WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes

Allison P. Heath,[1] Lydia E. Kavraki,[1,2,3]* and Cecilia Clementi[3,4]*

[1] Department of Computer Science, Rice University, Houston, Texas 77005

[2] Department of Bioengineering, Rice University, Houston, Texas 77005

[3] Department of Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas 77030

[4] Department of Chemistry, Rice University, Houston, Texas 77005

ABSTRACT

Multiscale methods are becoming increasingly promising as a way to characterize the dynamics of large protein systems on biologically relevant timescales. The underlying assumption in multiscale simulations is that it is possible to move reliably between different resolutions. We present a method that efficiently generates realistic all-atom protein structures starting from the C_α atom positions, as obtained for instance from extensive coarse-grain simulations. The method, a reconstruction algorithm for coarse-grain structures (RACOGS), is validated by reconstructing ensembles of coarse-grain structures obtained during folding simulations of the proteins src-SH3 and S6. The results show that RACOGS consistently produces low energy, all-atom structures. A comparison of the free energy landscapes calculated using the coarse-grain structures versus the all-atom structures shows good correspondence and little distortion in the protein folding landscape.

Proteins 2007; 68:646–661.

Key words: *multiscale; coarse grain; minimalist model; all-atom reconstruction; protein folding; free energy landscape.*

INTRODUCTION

Multiscale techniques have recently emerged as promising tools to combine the efficiency of coarse-grain simulations with the detail of all-atom simulations for the characterization of a broad range of molecular systems in fields such as material science and biophysics. Recent work has focused on the definition of strategies that combine different resolutions in different regions of the space during a single simulation.[1,2] For instance, this idea has been applied to a system of small molecules where some parts of the space use the all-atom representation and the rest of the space uses a coarse-grain representation.[1] Multiple resolution simulations have also been used to study membrane-bound ion channels by coarse graining the lipid and water molecules while using an all-atom representation for the polypeptide ion channel.[3] In the context of protein simulation, a similar idea has been applied to represent parts of the protein, such as the active site, in all-atom detail while using a coarse-grain model for the rest of the system.[2] Additional multiscaling strategies for protein systems focus on changing the whole system resolution during the same simulation. One of the first applications in this area used a simplified protein model as a starting point to evaluate the folding free energy of the corresponding all-atom model.[4] In a more recent example, the villin headpiece was studied using structures from coarse-grain simulations as initial configurations for all-atom simulations. This allowed for a larger sampling of the protein's conformations than using all-atom simulations alone.[5] Coarse-grain simulations have also been used to probe the putative folding transition state structures obtained from all-atom simulations.[6] Another idea, known as "resolution exchange" or "model hopping", allows movement between different levels of structural detail in order to cross energy barriers.[7–9]

The underlying assumption in the definition of multiscale techniques for protein simulation is that it is possible to reliably and efficiently move between coarse-grain and all-atom models. The coarse-grain model used must be physically realistic so that the protein structures being sampled implicitly represent relevant conformations of the protein. Rigorous mathematical procedures, such as renormalization group theory, have yet to be applied to the general definition of coarse-grain models. Therefore, the evaluation of coarse-grain protein models is usually obtained by comparison to experimental data. Even assuming a realistic coarse-grain model, a robust and efficient pro-

Grant sponsor: NSF; Grant numbers: CHE-0349303, LK GM078988, LK and CC CCF-0523908, CNS-0454333, CNS-0421109; Grant sponsor: Robert A. Welch Foundation; Grant number: C-1570; Grant sponsor: Sloan Foundation (LK)
*Correspondence to: Lydia E. Kavraki, Department of Computer Science, Rice University, 6100 Main St., MS-132, Houston, TX 77005, E-mail: kavraki@rice.edu or Cecilia Clementi, Department of Chemistry, Rice University, 6100 Main St., MS-60, Houston, TX 77005. E-mail: cecilia@rice.edu
Received 26 August 2006; Revised 14 November 2006; Accepted 7 December 2006
Published online 21 May 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21371

cedure is needed to add all-atom details to coarse-grain protein structures to produce realistic (i.e., physicochemically sound) all-atom structures. It has been shown that moving between coarse-grain and all-atom models for systems such as simple polymers and liquids is possible.[8,10–12] However, the feasibility of consistently adding all-atom detail to coarse-grain protein models has not been thoroughly tested. There have been no thorough studies on whether moving from coarse-grain to all-atom protein structures distorts the thermodynamic properties of the corresponding ensembles of structures. Several reconstruction methods have been proposed for protein systems in the context of structure prediction, homology modeling, and protein design. These procedures mainly deal with the reconstruction of native state structures. However, the reconstruction of coarse-grain structures from extensive simulations with minimalist models, spanning large regions of the protein folding landscape, requires addressing additional concerns.

In this paper we describe a method to add all-atom detail to "general" (folded, unfolded, or partially folded) coarse-grain structures defined only by the positions of C_α atoms. We first validate our method using structures selected from the Protein Data Bank (PDB).[13] Most existing reconstruction algorithms are tested in this manner and we show that our method performs comparably with previous methods in this context. However, the reconstruction of protein structures spanning large regions of the protein folding landscape cannot be truly tested only considering PDB structures. We evaluate our procedure by applying it to coarse-grain structures generated by extensive folding/unfolding simulations performed using a minimalist model.[14,15] The results show that our method consistently produces realistic all-atom structures. To show the consistency between the coarse-grain and all-atom models, the reconstructed ensembles of both sets of structures are used to compute free energy landscapes. The results show that there is good correspondence between the landscapes produced by the coarse-grain and all-atom structures. In contrast, comparison to state-of-the-art side-chain positioning (SCP) programs illustrates that methods designed and optimized to reconstruct PDB structures cannot be reliably used to recover all-atom details in large regions of the protein folding landscape.

The reintroduction of all-atom resolution into protein configurations generated by coarse-grain simulations allows one to "zoom in" on the details in particular regions of a protein folding landscape. A closer look at the misfolded structures visited during the folding of a mutant of S6 ($S6^{Alz}$) show that the mechanism of stabilization of nonnative structures for this protein is fully in agreement with experimental evidence. Similarly, the analysis of the all-atom reconstructed transition state ensemble (TSE) of src-SH3 depicts a folding mechanism consistent with previous experimental and computational studies. To our knowledge this is the first large scale reconstruction experiment on a protein system, and provides solid groundwork for future development on multiscale modeling of protein systems.

EXISTING RECONSTRUCTION METHODS

As mentioned above, algorithms for the reconstruction of all-atom protein structures from coarse-grain structures have mainly been developed for applications in structure prediction, homology modeling, and protein design. Our goal of reconstructing large coarse-grain simulations spanning large regions of the protein folding landscape presents a different set of problems. We discuss these problems and how they were addressed in the Section Reconstruction Algorithm for Coarse-Grain Structures (RACOGS). In this section we briefly review the main ideas and previous work on reconstruction methods. When reconstructing protein structures using only the positions of the C_α atoms the problem is usually broken down into two parts. The backbone atoms are added to the structure first and then side-chain atoms are added to the reconstructed backbone.

Backbone reconstruction algorithms

The problem of determining the backbone atom positions of a protein with only the knowledge of the C_α atoms appears in the literature multiple times for different applications. Existing approaches to place backbone atoms use a variety of techniques such as analytical methods,[16–18] using known structures or peptide fragments explicitly[19–23] and more general statistical methods[24–26] based on a large number of known structures. Many of these methods have been validated on structures from the PDB, where they have been shown to position the backbone atoms efficiently and with a high degree of accuracy.

SCP algorithms

After the backbone atoms have been added to the coarse-grain structures the next step is to place in the side-chain atoms. The SCP problem has been heavily studied because of its applications in predicting and designing protein structures. Methods addressing the SCP problem usually discretize possible side-chain conformations into rotamers. Each rotamer represents one conformation of a side-chain. A set of these rotamers for all of the amino acids is called a "rotamer library." Much recent research has focused on producing rotamer libraries that realistically represent the conformations of side-chains.[27] In this context, the SCP problem is normally defined as given the positions of the backbone atoms, a set of possible rotamers for each residue, and an energy

function, find a rotamer for each residue such that the final structure containing the positions of all of the side-chain atoms has the lowest global energy. It has been shown that the SCP problem is NP-complete[28] and that the solution cannot be approximated within any error bound.[29] Previous work has also shown that there are limits on SCP accuracy on native and near-native backbones.[30,31] However, practical results have suggested that good conformations can be produced readily. Many methods have been proposed to solve this problem based on techniques such as dead-end elimination and its variants,[32–38] Monte Carlo methods,[22,39] simulated annealing,[40] local optimization,[30,41,42] genetic algorithms,[43,44] mean field optimization,[45–47] graph theoretical algorithms,[48–51] integer linear programming,[52,53] consensus modeling,[31] and other approaches.[54] These methods have been mainly tested upon their performance by how well they reconstruct PDB structures given only backbone coordinates.

RECONSTRUCTION ALGORITHM FOR COARSE-GRAIN STRUCTURES

RACOGS was designed for the purpose of multiscale modeling of protein landscapes. In this vein, several considerations were taken into account during the development of the method. Most importantly, the all-atom structures produced by the algorithm must be physically realistic. Coarse-grain structures obtained through folding simulations using minimalist models differ from PDB structures because they can be far away from the native state of the protein. PDB structures are usually native or near-native structures, while coarse-grain simulations can contain more unstructured conformations found in the unfolded and transition states of the protein. A good reconstruction method should be able to handle any legitimate conformation of the protein.

An additional problem is presented by the fact that there are no original all-atom structures to compare to when reconstructing coarse-grain structures spanning large regions of the protein folding landscape. Therefore, a metric needs to be chosen to assess how realistic a reconstructed structure is. In this work we evaluate the "goodness" of protein structures by their relative potential energy, according to the Boltzmann criterion. We use a standard force field, AMBER99[55] with a generalized born/solvent accessible (GB/SA) implicit water model,[56] to evaluate the energy of the all-atom structures.

For the reconstruction method to be useful there must be a high probability that the coarse-grain structure will produce a reasonable all-atom structure. During multiscale modeling any valid coarse-grain structure may be considered a candidate for reconstruction. At the same time, the method must also be efficient enough to reconstruct hundreds of thousands of coarse-grain structures in a reasonable amount of time. Therefore, the reconstruction method should be able to efficiently produce relatively low-energy (i.e., statistically significant when Boltzmann-weighted) all-atom structures from most of the coarse-grain structures, even if very far from the native state. This is a key difference from previous reconstruction methods: While existing methods focus on the recovery of a native-like geometry, RACOGS was designed specifically to obtain physically realistic all-atom structures in any region of the folding landscape visited by coarse-grain protein simulations. We designed RACOGS to use only the C_α atom positions to produce a structure containing all heavy atom positions of the protein.

RACOGS combines previous methods by Feig et al.[25] to handle the backbone reconstruction and a modified version of the method described by Xiang and Honig[31] to perform SCP. A novel side-chain minimization step has been added after the SCP step. We show that this step represents a crucial component of the method as it improves its performance greatly, and efficiently produces realistic all-atom structures even in regions far from the native state. The final step adds hydrogens to the structure and performs a short all-atom minimization. The steps of RACOGS are detailed in the following sections and illustrated in Figure 1.

Backbone reconstruction

The first step in RACOGS is to position the backbone atoms given only the C_α atom positions. The C, O, N, and C_α atoms of each amino acid are considered backbone atoms. The first step in Figure 1 corresponds to the backbone reconstruction step. The backbone reconstruction step of RACOGS is performed using the method previously proposed by Feig et al.,[25] which is in turn based on the work of Milik et al.[24] This is a statistical method that compiles the average positions of the backbone atoms of an amino acid based on the distances to the neighboring C_α atoms. Then these average positions are used to place the atoms in a coarse-grain structure. To compile the statistics 4013 nonredundant protein structures were selected from the PDB.

Side-chain positioning

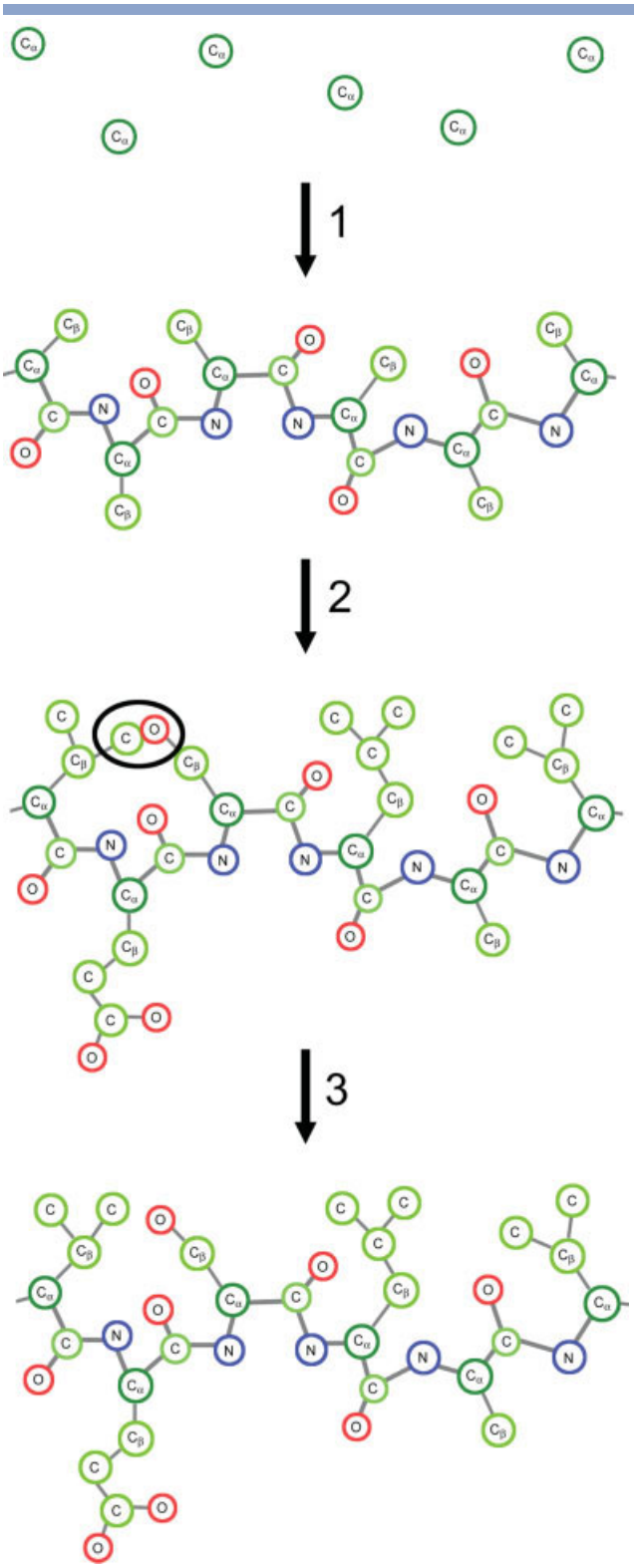
After the backbone atoms have been added, the next step is to position the side-chain for each residue, as shown in Figure 1. We have modified the method described by Xiang and Honig[31] for use in RACOGS. This method was chosen because it has been shown to perform well when reconstructing PDB structures, and is fairly efficient. The rotamer library used in this work is the most extensive backbone dependent, coordinate rotamer libraries described in Ref. 31. However, any rotamer library can

DOI 10.1002/prot

be used in the method. The energy function used in the SCP step consists of the van der Waals and dihedral energy terms as defined by the AMBER99 force field.[55]

The method to place the side-chains is a straightforward hill climbing algorithm. It starts by generating an initial structure that contains positions for all of the side-chains based on the backbone atom positions. The initial structure is constructed by placing the rotamer on each residue that has the minimal energy between the side-chain and the backbone atoms of the other residues. During this procedure any rotamer that has interaction energy with the backbone higher than a user defined cutoff is discarded and is no longer considered in further iterations. The energy cutoff helps to improve the efficiency by eliminating any side-chains that have steric clashes with atoms in the backbone. In this work an energy cutoff of 100 kcal/mol was used. Unlike the Xiang and Honig method, we only use this one structure as the initial conformation rather than generating 120 starting conformations. This was done to improve efficiency, even if it could cause a slight decrease in accuracy. However, the results show that the method still performs well. Additionally, there is evidence that most side-chains can be placed correctly by only using their interactions with the backbone[57] and other methods also use this as the initial structure.[54]

Starting from the initial structure an iterative procedure is used to find side-chains with the lowest energy. Each side-chain is selected in turn, and the interaction energy between the possible rotamers of currently selected side-chain and all of the other currently placed side-chains and the backbone is considered. If there is another rotamer that has a lower energy, the current rotamer is replaced by the lower energy rotamer. This continues until after a full iteration over the entire protein none of the rotamers are replaced or until a user specified maximum number of iterations is reached. In the results of this paper the maximum number of iterations allowed was 10, which was never reached when reconstructing the coarse-grain simulations of src-SH3 and S6. The rotamers can be considered sequentially down the chain or in a random order. We found that for the proteins studied in this paper the order of iteration did not affect the results.



Side-chain minimization step

After the SCP a number of side-chains in very high-energy conformations were detected. The all-atom minimization could only fix a small fraction of the high-energy

Figure 1

Cartoon illustration of the RACOGS method for a short peptide with the sequence VAL-ASP-SER-LEU-VAL. (1) Starting from the C_α atoms the backbone atoms are added. (2) After the backbones are added the side-chains are placed. (3) The first and third amino acids, circled, are clashing and causing a high energy interaction. The side-chain minimization step is performed on the first amino acid and resolves the clash. The last step of adding hydrogens and performing an all-atom minimization is not shown. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

interactions, producing many structures with unreasonably high energy. The all-atom minimization is also the most expensive part of the procedure and is performed for a fixed number of steps. Therefore, the energetic frustration introduced in the SCP needs to be removed to improve the overall performance of the method. The side-chain minimization step was developed to address this issue. After side-chain placement, if the energy between a side-chain and the rest of the protein is greater than a user specified cutoff then minimization is performed on the high-energy side-chain. This is performed by fixing the rest of the protein in place and only allowing the high energy side-chain to move during minimization. The side-chain minimization step is illustrated in Step 3 of Figure 1. The side-chain minimization was performed with the van der Waals, dihedral, bond and angle terms from the AMBER99 force field.[55] These terms were chosen to eliminate steric clashes without causing undue bond or angle stretching. The minimization was performed using the conjugate gradient method from the standard optimization package OPT++[58] with a maximum number of iterations set to 100.

Our independently developed side-chain minimization step is similar to a method recently proposed for side-chain modeling in protein–protein docking.[59] The purpose of this step is to overcome the limitations of using a rotamer library by allowing the side-chain to move through a continuous space. As rotamer libraries are built upon statistics on relative positions of side-chain atoms in native protein structures, they may introduce a strong bias in the positioning of side-chain in nonnative configurations, where the local packing is not as tight. Overall, the side-chain minimization step produces all-atom structures with lower initial energy and less steric clashes, which greatly improves the performance of the subsequent all-atom minimization over the whole protein. The results presented in next section show that including the side-chain minimization step significantly increases the number of low-energy structures obtained, particularly when considering configurations that are not necessarily close to the protein native state.

All-Atom minimization

The final step of the reconstruction process is to perform a short all-atom minimization over the whole protein. Once all of the side-chain heavy atoms are added to each structure, hydrogens are added using the leap program from the AMBER 8 suite.[60] Then the all-atom structures are minimized for up to 150 steps of conjugate gradient minimization using the program sander from the AMBER 8 molcular dynamics package. We observed that for src-SH3 and S6 the energy normally converges after roughly 100 steps of minimization. The energy function used is AMBER99 with the GB/SA implicit water model. The parameters used for GB/SA are from Onufriev et al.,[56] which were developed to improve accuracy in simulations with large conformational changes.

RESULTS

RACOGS was first tested on the reconstruction of PDB structures. The accuracy of RACOGS is presented in comparison to a SCP method known to perform well on PDB structures, SCWRL 3.0.[50] The second, and most important, part of the results focuses on the reconstruction of coarse-grain structures obtained from simulations. The coarse-grain model used has been extensively discussed and validated elsewhere and has been shown to produce results in good agreement with experimental data.[14,15] We demonstrate that RACOGS is able to reconstruct a high percentage of low-energy, all-atom structures from the coarse-grain structures. On the contrary, using a method developed for protein structure prediction applications produces a much lower percentage of all-atom structures when applied to coarse-grain simulations.

The comparison of the resulting free energy landscapes from the reconstructed all-atom structures and the coarse-grain structures shows that they are consistent. In addition, a closer look at the misfolded structures of a mutant of S6, and at the transition state structures of src-SH3 shows that the all-atom structures are in agreement with experimentally determined properties. All of the reconstruction experiments were run on Ada, a Cray XD1 system containing 316 dual core AMD Opteron 275 2.2 GHz processors located at Rice University.

Preliminary applications and performance evaluation

Reconstruction of PDB structures

As discussed in the section Existing Reconstruction Methods, many methods have been proposed for SCP. Among those we choose to compare RACOGS to SCWRL 3.0 because it is a popular, recent, readily available, and very fast method.[50] SCWRL 3.0 has been shown to have comparable accuracy to other recent methods.[39,52]

We tested RACOGS and SCWRL 3.0 on a set of 2945 nonredundant protein structures culled from the PDB, which contain the positions of all of the heavy atoms. To equalize the test we use only the C_α coordinates from the PDB structures as input for the backbone reconstruction method described in the section Backbone Reconstruction. The reconstructed backbones are then given either to the SCP step in RACOGS or to SCWRL 3.0. In both cases no all-atom minimization was performed. Then the side-chain RMSD between the structures produced by the two methods and the original PDB structures is compared. As shown in Figure 2, even if SCWRL 3.0 performs slightly

DOI 10.1002/prot

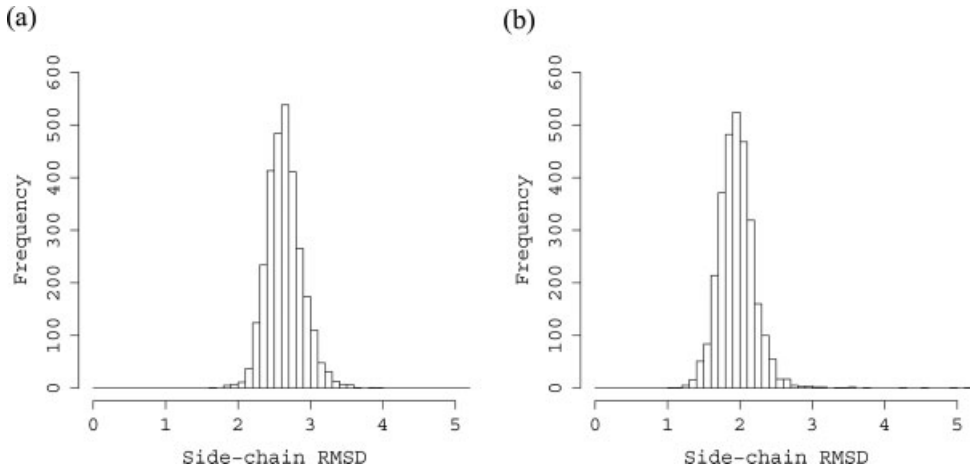


Figure 2

*Histograms of side-chain RMSD for PDB structure reconstruction: (**a**) Results obtained with RACOGS, with a median RMSD of 2.62Å; (**b**) Results obtained with SCWRL 3.0, with a median RMSD of 1.94 Å.*

better (with a median of 1.94 Å RMSD versus a median of 2.62 Å RMSD obtained with RACOGS) our method has competitive accuracy when reconstructing PDB structures. In the rest of the results we show that for the purpose of reconstructing coarse-grain simulations RACOGS performs drastically better.

Reconstruction of coarse-grain simulations

As stated above, the main purpose of RACOGS is to enable multiscale modeling of proteins by consistently reconstructing all-atom details from coarse-grain structures obtained by simulations using minimalist protein models. We present in this section the results of a large-scale reconstruction of the folding landscapes of two protein systems.

Model systems in coarse-grain simulations: src-SH3 and S6

The two proteins used in the coarse-grain simulations are the src-SH3 domain (residues 84–140 from PDB code 1FMK) and ribosomal protein S6 (PDB code 1RIS). The src-SH3 domain was chosen because its folding/unfolding process has been extensively studied by experiment, theory, and simulations. The protein contains β-sheets packed orthogonally, which form a hydrophobic core. The β-sheets are connected by the RT, n-src, and distal loops. The src-SH3 domain is found in proteins involved in signal transduction and cytoskeleton components.[61] It has been a model system in studying protein folding because it is relatively small, folds independently and can be modeled by two-state kinetics.[62]

The ribosomal protein S6 is one of many small protein subunits found in the ribosome. It binds to RNA and ribosomal protein S18 during the formation of the 30S ribosomal subunit.[63] It consists of four anti-parallel β-sheets and two α-helices, which create a hydrophobic core.[64] Experiments have shown that S6 can also be modeled by two-state kinetics.[65,66] We analyze both the wild-type S6 ($S6^{wt}$) and a mutant ($S6^{Alz}$) obtained upon the mutations EA41/EI42/RM46/RV47. This mutant is referred to as $S6^{Alz}$ as this set of mutations causes the protein to become highly homologous to the Alzheimer peptide and it has been shown to significantly increase the aggregation propensity of S6.[67]

Coarse-grain model used

All-atom reconstruction was performed on 606,000 coarse-grain structures for each of src-SH3, $S6^{wt}$, and $S6^{Alz}$. The coarse-grain structures were obtained from simulations using a minimalist protein model at the folding temperature of the proteins. The simulations extensively sample the folding landscape from the completely unfolded to the completely folded states. The details of the model used are described in Das et al.[14] In the case of S6 the model is augmented with experimental data as described in Matysiak and Clementi[15,68] The coarse-grain model takes into account both sequence information and energetic frustration to provide a realistic picture of a protein during the folding process. In this model the protein is represented by only the coordinates of the C_α atoms.

Comparison to an existing SCP method

The performance of RACOGS is again compared with SCWRL 3.0, but this time on coarse-grain structures instead of PDB structures. To compare the two methods we substitute SCWRL 3.0 for the SCP step of RACOGS.

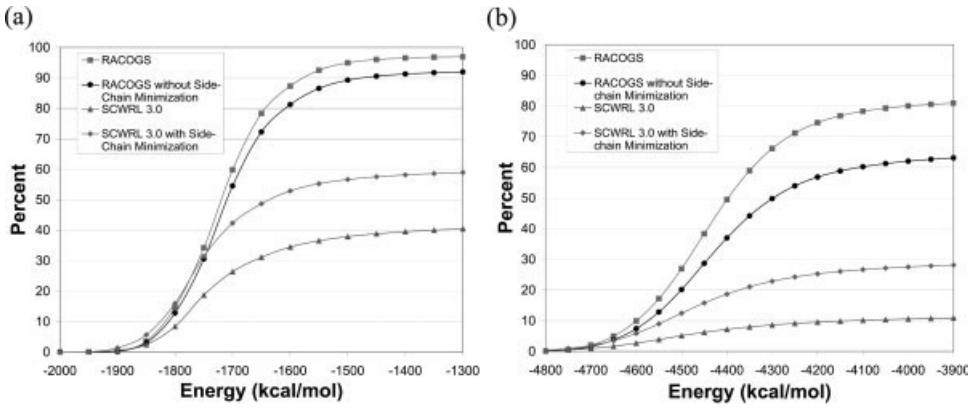


Figure 3

*Percentage of all-atom structures with energy below a given value as produced from coarse-grain structures of (**a**) src-SH3 and (**b**) S6^{wt}. Results from RACOGS and SCWRL 3.0 (both with and without side-chain minimization) are compared. The remaining structures have unreasonably high energy mainly because of steric clashes.*

The backbone atoms were added using the method described in the Backbone Reconstruction section to the coarse-grain structures. The reconstructed backbone was then used as input for SCWRL 3.0. The output structures from SCWRL 3.0 were minimized as described in the section All-Atom Minimization. The energies of these all-atom minimized structures were then compared with the energies of the all-atom minimized structures produced by RACOGS for the same coarse-grain structures.

As discussed previously, the performance is evaluated by considering the number of low-energy (according to the Boltzmann criterion), all-atom structures produced from the coarse-grain structures. The distribution of energy values for structures produced by each method are shown in Figure 3 for src-SH3 and S6^{wt}. RACOGS is able to produce substantially more low-energy structures than SCWRL 3.0. In the case of src-SH3 the number of low energy structures produced by SCWRL 3.0 with all-atom minimization is less than 50% of the total number of structures. Using RACOGS, 95% of the reconstructed structures of src-SH3 have low-energy. For S6^{wt} less than 20% of the structures have low energy when using SCWRL 3.0. In contrast, more than 80% of the all-atom structures of S6 have low energy when using RACOGS.

SCWRL 3.0's performance when reconstructing coarse-grain structures is somewhat surprising because this method performed well on the PDB structures and has been used quite successfully in several applications such as homology modeling and structure prediction.[69] However, the difference in performance may be explained by considering that while both methods on the surface address positioning of side-chains, they were designed with very different goals in mind. SCWRL 3.0 was designed to mainly be used in protein structure prediction applications. RACOGS was designed to reconstruct low-energy, all-atom structures from coarse-grain structures obtained from simulations where the protein undergoes large conformational changes, visiting regions far from the native state. Additionally, SCWRL 3.0 uses a simplified energy function to generate its output, but our final evaluation is done using AMBER99. Switching between two different energy functions may have an influence on the results. Further investigation into these issues is outside of the scope of this paper, but may help develop improved reconstruction methods in the future. These results show that developing and testing methods by how well they reconstruct PDB structures may not be the best measure to use when choosing a method to reconstruct coarse-grain simulations.

Effect of the side-chain minimization step

To assess how the side-chain minimization step influenced the reconstruction method, we analyzed two variants of RACOGS: one without the side-chain minimization step and one with the side-chain minimization step. This same comparison was made for the structures reconstructed by SCWRL 3.0. The side-chain minimization step was first performed on the output structures of SCWRL 3.0, and then the all-atom minimization was performed.

We compared the number of low-energy, all-atom structures produced by RACOGS and SCWRL 3.0, including and excluding the side-chain minimization step. The results are shown in Figure 3 for src-SH3 and S6^{wt}. The side-chain minimization step substantially improves the number of low energy all-atom structures recovered for both RACOGS and SCWRL 3.0. With the addition of the side-chain minimization step, SCWRL 3.0 still does not produce as many low-energy, all-atom structures as RACOGS.

DOI 10.1002/prot

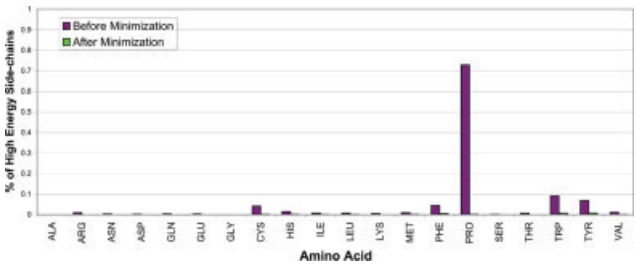


Figure 4

Percent of high energy (>10 kcal/mol) side-chain interactions before and after the side-chain minimization step, for each of the twenty amino-acids. These results correspond to the reconstruction of PDB structures. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

In the case of src-SH3, not performing the side-chain minimization step results in ~5% more of the RACOGS structures and 20% more of the SCWRL 3.0 structures having physically unrealistic high energy than with the step included. The results are more dramatic for $S6^{wt}$; about 20% more of both the RACOGS and the SCWRL 3.0 structures have unreasonably high energy when the side-chain minimization step is not used. A closer look at the interaction energies of specific amino acids provides a better understanding of how the side-chain minimization step is affecting the all-atom structures. PDB structures were first used to address the question and to assess how well the side-chain minimization step was able to reduce the energy for different types of amino acids. The same set of PDB structures used in the section Reconstruction of PDB Structures was used in this analysis.

The positions of the C_α atoms for each PDB structure were used as input for RACOGS. We minimized any side-chain that had energy greater than 10 kcal/mol. We then counted the number of times a side-chain caused interaction energy greater than 10 kcal/mol before and after the side-chain minimization step for each type of amino acid. The results are plotted in Figure 4. Proline caused by far the most high-energy interactions, followed by the bulky amino acids tryptophan, tyrosine and phenylalanine. However, the side-chain minimization step is able to successfully reduce these high energy interactions. As shown in Figure 4, almost all of the high energy side-chains interactions are eliminated after the side-chain minimization step. This result holds for other amino acids besides proline as well. The structures of src-SH3 and S6 reconstructed using RACOGS present the same trend. Figures 5 and 6 show the number of high energy side-chains for src-SH3 and S6, respectively, over all 606,000 structures before and after the side-chain minimization step during the reconstruction process. For both proteins we see that the proline side-chains are causing the large majority of high-energy interactions, consistent with the results found using the PDB structures. Again, the side-chain minimization step fixes many of the high energy side-chain interactions. This is one of the main reason that using side-chain minimization step produces a much larger fraction of all-atom structures with reasonable energy. Overall the side-chain minimization step is able to produce more low-energy structures without significantly increasing the running time of the algorithm.

Free energy landscape comparisons

The results presented in the previous sections suggest that RACOGS can be efficiently used to process large ensembles of coarse grain structures as starting points to characterize the dynamics of a protein system in all-atom detail. However, caution is needed when hopping between models at different resolutions. As significantly different energy functions are associated with the coarse grain and all-atom model, there is a priori no guarantee that the landscape sampled by one model is representative of the landscape corresponding to the other. A poor coarse grain model could mainly sample regions that are not significant when considered in the all-atom model. In such a case we expect the free energy surface defined by using the all-atom reconstructed structures to appear quite different from the corresponding free energy surface calculated from the coarse-grain structures, as a result of the different Boltzmann weights associated to the structures in the two models. From this point of view, the fact that the RACOGS-reconstructed all-atom and coarse-grain free energy landscape remain remarkably similar for both src-SH3 and S6 proteins (as shown in the following sec-

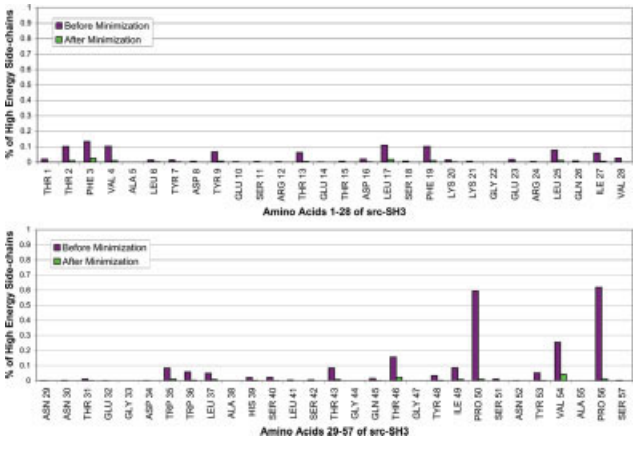


Figure 5

Percent of high energy (>10 kcal/mol) side-chain interactions before and after the side-chain minimization step for each amino-acid. These results correspond to the in the reconstruction of src-SH3 structures from coarse-grain configurations sampled during extensive folding/unfolding simulations. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

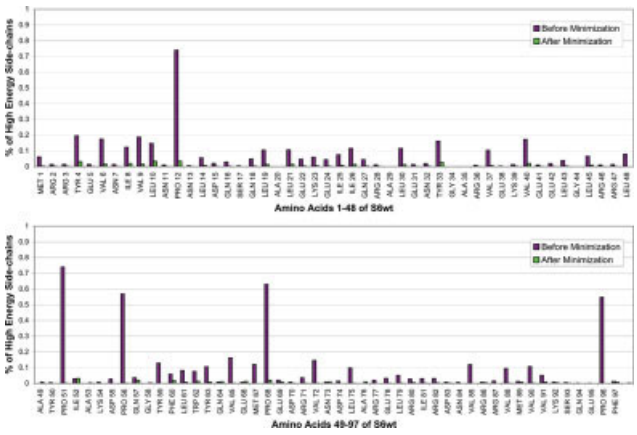


Figure 6

Percent of high energy (>10 kcal/mol) side-chain interactions before and after the side-chain minimization step for each amino acid. These results correspond to the in the reconstruction of structures of S6wt from coarse-grain configurations sampled during extensive folding/unfolding simulations. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

tions) represents a significant result. Clearly this does not represent a full proof of the consistency between the coarse grain and all-atom model used, as relevant configurations may have escaped the coarse sampling, and the detection of additional relevant regions of the landscape may be missed. However, the results presented in the following prove that it is possible to use a good coarse-grain model at least as a robust starting point for an extensive sampling of complex protein landscapes at all-atom resolution.

Although the determination of appropriate reaction coordinates for the definition of free energy surfaces is an area of active research,[70–72] the free energy presented in this section are all obtained as a function of the coordinates Q, the fraction of native, and A, the fraction of nonnative contacts. This choice of reaction coordinates is motivated by the fact that we want to compare the free energy landscapes associated with the all-atom reconstructed structures with the corresponding coarse grain landscapes, that have been originally calculated and validated by using this set of reaction coordinates, both for src-SH3[14] and S6.[15] In the definition of the parameters Q and A contacts are considered as native or nonnative based on their probability of forming in the native state.[73] A contact between a pair of residues is considered native if the probability of formation is >0.85 over all configurations with C_α RMSD < 2.5 Å from the crystal structure. If the probability of formation is <0.01 over the same set of structures the contact is considered nonnative.

The final energy after the all-atom minimization is used as input to the weighted histogram analysis method (WHAM)[74] to calculate the free energy in the all-atom model. The resulting free energy landscapes for the coarse-grain and all-atom models of src-SH3 are plotted in Figure 7. The free energy landscape computed using the low energy structures obtained using RACOGS is highly similar to the coarse-grain landscape. The folded, transition and unfolded states remain in place and no overall distortion is introduced into the landscape upon reconstruction.

The free energy barrier between the folded and unfolded states in the all-atom landscape of src-SH3 is calculated to be $\Delta G/RT_f \approx 2.5 \pm 0.3$. This value is in good agreement with the free energy barrier $\Delta G/RT_f \approx 2 \pm 0.4$ calculated using the coarse-grain model.[14] As the definition of reaction coordinates represents the main source of error in the calculation of free energy differences, the error reported on the free energy barrier is estimated as the largest difference obtained when considering different sets of reaction coordinates, as in previous work.[14,15]

A folding temperature can be estimated from the RACOGS-reconstructed all-atom structures of src-SH3, as

Figure 7

*The free energy landscape of src-SH3 at the folding temperature, obtained using (**a**) coarse-grain (**b**) all-atom structures reconstructed with RACOGS. The free energy is calculated as a function of the fraction of native contacts,* Q, *and the fraction of nonnative contacts,* A. *Each contour level marks a free energy change of 1 RT_f.*

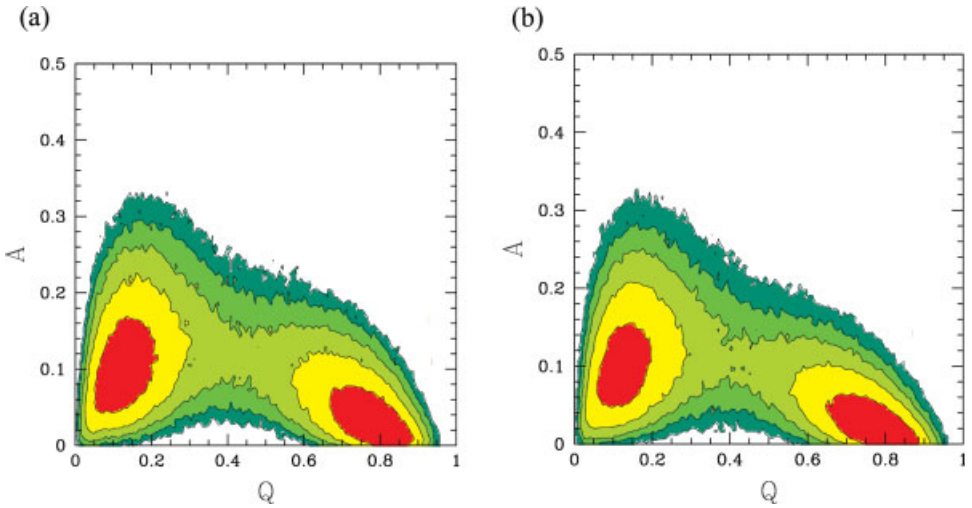
Figure 8

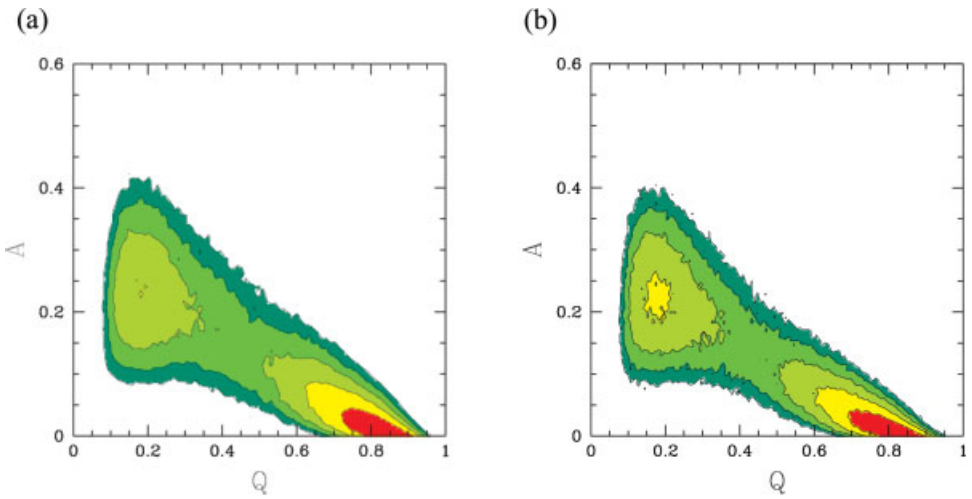
*The free energy surface of $S6^{wt}$ at the folding temperature, calculated using (**a**) coarse-grain (**b**) all-atom structures reconstructed with RACOGS. The free energy is shown as a function of the fraction of native contacts,* Q, *and the fraction of nonnative contacts,* A. *Each contour level marks a free energy change of 1 RT_f.*

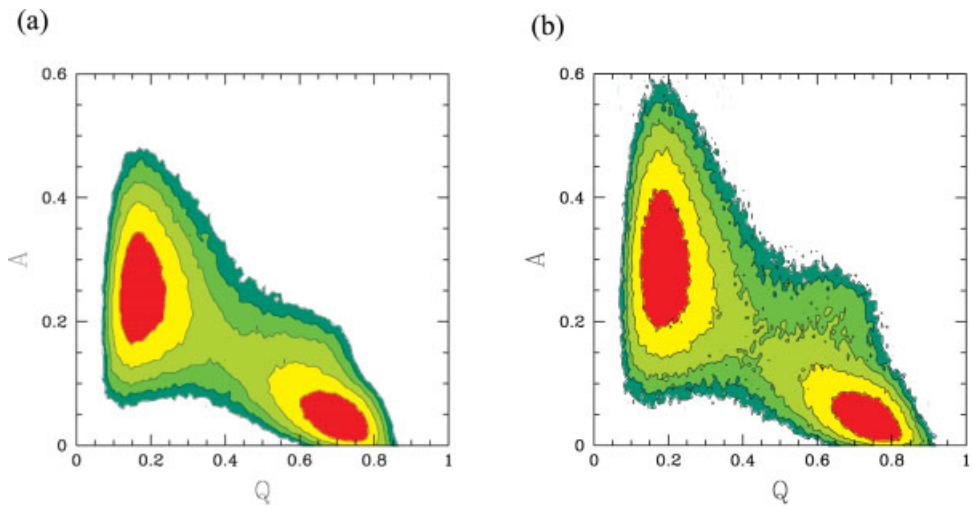
Figure 9

*The free energy landscape of $S6^{Alz}$ calculated using (**a**) coarse-grain or (**b**) all-atom structures, at the folding temperature. The free energy is plotted as a function of the fraction of native contacts,* Q, *and the fraction of nonnative contacts,* A.

DOI 10.1002/prot


(a)
(b)
A
Q
0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1


(a)
(b)
A
Q
0
0.2
0.4
0.6
0.8
1


(a)
(b)
A
Q
0
0.2
0.4
0.6
0.8
1

the peak of the heat capacity curve as a function of temperature. The resulting folding temperature is $T_f = 350 \pm 5$ K. This result is in remarkable agreement with the experimentally measured folding temperature of 356 K.[62]

The same free energy landscape comparison was performed for S6. The free energy plots were again calculated using the reaction coordinates Q and A along with the energy of the final all-atom minimized structure using WHAM. The landscapes for $S6^{wt}$ are shown in Figure 8, which shows a high level of similarity between the two free energy landscapes for the all-atom and the coarse-grain structures. The transition state occurs at the same place in both plots with the barriers remaining at a similar height: The free energy barrier in the all-atom landscape of $S6^{wt}$ is calculated to be $\Delta G/RT_f \approx 1.5 \pm 0.5$, in good agreement with the free energy barrier $\Delta G/RT_f \approx 1.7 \pm 0.7$ calculated using the coarse-grain model.[15] The folding temperature calculated by using the all-atom reconstructed structures of $S6^{wt}$ is $T_f = 384 \pm 5$ K. This value is again in remarkable agreement with the experimental folding temperature of 383 K.

"Zooming In" the misfolded states of $S6^{Alz}$

The free energy landscapes for $S6^{Alz}$ calculated using all-atom or coarse-grain structures are shown in Figure 9. Again, the folded and unfolded states stay quite similar between the coarse-grain and all-atom models. Previous studies on $S6^{Alz}$ have shown that this mutant can easily remain trapped in partially misfolded states during the folding process.[15] The population of these misfolded traps appear as a "bulge" in the free energy landscape around $Q \approx 0.7$ and $A \approx 0.2$, that is not present in the landscape of $S6^{wt}$ (see Figs. 8 and 9). Moreover, the comparison of Figures 8(a) and 9(a) shows that the position of the native state of $S6^{Alz}$ is shifted in the free energy landscape of $S6^{Alz}$ with respect to the native state of $S6^{wt}$. This shift is confirmed by a difference of ≈4.15 Å RMSD between the crystal structures of $S6^{wt}$ (PDB code 1RIS) and $S6^{Alz}$ (PDB code 1QJH) and can be explained by the increased flexibility and the formation of nonnative contacts detected in the native state of $S6^{Alz}$ (see Ref. 15 for detail). Figures 8(b) and 9(b) show that when the all-atom detail is added to the coarse-grain structure the main features associated with the free energy landscape of $S6^{Alz}$ are preserved. It is worth noting that the bulge in the landscape associated with the population of partially misfolded structures becomes larger and more distinct when all-atom detail is added, signaling that the misfolded states are partially stabilized in the all-atom structures. This is in good agreement with experimental results, detecting off-pathways traps and partially stable aggregates during the folding of $S6^{Alz}$.[67]

A closer look at the partially misfolded structures populated during the folding of $S6^{Alz}$ yields information on the misfolding mechanism. Figures 10(b,d) show two orientations of the structure representative of the most populated cluster emerging from a cluster analysis performed on all the structures in the bulge region of the all-atom reconstructed landscape. We used a simple "leader algorithm"[75] to perform the clustering. The distance between each structure was measured using the RMSD calculated over all of the heavy atoms, with the cutoff distance for each cluster set to RMSD = 5Å. The crystal structure (i.e., the native state of $S6^{wt}$) is shown for comparison in Figure 10(a,c) [with the same orientations as in Fig. 10(b,d), respectively]. Figure 10 reveals that while β-strands 1 and 2 do not interact in the correctly folded structures (as they reside at apposite sides of Strand 3 in the β-sheet), these two strands pack against each other in the misfolded structure. Moreover, while Strands 1, 3, and 4 retain an almost-native structure, larger differences are detected in Strand 2. In particular, Strand 2 migrates toward the interior of the protein, disrupting the packing of the hydrophobic core. Figure 10 shows that the reconstructed all-atom structure is stabilized by the formation of multiple interactions between the side-chains of Strands 1 and 2 and the overall repacking of the four β-strands. Interestingly, Strand 2 represents the part of the protein involved in the formation of interprotein interactions in the tetrameric crystal structure of $S6^{Alz}$.[67] The same stretch of residues is the segment of $S6^{Alz}$ with increased homology to the Alzheimer peptide β-AP, and it thought to be responsible for the increased aggregation propensity of the protein.[15,67] Moreover, recent experimental investigations have shown that Strand 2 is not part of the folding nucleus of S6, neither for the wild-type nor for any of its circular permutants (Mikael Oliveberg, personal communication), leaving open the possibility of populating a misfolded state characterized by the mispacking of Strand 2 either in the late stages of the folding process, or even as an alternative state accessible by fluctuation from the native state. The formation of this mispacked structure may trigger the observed misfolding and aggregation of $S6^{Alz}$.

The fact that the introduction of all-atom detail in protein configurations obtained with a coarse grain model creates physically relevant misfolded structures supports the idea that simplified models can indeed provide a robust starting point to characterize even complex folding scenarios, and multiscale strategies built on these models may offer a powerful tool to investigate the interplay between folding/misfolding/aggregation mechanisms.[15,76–79]

Probing the TSE of src-SH3

Coarse-grain simulations provide an extensive sampling of protein conformations over a long time scale. However, moving to an all-atom representation is necessary to perform a detailed structural analysis of the conformational states of the protein. RACOGS allows this analysis to be per-

DOI 10.1002/prot

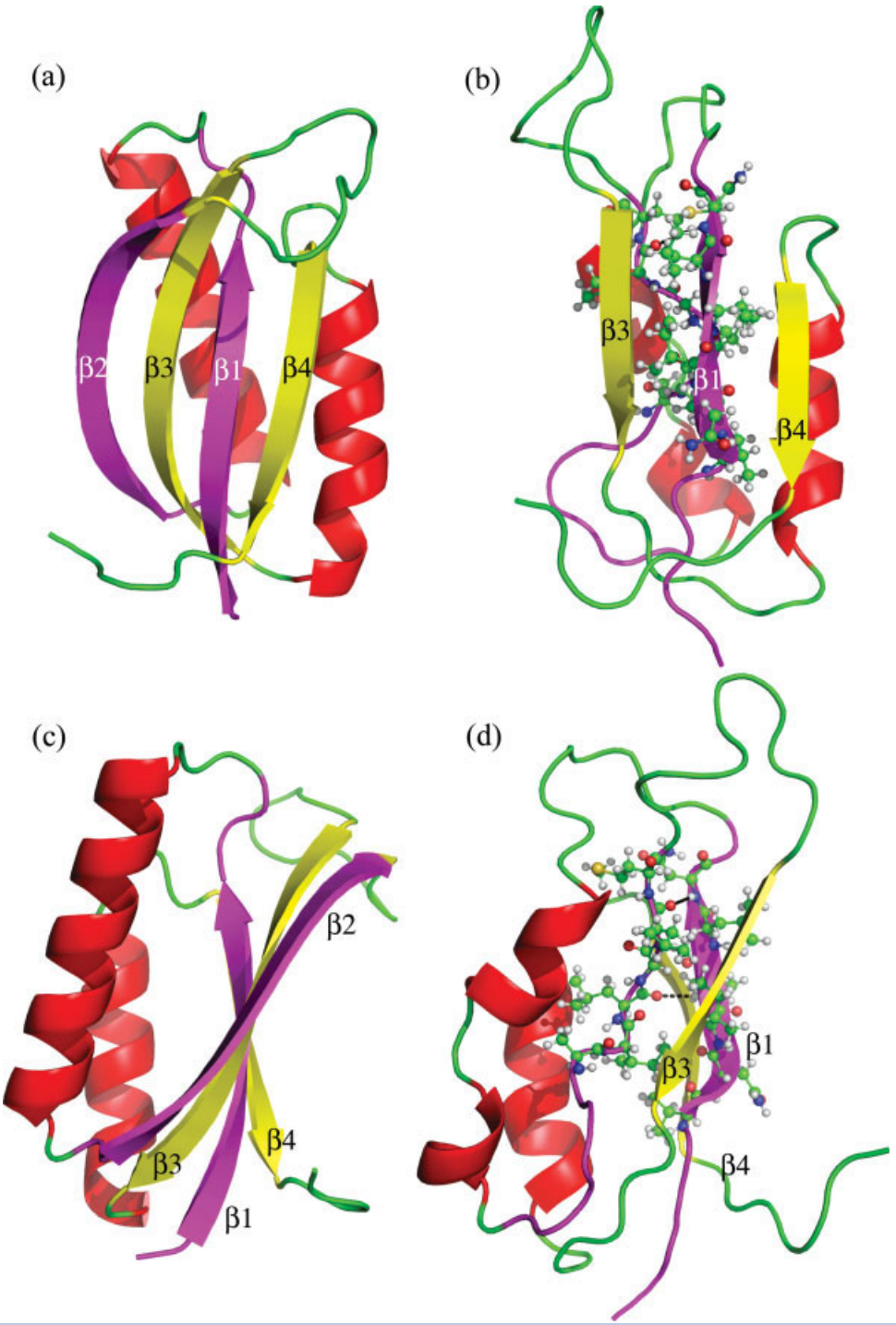


Figure 10

*Two orientations of a representative misfolded structure of $S6^{Alz}$ are compared with the corresponding orientations of the native structure of $S6^{wt}$. The front view of the native structure is shown in (**a**), and the corresponding view of the misfolded state is shown in (**b**), while the side view of the same native and misfolded structure are shown in (**c**) and (**d**), respectively. The comparison of these structures reveals that the misfolding is mainly caused by the mispacking of strand β2. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]*

formed on proteins simulated using coarse-grain techniques. Since there is a relatively large body of both experimental and theoretical knowledge on the folding mechanism of src-SH3,[62,73,80–91] we can compare our set of all-atom structures of src-SH3 with previous results, particularly on the characterization of the TSE of this protein. In

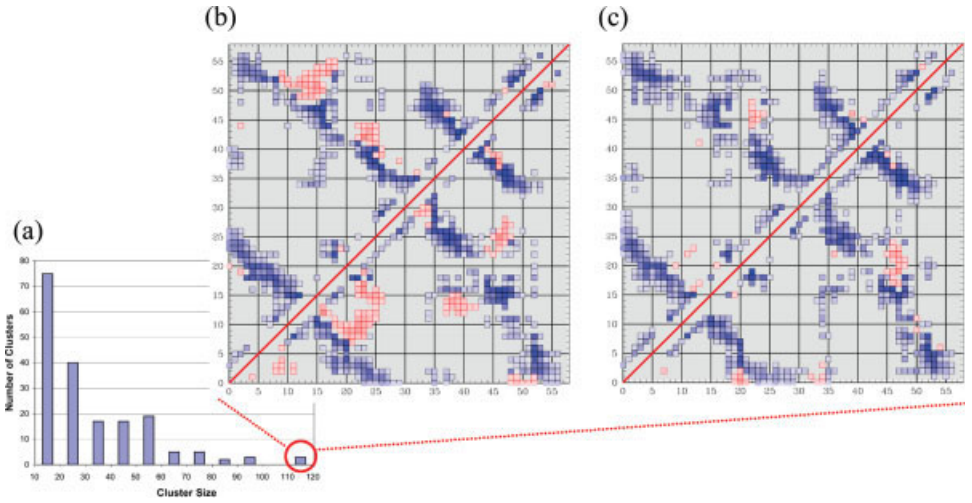


Figure 11

*(**a**) Distribution of cluster size for the cluster obtained in the analysis of the all-atom reconstructed transition state ensemble of src-SH3. The all-atom contact maps for the four most populated clusters are shown: (**b**) Top half of the map corresponds to the most populated cluster, the bottom half to the second most populated; (**c**) Top half corresponds to the third most populated cluster, bottom half to the fourth. Different shades of blue are used to illustrate different probability of formation for the native contacts, from white to deep blue (the contacts with higher probabilities are in deep blue, lower probabilities are in white). Different shades of red are used for the nonnative interactions, from white to deep red.*

the following we present a detailed comparison between the all-atom structures produced by RACOGS and what is known about the TSE of src-SH3. Following Das et al.,[14] the TSE was determined as the top of the free energy barrier, by using the reaction coordinates Q and A. Low energy all-atom structures with a value of $Q \in (0.4–0.5)$ and a value of $A \in (0.08–0.14)$ were considered TSE structures. This selection produced 10,044 structures.

Cluster analysis of the TSE structures

A cluster analysis was performed on the 10,044 all-atom structures representing the transition state of src-

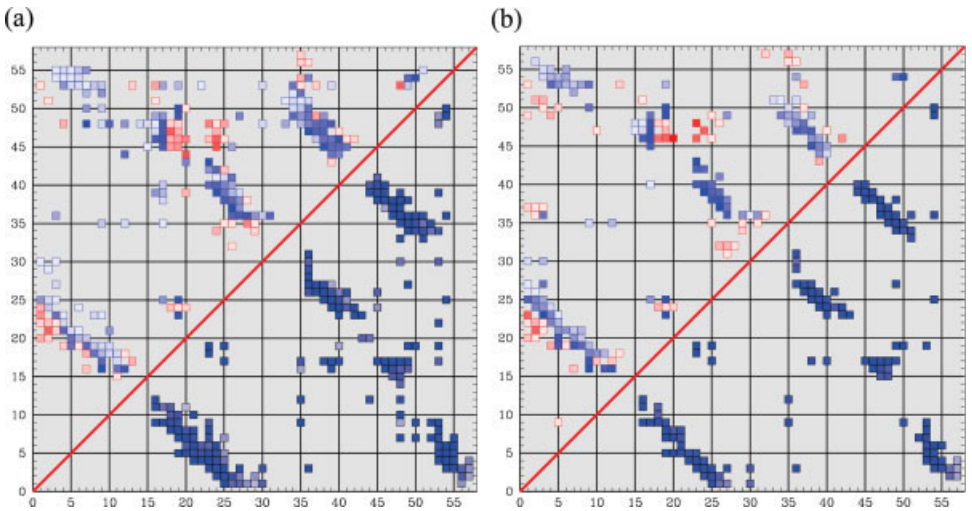


Figure 12

*Contact map for (**a**) coarse-grain structures (**b**) all-atom structures of src-SH3 reconstructed with RACOGS. The upper half of each map corresponds to the transition state ensemble; the lower half to the folded state. The coloring is the same as in Figure 11.*

DOI 10.1002/prot

SH3, to identify which structural components more probably formed in the TSE. As in the cluster analysis for S6, a simple "leader algorithm"[75] was used for the clustering, with the cutoff distance for each cluster set to RMSD = 5 Å. The analysis produced a total of 1745 clusters, with sizes distributed as illustrated by Figure 11(a). Figure 11(b,c) shows the contact maps associated with the four most populated clusters (each representing more than 100 structures). The contacts in Figure 11 are "all-atom contacts", that is, two amino acids are considered in contact if any of their heavy atoms are within 4.5 Å of each other. All-atom contacts are also identified as native or nonnative by their probability of forming in the native state. If the probability of a contact forming in the native state is higher than in the transition state then the contact is considered native, otherwise it is considered nonnative.

These contact maps illustrate similar secondary structure formation in the most highly populated clusters. The central three-stranded β-sheet is well formed, together with the interactions involving the distal loop and diverging turn, which is in good agreement with experimental results[86] and previous computational studies.[80,87–89] The main difference between the clusters illustrated Figure 11(b) consists in the formation of different set of nonnative contacts, although there is an overall tendency to form nonnative interactions within the hydrophobic core of the protein in full agreement with the experimentally detected formation of a nonspecific hydrophobic cluster of nonnative contacts in the TSE of SH3.[14,73,83,91]

Contact map analysis of the TSE

The overall average features of the TSE associated with the coarse-grain and all-atom landscapes can be compared by means of C_α contact maps. Figure 12 illustrate the results. The average C_α contact map computed over all the TSE structures obtained from the coarse-grain simulation of src-SH3[14] is shown in part (a) of Figure 12, while the average map computed from the RACOGS all-atom structures is shown in part (b). In each map, the bottom right half represents the native state of the protein and the top left half is the transition state. Figure 12(a) shows that a cluster of nonnative contacts is formed around residues 40–50 and residues 15–25 in the TSE associated with the coarse-grain folding simulation of src-SH3 (see Ref. 14 for detail). As mentioned in the previous section, it has been shown that the formation of nonnative contacts in the hydrophobic core plays an important role in the folding process of src-SH3.[73,83,91] Figure 12(b) shows that the all-atom TSE reconstructed by RACOGS retains this cluster of nonnative contacts, consistently with the results obtained from the cluster analysis. Similar to what observed for $S6^{Alz}$, a few nonnative contacts appear stabilized upon the reinsertion of all-atom detail.

CONCLUSIONS

This work provides a solid starting point for multiscale protein simulations by examining the transition from coarse-grain to all-atom models. A number of coarse-grain models developed in the last decade reduce the structural detail of proteins in order to reach longer timescales in simulations. However, there has been little work on how to go in the opposite direction, from coarse-grain models to all-atom models, which is essential for multiscale techniques. We have filled this gap by providing an efficient and reliable method for producing low energy, all-atom structures from coarse-grain protein configurations called RACOGS. RACOGS was thoroughly tested and validated on both PDB and coarse-grain structures from simulations. The results showed that a key step of the method is the side-chain minimization, which substantially increased the number of low-energy, all-atom structures produced from coarse-grain structures, particularly in regions far from the native state. The reconstructed all-atom structures were used to calculate free energy landscapes for src-SH3 and S6. A comparison with the free energy landscapes calculated using the coarse-grain structures showed no apparent distortion. Additionally, further examination of the all-atom structures obtained in the misfolded region of $S6^{Alz}$ and in the TSE src-SH3 showed good agreement with previous experimental and computational evidence.

By demonstrating that it is feasible to reliably and quickly move between a coarse-grain model and an all-atom model, this work has opened a door for future work on multiscale simulations.

ACKNOWLEDGMENTS

APH is supported by a NSF Graduate Research Fellowship. The Rice University Cray XD1 Research Cluster used for the calculations is supported in part by a Major Research Infrastructure grant from NSF, Rice University and partnerships with AMD and Cray. We acknowledge Payel Das for her contributions to the initial stages of this project, and Silvina Matysiak for her help on the coarse-grain modeling of S6. We are grateful to Mikael Oliveberg for insightful discussions and for sharing with us the experimental data on S6.

REFERENCES

1. Praprotnik M, Delle Site L, Kremer K. Adaptive resolution molecular-dynamics simulation: changing the degrees of freedom on the fly. J Chem Phys 2005;123:224106.
2. Neri M, Anselmi C, Cascella M, Maritan A, Carloni P. Coarse-grained model of proteins incorporating atomistic detail of the active site. Phys Rev Lett 2005;95:218102.
3. Shi Q, Izvekov S, Voth GA. Mixed atomistic and coarse-grained molecular dynamics: simulation of a membrane-bound ion channel. J Phys Chem B 2006;110:15045–15048.

4. Fan ZZ, Hwang JK, Warshel A. Using simplified protein representation as a reference potential for all-atom calculations of folding free energy. Theor Chem Acc 1999;103:77–80.
5. De Mori GM, Colombo G, Micheletti C. Study of the Villin headpiece folding dynamics by combining coarse-grained Monte Carlo evolution and all-atom molecular dynamics. Proteins 2005;58:459–471.
6. Ding F, Guo W, Dokholyan NV, Shakhnovich EI, Shea JE. Reconstruction of the src-SH3 protein domain transition state ensemble using multiscale molecular dynamics simulations. J Mol Biol 2005;350:1035–1050.
7. Lyman E, Ytreberg FM, Zuckerman DM. Resolution exchange simulation. Phys Rev Lett 2006;96:028105.
8. Christen M, van Gunsteren WF. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. J Chem Phys 2006;124:154106.
9. Kwak W, Hansmann UH. Efficient sampling of protein structures by model hopping. Phys Rev Lett 2005;95:138102.
10. Hess B, Len S, van der Vegt N, Kremer K. Long time atomistic polymer trajectories from coarse grained simulations: bisphenol-A polycarbonate. Soft Matter 2006;2:409–414.
11. Praprotnik M, Delle Site L, Kremer K. Adaptive resolution scheme for efficient hybrid atomistic-mesoscale molecular dynamics simulations of dense liquids. Phys Rev E 2006;73:066701.
12. Bedrov D, Ayyagari C, Smith G. Multiscale modeling of poly(ethylene oxide)-poly(propylene oxide)-poly(ethylene oxide) triblock copolymer micelles in aqueous solution. J Chem Theory Comput 2006;2:598–606.
13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
14. Das P, Matysiak S, Clementi C. Balancing energy and entropy: a minimalist model for the characterization of protein folding landscapes. Proc Natl Acad Sci USA 2005;102:10141–10146.
15. Matysiak S, Clementi C. Minimalist protein model as a diagnostic tool for misfolding and aggregation. J Mol Biol 2006;363:297–308.
16. Bassolino-Klimas D, Bruccoleri RE. Application of a directed conformational search for generating 3-D coordinates for protein structures from α-carbon coordinates. Proteins 1992;14:465–474.
17. Rey A, Skolnick J. Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. J Comput Chem 1992;13:443–456.
18. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. Calculation of protein backbone geometry from α-carbon coordinates based on peptide-group dipole alignment. Protein Sci 1993;2:1697–1714.
19. Jones TA, Thirup S. Using known substructures in protein model building and crystallography. EMBO J 1986;5:819–822.
20. Claessens M, Van Cutsem E, Lasters I, Wodak S. Modelling the polypeptide backbone with 'spare parts' from known protein structures. Protein Eng 1989;2:335–345.
21. Reid LS, Thornton JM. Rebuilding flavodoxin from Cα coordinates: a test study. Proteins 1989;5:170–182.
22. Holm L, Sander C. Database algorithm for generating protein backbone and side-chain co-ordinates from a Cα trace: application to model building and detection of co-ordinate errors. J Mol Biol 1991;218:183–194.
23. Payne PW. Reconstruction of protein conformations from estimated positions of the Cα coordinates. Protein Sci 1993;2:315–324.
24. Milik M, Kolinski A, Skolnick J. Algorithm for rapid reconstruction of protein backbone from α carbon coordinates. J Comput Chem 1997;18:80–85.
25. Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL. Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. Proteins 2000;41:86–97.
26. Wang Y, Huq HI, de la Cruz XF, Lee B. A new procedure for constructing peptides into a given calpha chain. Fold Des 1998;3:1–10.
27. Dunbrack RL. Rotamer libraries in the 21st century. Curr Opin Struct Biol 2002;12:431–440.
28. Pierce NA, Winfree E. Protein design is NP-hard. Protein Eng 2002;15:779–782.
29. Chazell B, Kingsford C, Singh M. A semidefinite programming approach to side chain positioning with new rounding strategies. INFORMS J Comput 2004;16:380–392.
30. Huang ES, Koehl P, Levitt M, Pappu RV, Ponder JW. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. Proteins 1998;33:204–217.
31. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
32. Desmet J, De Maeyer M, Hazers B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature 1992;356:539–542.
33. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophys J 1994;66:1335–1340.
34. Keller DA, Shibata M, Marcus E, Ornstein RL, Rein R. Finding the global minimum: a fuzzy end elimination implementation. Protein Eng 1995;8:893–904.
35. De Maeyer M, Desmet J, Lasters I. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. Fold Des 1997;2:53–66.
36. Pierce NA, Spriet JA, Desmet J, Mayo SL. Conformational splitting: a more powerful criterion for dead-end elimination. J Comput Chem 1999;21:999–1009.
37. Gordon DB, Mayo SL. Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure 1999;7:1089–1098.
38. Looger LL, Hellinga HW. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. J Mol Biol 2001;307:429–445.
39. Liang S, Grishin NV. Side-chain modeling with an optimized scoring function. Protein Sci 2002;11:322–331.
40. Lee C, Subbiah S. Prediction of protein side-chain conformation by packing optimization. J Mol Biol 1991;217:373–388.
41. Bower MJ, Cohen FE, Dunbrack RL. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
42. Dunbrack RL, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 1993; 230:543–574.
43. Tuffery P, Etchebest C, Hazout S, Lavery R. A new approach to the rapid determination of protein side chain conformations. J Biomol Struct Dyn 1991;8:1267–1289.
44. Pedersen JT, Moult J. Genetic algorithms for protein structure prediction. Curr Opin Struct Biol 1996;6:227–231.
45. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol 1994;239:249–275.
46. Lee C. Predicting protein mutant energetics by self-consistent ensemble optimization. J Mol Biol 1994;236:918–939.
47. Mendes J, Baptista AM, Carrondo MA, Soares CM. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. Proteins 1999;37:530–543.
48. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. Proteins 1998;33:227–239.
49. Samudrala R, Moult J. A graph-theoretic algorithm for comparative modeling of protein structure. J Mol Biol 1998;279:287–302.
50. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12: 2001–2014.
51. Dukka Bahadur KC, Tomita E, Suzuki J, Akutsu T. Protein side-chain packing problem: a maximum edge-weight clique algorithmic approach. J Bioinform Comput Biol 2005;3:103–126.
52. Kingsford CL, Chazelle B, Singh M. Solving and analyzing side-chain positioning problems using linear and integer programming. Bioinformatics 2005;21:1028–1036.

 DOI 10.1002/prot

53. Eriksson O, Zhou Y, Elofsson A. Side chain-positioning as an integer programming problem. Proceedings of the first international workshop on algorithms in bioinformatics. London, UK: Springer-Verlag; 2001. pp 128–141.
54. Desmet J, Spriet J, Lasters I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. Proteins 2002;48:31–43.
55. Wang J, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? J Comput Chem 2000;21:1049–1074.
56. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized born model. Proteins: Struct Funct Genet 2004;55:383–394.
57. Eisenmenger F, Argos P, Abagyan R. A method to configure protein side-chains from the main-chain trace in homology modelling. J Mol Biol 1993;231:849–860.
58. Meza JC. OPT++: an object-oriented class library for nonlinear optimization. Technical Report, Technical Report SAND94-8225. Albuquerque, NM: Sandia National Laboratories; 1994.
59. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein–protein docking. Protein Sci 2005;14:1328–1339.
60. Case DA, Darden TA, Cheatham TE, III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Wang B, Pearlman DA, Crowley M, Brozell S, Tsui V, Gohlke H, Mongan J, Hornak V, Cui G, Beroza P, Schafmeister C, Caldwell JW, Ross WS, Kollman PA. AMBER 8. San Francisco, CA: University of California; 2004.
61. Pawson T. Protein modules and signalling networks. Nature 1995;373:573–580.
62. Grantcharova VP, Baker D. Folding dynamics of the src SH3 domain. Biochemistry 1997;36:15685–15692.
63. Agalarov SC, Sridhar Prasad G, Funke PM, Stout CD, Williamson JR. Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. Science 2000;288:107–113.
64. Lindahl M, Svensson LA, Liljas A, Sedelnikova SE, Eliseikina IA, Fomenkova NP, Nevskaya N, Nikonov SV, Garber MB, Muranova TA. Crystal structure of the ribosomal protein S6 from *Thermus thermophilus*. EMBO J 1994;13:1249–1254.
65. Otzen DE, Kristensen O, Proctor M, Oliveberg M. Structural changes in the transition state of protein folding: alternative interpretations of curved chevron plots. Biochemistry 1999;38:6499–6511.
66. Otzen DE, Oliveberg M. Conformational plasticity in folding of the split β-α-β protein S6: evidence for burst-phase disruption of the native state. J Mol Biol 2002;317:613–627.
67. Otzen DE, Kristensen O, Oliveberg M. Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. Proc Natl Acad Sci USA 2000;97:9907–9912.
68. Matysiak S, Clementi C. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? J Mol Biol 2004;343:235–248.
69. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Sci 2005;14:1315–1327.
70. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES. On the transition coordinate for protein folding. J Chem Phys 1998;108:334–350.
71. Cho SS, Levy Y, Wolynes PG. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. Proc Natl Acad Sci USA 2006;103:586–591.
72. Das P, Moll M, Stamati H, Kavraki LE, Clementi C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc Natl Acad Sci USA 2006;103:9885–9890.
73. Clementi C, Plotkin SS. The effects of nonnative interactions on protein folding rates: theory and simulation. Protein Sci 2004;13:1750–1766.
74. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenburg JM. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 1992;13:1011–1021.
75. Hartigan J. Clustering algorithms. New York: Wiley; 1975.
76. Fawzi NL, Chubukov V, Clark LA, Brown S, Head-Gordon T. Influence of denatured and intermediate states of folding on protein aggregation. Protein Sci 2005;14:993–1003.
77. Tarus B, Straub JE, Thirumalai D. Probing the initial stage of aggregation of the Aβ(10-35)-protein: assessing the propensity for peptide dimerization. J Mol Biol 2005;345:1141–1156.
78. Friedel M, Shea JE. Self-assembly of peptides into a β-barrel motif. J Chem Phys 2004;120:5809–5823.
79. Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. J Mol Biol 2002;324:851–857.
80. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.
81. Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. Proc Natl Acad Sci USA 1999;96:11305–11310.
82. Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–11304.
83. Cobos ES, Filimonov VV, Vega MC, Mateo PL, Serrano L, Martinez JC. A thermodynamic and kinetic analysis of the folding pathway of an SH3 domain entropically stabilised by a redesigned hydrophobic core. J Mol Biol 2003;328:221–233.
84. Viguera AR, Serrano L. Bergerac-SH3: "frustation" induced by stabilizing the folding nucleus. J Mol Biol 2001;311:357–371.
85. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, Baker D. Experiment and theory highlight role of native state topology in SH3 folding. Nat Struct Biol 1999;6:1016–1024.
86. Grantcharova VP, Riddle DS, Baker D. Long-range order in the src SH3 folding transition state. Proc Natl Acad Sci USA 2000;97:7084–7089.
87. Guo W, Lampoudi S, Shea JE. Temperature dependence of the free energy landscape of the src-SH3 protein domain. Proteins 2004;55:395–406.
88. Shea JE, Onuchic JN, Brooks CL. Probing the folding free energy landscape of the Src-SH3 protein domain. Proc Natl Acad Sci USA 2002;99:16064–16068.
89. Gsponer J, Caflisch A. Molecular dynamics simulations of protein folding from the transition state. Proc Natl Acad Sci USA 2002;99:6719–6724.
90. Di Nardo AA, Korzhnev DM, Stogios PJ, Zarrine-Afsar A, Kay LE, Davidson AR. Dramatic acceleration of protein folding by stabilization of a nonnative backbone conformation. Proc Natl Acad Sci USA 2004;101:7954–7959.
91. Viguera AR, Vega C, Serrano L. Unspecific hydrophobic stabilization of folding transition states. Proc Natl Acad Sci USA 2002;99:5349–5354.