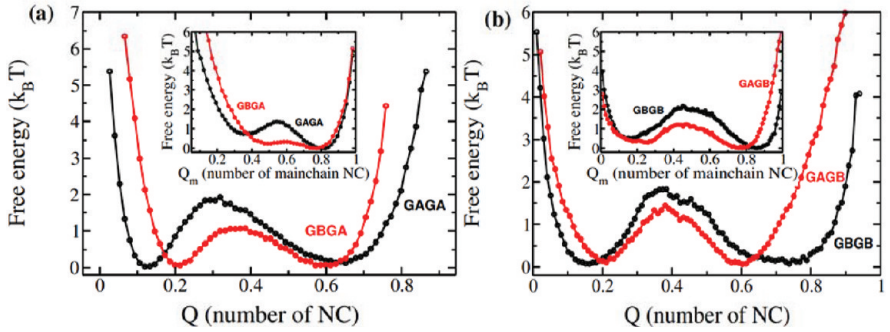


Figure 5. Free energy as a function of the number of native contacts, Q, for (a) GAGA and GBGA and (b) GBGB and GAGB. The inset shows the free energy as a function of the main-chain contacts. Data are from all-atom Go-model simulations at the respective folding temperatures of $T_f^{GAGA} = 125$, $T_f^{GBGA} = 125$, $T_f^{GAGB} = 113$, and $T_f^{GBGB} = 110$. Here, GAGB marks a Go-model constructed such that the GA sequence has the GB fold as the lowest energy state.

main-chain contacts (i.e., contacts only between backbone atoms) and side-chain contacts (contacts that involve side-chain atoms). Once the number of main-chain contacts grows above ≈30, the number of side-chain contacts increases rapidly, whereas below this threshold, the number of side-chain contacts grows only slowly.

The corresponding free-energy landscapes for the B domain are shown in Figure 4d–f. The free energy as a function of the number of native contacts and rmsd is less L-shaped than for the A domain. The transition happens at a larger rmsd value (≈10 Å for GB compared to ≈7 Å for GA), and the rmsd decreases strongly, even in the folded region with an increasing number of contacts. Note also that the two regions are more separated than in the GA case. This can also be seen in the free energy landscape projected on contact order and number of native contacts. There is a very sharp transition between a region of a small number of contacts and a low contact order and a region where the contact order is much larger and stays constant with an increasing number of contacts. The larger values of the contact order result from the β-sheets in GB that are inherently nonlocal.

Formation of the N-terminal β-hairpins (S1 with S2) seems to be the limiting step in the folding process, as can also be seen from transition states in Figure 6. Once formed, the hairpin seems to initiate formation of the S4 strand, followed by the S3 strand. The helix is formed and dissolves intermediately, but is stable only once it is in contact with the four β-sheets. As in the case of GA, side-chain ordering and formation of side-chain

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645–6653

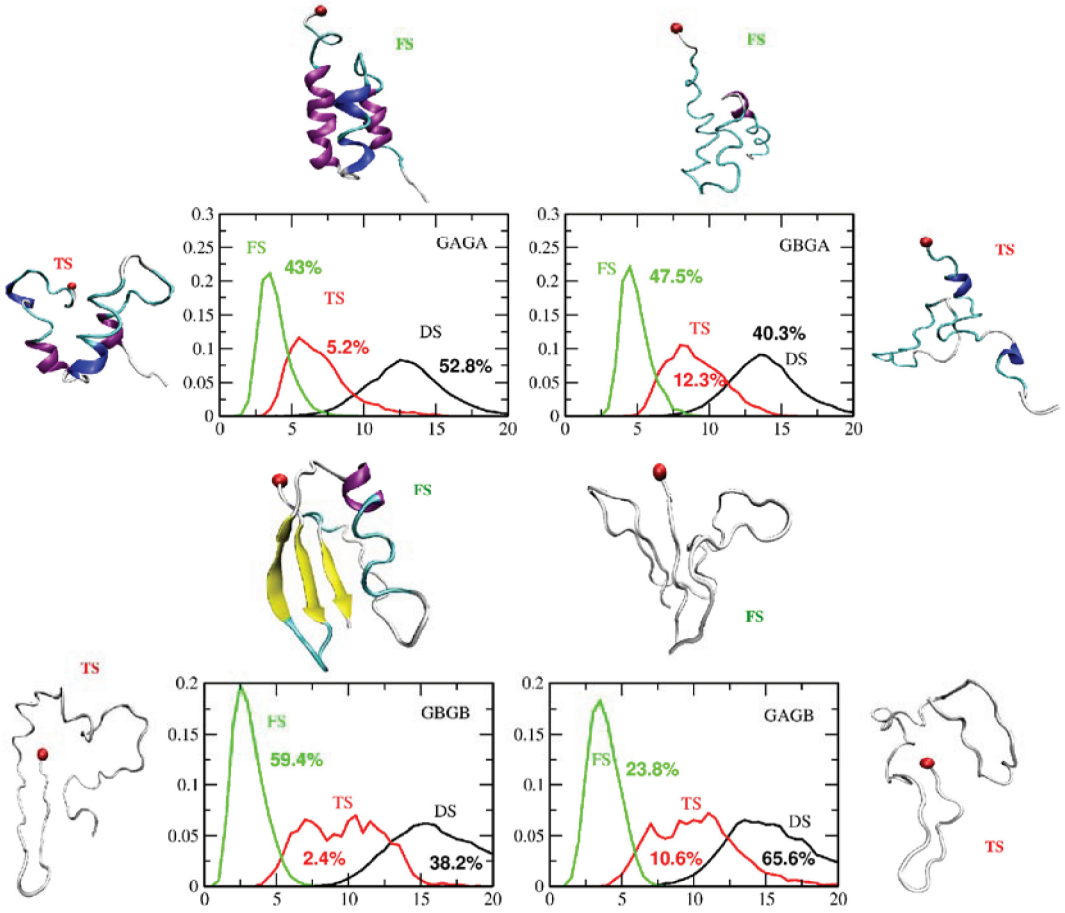


Figure 6. Percentage of native state (FS), transition state (TS), and denatured state (DS) as a function of the rmsd as obtained in simulations of the four all-atom Go-models GAGA, GBGA, GAGB, and GBGB at the respective folding temperatures: $T_f^{GAGA} = 125$, $T_f^{GBGA} = 125$, $T_f^{GAGB} = 113$, and $T_f^{GBGB} = 110$. The curves are normalized such that the area under the curve is 1. The absolute frequencies of state are listed above each curve. Also shown are typical folded and transition state configurations. The N terminal of each configuration is marked by a dot.

contacts seem to happen only after it has assumed its fold. The sequence of events is somehow different from the one observed in coarse-grain simulations of the protein by the Kolinski lab,[34] which found that folding starts with formation of the C-terminal hairpin (S3−S4), followed by contacts of this hairpin with the helix. A third step is the formation of the N-terminal hairpin S1−S2, followed by completion of the correct fold with formation of the S1−S4 contacts. It is not clear to us whether these differences result from the differing dynamics (Monte Carlo moves versus molecular dynamics) or from the choice of force fields.

An inherent problem with Go-model simulation is that the energy function is not physical. By energetically favoring contacts that appear in the native structure, one already makes assumptions on the folding mechanism; hence, there is a danger that the observed folding mechanisms are artifacts of the energy function. To probe how much the form of our energy function influences our results, we made the following test: We took the GA (GB) sequence and forced it into the GB (GA) fold. For this purpose, we started with the GB (GA) fold and changed the sequence of amino acids in 47 subsequent in silico "mutations" from GB (GA) to GA (GB). After each "mutation" done with Modeler,[35] the protein configuration was minimized carefully to avoid steric clashes while at the same time preserving the parent fold. We used the contacts in the so-generated two configurations to derive two all-atom Go-models, which we call GAGB and GBGA, that enforce for the GA sequence the GB fold as the lowest-energy state and vice versa for the GB sequence the GA fold. We then compared these artificial Go-models with the realistic ones, GAGA and GBGB, in which the GA (GB) sequence folds into the GA (GB) fold. In Figure 5 we contrast the free energy landscape as a function of native contacts at the respective folding temperatures for GAGA with GBGA, and for GBGB with GAGB. The inset displays the same quantities as the function of the number of native main chain contacts. Note that in both cases, the free energy barriers between the folded and unfolded states is substantially lower, and the transition is broader and less pronounced. The difference is especially prominent in the free energy as a function of main-chain native contacts. We observe such behavior also in all other free energy plots that we have considered (data not shown).

Although the folding mechanisms did not change, we find a less clear order in that long-range contacts are formed. This can also be seen in the distribution of transition states displayed in Figure 6. In this figure, we show for GAGA, GBGA, GAGB, and GBGB the distribution of folded states, transition states, and denatured states as functions of the rmsd. From an analysis of Figure 4 and the corresponding plots for GAGB and GBGA (not shown), we define here a state as a transition state if its free energy lies in the range close to the maximum free energy as a function of the number of native contacts. Structures from

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645−6653

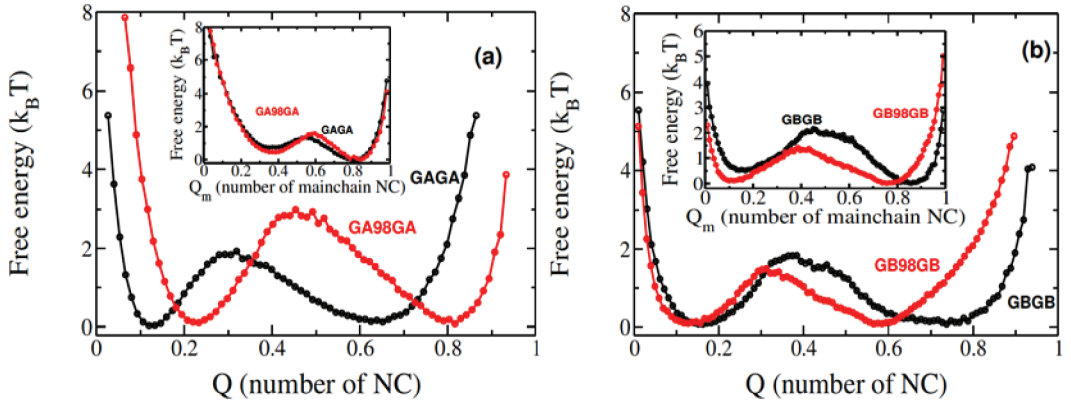


Figure 7. Free energy as a function of the number of native contacts, Q. In part a, we compare GAGA with GA98GA; in part b, GBGB with GB98GA. The inset shows the same quantity as a function of main-chain contacts only. All results are from all-atom Go-model simulations at their folding temperatures: T_f^{GAGA} = 125, T_f^{GA98GA} = 125, T_f^{GA98GB} = 110, T_f^{GBGB} = 110, T_f^{GB98GB} = 110, and T_f^{GB98GA} = 123.

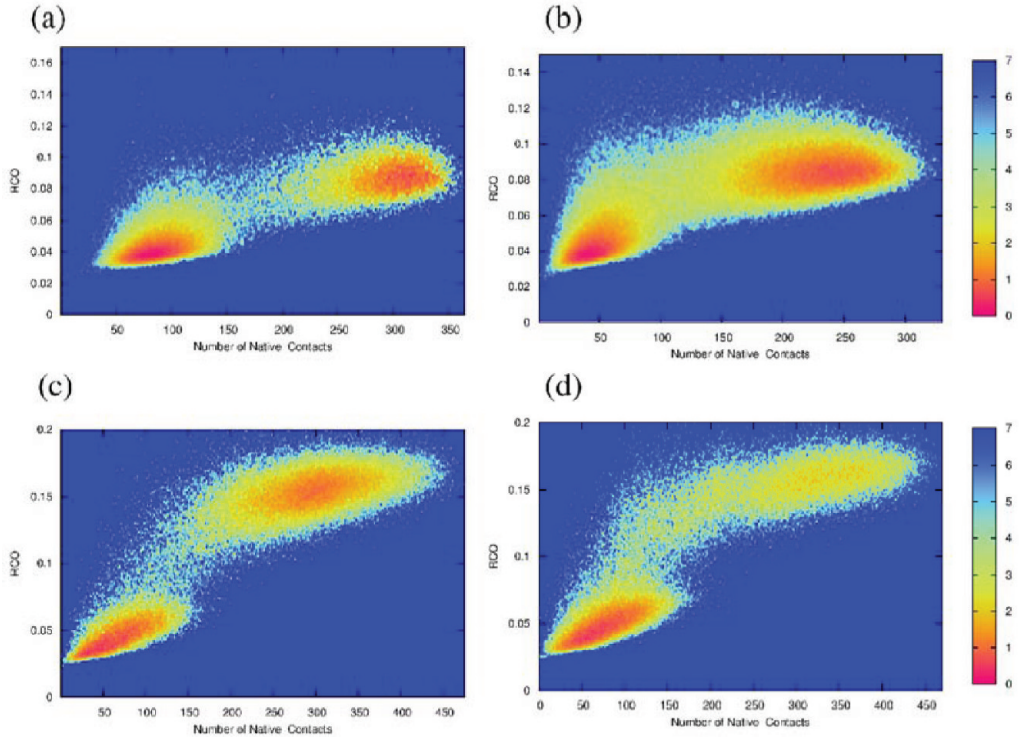


Figure 8. Free energy as a function of contact order RCO and number of native contacts, Q, for (a) GA98GA, (b) GAGA, (c) GB98GB, and (d) GBGB. All data are from all-atom Go-model simulations at the folding folding temperatures: T_f^{GA98GA} = 125, T_f^{GAGA} = 125, T_f^{GB98GB} = 110, and T_f^{GBGB} = 110.

regions with smaller or greater free energy are referred to as unfolded or folded, respectively. Note that the Y axis shows relative weights; that is, the area under each curve is set to 1. The total frequency of each group is on top of the respective curves. Also shown in the figures are typical folded (top or bottom) and transition state (left or right side) configurations. Note that not only is the absolute frequency of the transition states higher in GBGA (GAGB) than in the native GAGA (GBGB), but also the distribution is broader.

The transition states and folded states themselves appear to be less defined for the artificial models. We believe that this is because the sequence of a protein also optimizes the side chain contacts in its native state. Forcing the GA (GB) sequence into the GB (GA) fold leads during the late stages of the folding process to a large number of competing non-native interactions that lead to the broadening of the transition state distribution; hence, interestingly, our all-atom Go-model seems to distinguish between a sequence that "fits" a certain fold and one that does not.

We now extend our analysis to the two mutants GA98 and GB98, in which the GA and GB sequences are mutated toward two sequences that differ only in a single residue (45LEU in GA98 vs 45TYR in GB98), but keep both their original fold and function. In Figure 7a and b, we show the free energy landscape as a function of the number of native contacts, Q, for GA98 and GB98. The inset shows the same quantity as a function of the main chain contacts only. Data are again from a simulation at the folding temperatures (T_f^{GAGA} = 125 and T_f^{GB98} = 110).

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645–6653

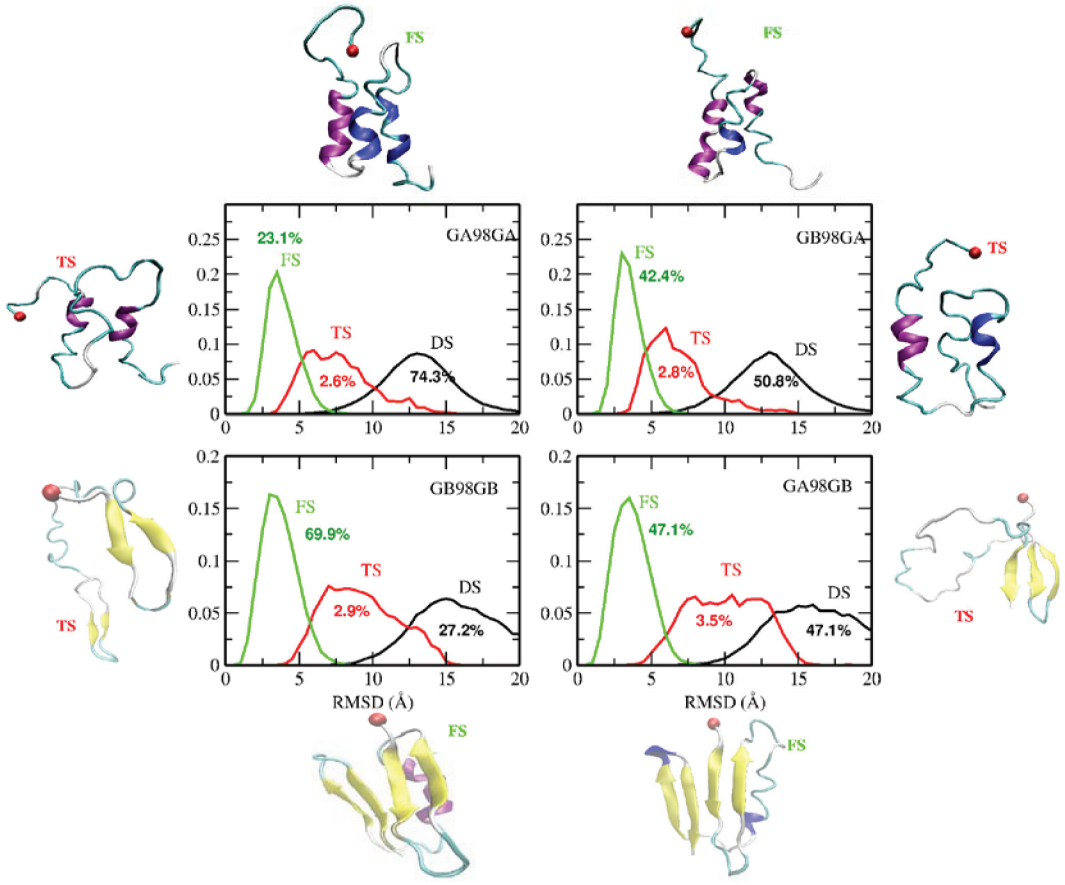


Figure 9. Percentage of native state (FS), transition state (TS), and denatured state (DS) as a function of the rmsd as obtained in simulations of the four all-atom Go-models GA98GA, GB98GA, GA98GB, and GB98GB at the respective folding temperatures: $T_f^{GA98GA} = 125$, $T_f^{GB98GA} = 123$, $T_f^{GA98GB} = 110$, and $T_f^{GB98GB} = 110$. The curves are normalized such that the area under the curve is 1. The absolute frequencies of state are listed above each curve. Also shown are typical folded and transition state configurations. The N terminal of each configuration is marked by a dot.

The position of the free energy barrier for GA98 is shifted to the right of that of GA, and its height is raised by about 1 k_BT. On the other hand, for GB98, the barrier is slightly lower and shifted to the left of that in GB. These shifts result from the differences in the number and kind of side-chain contacts. When looking into the free energy as a function of the main chain contacts, the shifts are smaller, and for Ga98, the barrier height differs little from that of GA. On the other hand, the barrier in free energy as function of main chain contacts is lowered for GB98 by about 1 k_BT over GB. Note that for the wild types (GA and GB), the height of the barrier differs little between the case in which the free energy is projected onto the total number of native contacts and the case in which it is projected onto the number of native main-chain contacts.

On the other hand, there are larger differences between the two cases for the mutants (GA98 and GB98). This indicates that arrangement of side chains plays a more important role for the folding of the mutants than for that of the wild types. We remark that unlike for the wild types, the corresponding figures for GA98GB and GB98GA (i.e., the GA98 (GB98) sequence forced into the GB (GA) fold) look very similar to that of GA98GB and GB98GB (data not shown). This is not unexpected because the two sequences GA98 and GB98 differ in only one residue.

When the energy landscape is plotted as a function of two variables, the transition appears to be sharper for the mutants than for the wild types. This can be seen in Figure 8, where we plot the free energy as a function of the relative contact order, RCO, and the number of native contacts for (a) GA98, (b) GA, (c) GB98, and (d) GB. The transition region is considerably less populated for GA98 and GB98 than in the corresponding wild types, with the difference more pronounced for GA and GA98. This can be also seen from the frequency in transition states shown in Figure 9 and indicates that for the mutants, folding is more of an all-or-nothing than for the wild type and requires narrow pathways in the funnel landscape.

On the other hand, little difference is found in the folding kinetics. As the wild type (GA), GA98 folds by first forming the three helices that then arrange themselves in the final fold, with contacts between the central helix and the terminal helices forming before that between the two terminal helices. In the same way, GB98 forms as GB first the N-terminal hairpin S1−S2, which then initiates formation of the S4 strand, followed by the S3 strand, and the helix, forming transiently but being stable only after forming contacts with the four β-sheets. The similarity in the transition states between Figure 6 and Figure 9 is consistent with this picture of the same folding mechanisms in wild types and mutants.

By construction, Go-models favor energetically the formation of contacts that are observed in the (known) native structure of the protein over that of other contacts. Hence, these models are especially suitable for simulating proteins with a single folding funnel. However, we can expect in the case of GA98 and GB98 that the free energy landscape is more diverse. Because the two

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645−6653

sequences differ only in a single residue, we conjecture that the landscape consists of two funnels, with that specific residue acting as a gatekeeper between the two basins of attraction. Such behavior cannot be modeled with a simple Go-model, and our above presented results for Ga98 and Gb98 assume implicitly that the effects of the secondary funnel are minor. However, in the case of GA98, the competing structure (resembling the B domain of protein G instead of the A domain) is also observed experimentally,[8] indicating that this assumption is not valid.

For this reason, we have also studied the two mutants GA98 and GB98 with Go-models that explicitly assume a folding landscape with two funnels (leading in either the GA or the GB fold). In Figure 10, we show the time series and energy

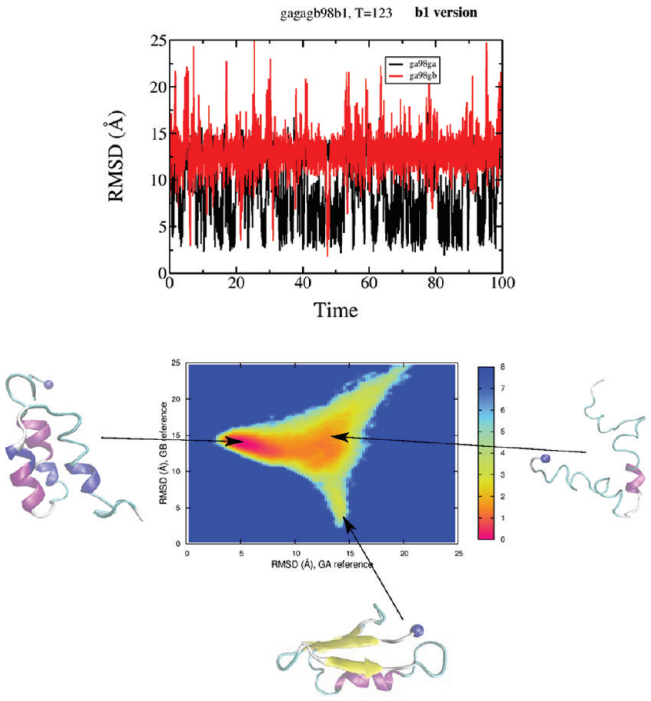


Figure 10. (a) The rmsd with respect to the GA98 (black) and GB98 (red) structure as a function of time. The corresponding free energy landscape is shown in part b, together with typical structures in the main local minima. The data are from an all-atom Go-model simulation at a folding temperature of $T_f^{GA98/GAGB}$ = 111. The Go-model was modified such that the model has two two global minima corresponding to the GA98 and GB98 structures.

landscape for a simulation of GA98 in such a two-funnel, all-atom Go-model. Both time series of rmsd (with respect to GA and GB fold) and free energy landscape demonstrate that GA98 can assume in this model both the GA and the GB fold, but the free energy of the protein in the GB fold is about 3 k_BT higher than that of the protein in the GA fold and comparable to that of the unfolded protein. Note that the free energy changes continuously between the various states, and no obvious barriers can be seen. On the other hand, the corresponding figure for GB98 in Figure 11 unveils a different behavior. Again, the protein can assume both folds, and again the "correct" fold (GB) is favored by about 3 k_BT over the competing GA fold whose free energy is also comparable to the unfolded states, but in this case, there is a clearly defined barrier isolating the GA states. This is consistent with the experimental

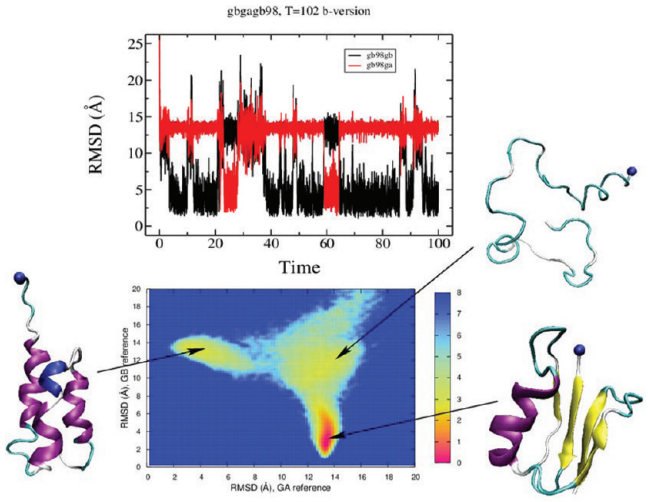


Figure 11. (a) The rmsd with respect to the GA98 (red) and GB98 (black) structure as a function of time. The corresponding free energy landscape is shown in part b, together with typical structures in the main local minima. The data are from an all-atom Go-model simulation at a folding temperature of $T_f^{GB98/GAGB}$ = 102. The Go-model was modified such that the model has two two global minima corresponding to the GA98 and GB98 structures.

results in which GA98 is also observed with a small frequency in the GB fold but GB98 is not found in a GA fold.

For both GA98 and GB98, a switch from one to the other configuration implies going through the ensemble of unfolded configurations. In the case of GB98, there is a large barrier between the ensemble of unfolded states and the ensemble of configurations with the GA fold and a much smaller barrier between unfolded states and such with a GB fold. On the other hand, for GA98, no such large barrier exists that separates the ensemble of unfolded states from that of configurations in the competing GB fold. Because both proteins differ by a single residue (45LEU in GA98 vs 45TYR in GB98), we can expect that the behavior of this residues leads to the different folding landscapes, that is, creating the barrier for GB98 but not for GA98.

Our first assumption was that residue 45 selects the conformation of the C-terminal segment, which afterward acts as a template for the structure of the N-terminal segment. Such a mechanism was observed in unfolding simulations of the similar mutants GA59 and GB59 (which have only 59% of sequence identity instead of the 98% identity between GA98 and GB98) in ref 36. However, in our simulations, the observed folding events indicate that folding into the GA structure starts with formation of the three helices that then form interhelical contacts, whereas such into the GB-fold starts with formation of the N-terminal hairpin S1−S2, followed by contacts S1−S4 and later S3−S4 and helix-strand contacts. Although the folding of GA98 is similar to the one observed in unfolding simulations of GA59 by the Daggett group,[36] this is not the case for GB98. The Daggett group concluded that GB59 starts folding by forming first the S3−S4, which becomes stabilized by contacts with the helix and prevents folding toward the helix bundle of the GA fold. Instead, our results indicate that the differences between GA98 and GB98 result from the disparity in that 45LEU and 45TYR form long-range contacts, not from the way they form their local secondary structure contacts.

dx.doi.org/10.1021/jp210497h | J. Phys. Chem. B 2012, 116, 6645−6653

This is consistent with an analysis of contacts formed by these residues in the transition regions of Figures 10 and 11, which indicates that typical helical short-range contacts such as between residues 49 and 49 are formed with similar frequency: 65% in GA98 versus 68% in GB98. The same is true for sheet contacts such as between residues 45 and 52, which are found for GA98 with 33% probability and with 32% for GB98. Similarly, we also find little difference between the transition regions in the frequency of the long-range contacts that stabilize the GB fold. For instance, we find contacts between residues 23 and 45 (helix strand 3) in GA98 with 5% and in GB98 with 4% frequency. However, we find a pronounced difference in contacts between residues 32−35 with residue 45. In the GA fold, the central helix interacts with the C-terminal helix through these contacts. For instance, the hydrophobic contacts between 33ILE and 45LEU are found in GA98 at 48%, but such between 33ILE and 45TYR in GB98 only at 29%. Formation of such contacts between the central and the C-terminal helix follows in the GA fold formation of the secondary structure elements. We conjecture that the reduced probability by 45TYR to form such interhelical contacts is what leads to the barrier in GB98 that protects against the GA fold in this protein. On the other hand, switching from 45TYR to 45LEU increases the frequency of such intrahelical contacts (i.e., favoring the GA fold), but does not change the frequency of the long-range contacts typical for the GB fold. In addition, these contacts (as, for instance, the helix-strand contacts between residues 23 and 45) form late in the process of folding to the GB structure. Formation of these contacts therefore does not constitute a bottleneck in the folding. Hence, the barrier that separates the GB-fold in GA98 is much lower than the one separating the GA-fold in GB98.

CONCLUSIONS

We have simulated wild type and mutants of the A and B domain of protein G using all-atom Go-models. Although Go-models by construction lead to preselected structures as lowest-energy states, our simulations show a clear difference between sequences that "fits" a certain fold and ones that do not. For the wild type GA and GB, simple all-atom Go-model simulations allow study of the folding mechanism of these proteins; however, such models that by construction have only a single folding funnel will fail when the energy landscape of a protein is more complex.

In the case of the mutants GA98 and GB98 that differ only in a single residue but have very different distributions of folded structures, we tried a modified Go-model that incorporates folding funnels to both GA and GB folds. This model not only correctly reproduced the experimentally observed distributions but also revealed details on the folding mechanism in these two mutants. Such multifunnel Go-models may therefore be suitable to study the evolutionary history of proteins or structural transitions in proteins or to incorporate ensemble information into a simulation. For instance, in an upcoming study (Ping, J.; Hansmann, U. H. E.; unpublished results), we study a protein taking into account structural information from a whole NMR ensemble instead of only a single structure.

AUTHOR INFORMATION

Corresponding Author

*E-mail: mkouza@mtu.edu; uhansmann@ou.edu.

ACKNOWLEDGMENTS

This work was supported, in part, by research Grant CHE-08090002 of the National Science Foundation and GM62838 of the National Institutes of Health (USA).

REFERENCES

(1) Anfinsen, C. B. *Science* **1973**, *181*, 223−230.
(2) Alexander, P.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 21149−21154.
(3) Alexander, P.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11963−11968.
(4) Leopold, P.; Montal, M.; Onuchic, J. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721−8725.
(5) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70−75.
(6) Liwo, A.; Khalili, M.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2362−2367.
(7) Mohanty, S.; Meinke, J. H.; Zimmermann, O.; Hansmann, U. H. E. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 8004.
(8) Bryan, P.; Orban, J. *Curr. Opin. Struct. Biol.* **2010**, *20*, 482−488.
(9) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140−150.
(10) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.
(11) Hansmann, U. H. E.; Okamoto, Y. *J. Comput. Chem.* **1993**, *14*, 1333−1338.
(12) Zimmermann, O.; Hansmann, U. H. E. *Biochim. Biophys. Acta* **2008**, *1784*, 252−258.
(13) Hao, M. H.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *98*, 4940−4948.
(14) Nanias, M.; Czaplewski, C.; Scheraga, H. A. *J. Chem. Theor. Comput.* **2006**, *2*, 513−528.
(15) Scheraga, H. A.; Khalili, M.; Liwo, A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 57−83.
(16) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovski, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849−873. Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovski, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874−887.
(17) Karanicolas, J.; Brooks, C. L. III. *J. Mol. Biol.* **2003**, *334* (2), 309−325.
(18) Clementi, C.; Garsia, A.; Onuchic, J. *J. Mol. Biol.* **2003**, *326*, 933−954.
(19) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937−953.
(20) Whitford, P.; Noel, J.; Gosavi, S.; Schug, A.; Sanbonmatsu, K.; Onuchic, J. *Proteins* **2009**, *75*, 430−441.
(21) Meinke, J. H.; Hansmann, U. *J. Phys.: Condens. Matter* **2007**, *19*, 285215.
(22) Luo, Z.; Ding, J.; Zhou, Y. *Biophys. J.* **2007**, *93*, 2152−2161.
(23) Kleiner, A.; Shakhnovich, E. *Biophys. J.* **2007**, *92*, 2054−2061.
(24) Kouza, M.; Chang, C. F.; Hayryan, S.; Yu, T. H.; Li, M. S.; Huang, T. H.; Hu, C. K. *Biophys. J.* **2005**, *89*, 3353−3361.
(25) Li, M. S.; Kouza, M. *J. Chem. Phys.* **2009**, *130*, 145102.
(26) Lammert, H.; Schug, A.; Onuchic, J. *Proteins* **2009**, *77*, 881−891.
(27) Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. *Nucleic Acids Res.* **2010**, DOI: 10.1093/nar/gkq49.
(28) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435.
(29) Plaxco, K. W.; Simon, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.
(30) Ptitsyn, O. B. *Dokl. Akad. Nauk SSSR* **1973**, *210*, 1213−1215.
(31) Kim, P. S.; Baldwin, R. L. *Ann. Rev. Biochem.* **1982**, *51*, 459−489.
(32) Boczko, E. M.; Brooks, C. L. III. *Science* **1995**, *269*, 393.
(33) Trebst, S.; Hansmann, U. H. E. *Eur. Phys. J. E* **2007**, *24*, 311−316.
(34) Kmiecik, S.; Kolinski, A. *Biophys. J.* **2008**, *94*, 726−736.
(35) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779−815.
(36) Scott, K. A.; Daggett, V. *Biochemistry* **2007**, *46*, 1545−1556.