Folding pathway of the B1 domain of protein G explored by a multiscale modeling

Sebastian Kmiecik and Andrzej Kolinski

Faculty of Chemistry, University of Warsaw, L. Pasteura 1, 02-093 Warsaw, Poland

Correspondence should be addressed to A.K. *email: kolinski@chem.uw.edu.pl*

Summary

The understanding of the folding mechanisms of single domain proteins is an essential step in the understanding of protein folding in general. Recently, we developed a mesoscopic CABS protein model which was successfully applied in protein structure prediction, studies of protein thermodynamics and modeling of protein complexes. In the present research this model is employed in a detailed characterization of the folding process of a simple globular protein, the B1 domain of IgG-binding protein G (GB1). There is a vast body of experimental facts and theoretical findings for this protein. Performing unbiased, *ab initio* simulations, we demonstrated that the GB1 folding proceeds via the successive formation of an extended folding nucleus, followed by slow structure fine-tuning. Remarkably, a subset of native interactions drives the folding from the very beginning. The emerging comprehensive picture of GB1 folding perfectly matches and extends the previous experimental and theoretical studies.

Abbreviations used: cRMSD, coordinate root mean square deviation; MD, Molecular Dynamics; MC, Monte Carlo; T, reduced temperature; Tt, transition temperature.

Introduction

Significant theoretical and experimental research efforts are devoted to understand how proteins fold into their native structures. Determination of the folded structure is a priority for complete biochemical protein characterization. However, a detailed understanding of the folding process requires characterization of all alternative protein conformations that emerge along the folding pathway, including the unfolded state and partially folded intermediates. Elucidation of the principles governing the folding mechanism will have broad implications for predicting structure from sequence, protein design, and understanding of the formation and propagation of prions and amyloids.

Theoretical studies lead to better understanding of experimental results providing easy-to-interpret structural models. Molecular mechanics is a powerful method for studying complex molecular systems. However, there is a gap between the timescales of classical Molecular Dynamics (MD) simulation and the timescales of protein folding. Only small and ultra-fast folding (in a range of microseconds) proteins are now tractable by means of classical MD simulations[1]. An average protein folds orders of magnitude slower. Perhaps, the largest protein folded using the all-atom potential and electrostatic-driven Monte Carlo (MC) procedure was 46-residue staphylococcal protein A[2].

So far, for larger proteins, all-atom simulations of the entire folding process, from a random coil to the native state, are possible only for Go models. There have been a number of Go potential based studies of protein G using the simplified model[3], the all-atom model[4] or with a weak Go-like contribution to the applied force field[5]. In Go models, only native interactions are taken into account. Consequently, the lowest energy of the native conformation is guaranteed. The obvious weak point of such an approach is that the knowledge of the native structure is needed to construct the Go potential. A significant shortcoming also results from neglecting non-native interactions, thereby ignoring their sometimes important role in the folding mechanisms[6].

Due to the time scale limitations of the all-atom molecular mechanics reduced models offer the most promising possibilities to study large scale protein rearrangements, as it was recently demonstrated by Liwo et al.[7] Langevin dynamics with the physics-based united-residue force field (UNRES) was applied successfully to the folding of real proteins.

We have used a reduced protein lattice model and MC dynamics to perform equilibrium folding simulations at various folding stages, beginning from the denatured state.

The use of a reduced representation of polypeptide chains led to a significant reduction of the conformational space[8], thereby enabling the search for the native state at a reasonable time scale. Compared with the experimental results, we have obtained a similar sequence of folding events and have identified the interactions critical for the folding process.

Our simulations show that the folding of the GB1 domain is initiated by the formation of a specific nucleus involving the hydrophobic core residues. These residues were previously found by Shakhnovich and coworkers to participate in the specific nucleation event. The study employed all-atom MC simulations with the Go potential[4] and phi values derived from protein engineering as restraints[9]. Moreover, they have shown that the nucleus residues are evolutionary conserved among the proteins that share a similar fold but have very little sequence similarity[9]. This finding strongly suggests that the native fold topology is a main factor determining the character of the transition state. In this paper we present *ab initio* folding studies that are not driven by any restraints, nor by structure-specific potentials. This study confirms the specific nucleation process and provides a detailed description of the sequence of nucleation and folding events, from the highly denatured state to the formation of the native-like globule.

The CABS (CA – CB – Side chain) model[10] used in this work was successfully employed by the Kolinski-Bujnicki group during the 6th edition of CASP (Critical Assessment of Protein Structure Prediction), a community-wide testing experiment of protein structure prediction methods[11]. The approach ranked second best overall, and more importantly in the new fold category after Rosetta, the recombination of short fragments extracted from known protein structures[12]. Both methods employed a MC search of the conformational space. In comparative modeling cases CABS force field has been supplemented with weak spatial restraints derived from homologous templates. Based on the previous denatured state simulations, it was postulated that the CABS model can be used not only for protein structure prediction, but also to study the folding mechanism[13]. In the present study, we employ knowledge-based statistical potentials only, the same as in the *ab initio* protein structure prediction. The only way for the evolutionary information to enter the force field is the expected secondary structure (in the commonly used three-letter code), providing a weak bias toward protein-like local interactions. This is the only structure-specific input in the CABS *ab initio* modeling procedure, introduced to take advantage of a relatively high level of accuracy of the contemporary secondary structure prediction methods.

Results and Discussion

Key role of the second hairpin in GB1 folding

The B1 domain of streptococcal protein G is a small, very regular α/β structure composed of 56 amino acids. The fold[14] consists of a four-stranded β-sheet and an α-helix tightly packed against the sheet. The sheet can be described as consisting of two symmetrically spaced hairpins: the first one, formed by N-terminal strands and the first beta turn (β1-turn1-β2), and the second one, formed by C-terminal strands and the second beta turn (β3-turn2-β4). Numerous experimental and theoretical studies highlight the early formation of the second hairpin and its key role in the GB1 folding. The second hairpin was found to be stable in isolation[15], and protected from hydrogen/deuterium exchange early during the folding of the entire protein[16]. The isolated second β-hairpin folds on the 10 μs time scale, which is two orders of magnitude faster than the intact protein folding[17]. The second hairpin fragment became a model short peptide and its folding was extensively studied by MD and MC simulations[18]. The importance of the second hairpin was also confirmed by phi value analysis[19] which suggested presence of the well-formed second β-turn in the transition state ensemble.

The role of the supersecondary structural elements in GB1 folding was investigated by coarse-grained MC folding and unfolding MD simulations which indicated that the helix-second hairpin fragment can stabilize itself to some extent independently of the rest of the protein, while the first hairpin cannot[20].

Hydrogen-deuterium exchange and NMR studies suggested that three regions of GB1 may correspond to nucleation sites: the second β-hairpin, the middle of the α-helix and the N-terminus of the α-helix[16,21]. These findings were also found to be consistent with MD unfolding studies[22]. Full atom molecular dynamics in explicit solvent suggested that the β-sheet is more mobile, and might be expected to unfold earlier than the helix[23]. According to protein engineering studies of protein G, complex consequences of the folding kinetics of single point mutations in the helix may suggest its structural diversity during the folding[19]. However, effects of several mutations suggested that the helix's C-terminus is better defined than the rest of the helix at the folding transition state ensemble[19].

Protein G unfolding dynamics was also investigated by the older version of the CABS force field with a different chain representation (SICHO model)[24]. The first β-hairpin was found to unfold first and to be significantly less stable than the second β-hairpin. Here, we

present a much more detailed study. Extensive isothermal simulations from a highly denatured state to the folding temperature were performed, providing a detailed overview of the entire folding process. Changes of the protein properties with the system temperature (T) measured by CABS energy, similarly to the native structure (cRMSD) and radius of gyration, are shown in Fig. 1. Temperature dependence of the folding progress, with respect to the particular secondary structure elements, measured by the number of native contacts within the given substructure, is illustrated in Fig. 2. According to Fig. 2 the second hairpin and the terminal strands fold in a cooperative two-state fashion, while the helix in a continuous manner. Furthermore, the native arrangement of the terminal strands seems to have the largest contribution to the overall folding cooperativity. The first hairpin is definitely the least ordered substructure after the folding transition.

The early formation of the helix and the second β-hairpin is also apparent in residue-specific analysis of the MC trajectories. A native-like arrangement of the side-chain contacts in the transition structures at the transition temperature (Tt) can be clearly seen in Fig. 3d. Definitely, the helix is the most ordered residual structure at Tt. It is stabilized by a number of short-range contacts between the side chains separated in the sequence by 3 or 4 positions. Another noteworthy fragment of the local structure at Tt is the second β-hairpin (see Fig. 3d). However, with respect to the long-range interactions, the second hairpin is the main nucleation site, responsible for the early structure formation.

Nucleus residues: Y3, L5, F30, W43, Y45, F52

Changes of the protein properties with the system temperature (Fig. 1) suggest a two-state folding kinetics. Two main assemblies are present along the temperature coordinate which correspond to the denatured and more native-like conformations (definitely less heterogeneous than denatured). Transient conformations between these two assemblies were characterized at Tt by residue-specific contact studies (Fig. 3) and by an average side chain contact map (Fig. 4b). This analysis and side-chain contact studies presented below show that the folding transition is associated with a specific nucleation event.

Six residues: Y3, L5, F30, W43, Y45, F52 were found to form the folding nucleus during the MC dynamics simulations described here. The nucleus geometry is shown in Fig. 5 (bottom). The composition of the nucleus is in agreement with the all-atom MC simulations using phi-values as restraints performed by Shakhnovich and coworkers[9]. Moreover, the same set of residues was identified by these authors via an inspection of residue conservation in proteins that share fold but not sequence similarity to protein G. The analysis of conservation

has shown that all the six nucleus residues are among the ten most frequently conserved in the protein G-like folds. The remaining four of the ten most conserved ones are T18, A23, A26 and V54. Interestingly, in their secondary structure subunits A26 and V54 are second most frequently long-range interacting residues (at Tt): A26 after F30 in the helix, and V54 after F52 in the second hairpin (Fig. 3b,d). As the authors noted, the nucleus residues are conserved with a high statistical significance in respect to the rest of the sequence, which is consistent with similar observations for other fold families[25]. These findings support the idea that the transition state structures and folding mechanisms are determined by the fold topology of native proteins[26]. There is a very valuable confirmation of this idea in folding/denaturation studies of ubiquitin[27] which identified a quite similar folding nucleus to that observed by us in GB1. Both proteins are sequentially completely different, although they share the same overall fold, and therefore it is very likely that their folding mechanisms are similar.

Folding mechanism

Native contact clusters analysis (in the long-range interactions terms, Table 1) reveals a well-defined sequence of folding events. The folding process can be described as a successive assembly of the elements of the supersecondary structure, with the key residues attaching sequentially to the folding nucleus:

(i) The first folding event is the formation of the second β-hairpin (strongly stabilized by three hydrophobic residues: W43, Y45, F52).

(ii) With the decrease of T, contacts between the α-helix (F30) and the second β-hairpin (W43, Y45, F52) are increasingly stronger, and finally at the Tt they become more persistent than the key contacts within the second hairpin.

(iii) The next folding event is the nucleation of the β-sheet residues between β1 and β4 strands beginning from L5 (β1) and F52 (β4), at the beginning assisted by W43 (β3).

(iv) The involvement of the last nucleus residue in the nucleation process, Y3, results in the formation of the central part of the β-sheet (β1-β4) and the correct fold topology. A fluctuating native-like globule is formed.

Temperature dependence of the native contact clustering and the sequential acquisition of the native-like secondary structure (Fig. 2) are consistent with protein engineering studies

of the β-sheet formation. Mutations in β1, β3, β4 strands have intermediate phi values, suggesting partial formation of a three stranded β-sheet composed of these strands in the folding transition state ensemble, while the first hairpin and the helix have relatively low phi values[19]. A similar picture of the transition state can also be inferred from the observed subsets of long-range (i-ii≥5) native-like side-chain contacts seen in the transition structures at Tt (Fig. 3d). Strands β1, β3 and β4 are the most native-like. The internal ordering of the helix is stabilized mainly by local (short-range) interactions and therefore single mutations within the helix shouldn't have much impact on the whole GB1 folding mechanism.

Highly denatured state exhibits native-like long-range interactions

At T=2.1 GB1 is highly unfolded with the average radius of gyration, R_g= 16.4 Å and the chain expands occasionally to an R_g range of 35 Å (for R_g distribution at different temperatures see Fig. 1d). At this highly swollen state the protein begins to form first native-like, long-range interactions (for the side chain contact map see Fig. 4a). The average size at T=2.1 is smaller than the GB1 random coil size (23 Å)[28], mostly due to the partial formation of the helix. As can be seen in Fig. 4a, the center of the helix is the most persistent fragment of the residual structure, present in a more than half of snapshots. The rest of the structure, especially in respect to the non-local interactions, remains highly disordered. Remarkably, the most frequently occurring long-range side chain contacts and their fluctuating clusters involve native contacts between the residues found to participate in the folding nucleus (see Table 1, T=2.1). The most frequent non-native long-range (i-ii≥5) contacts are also observed for the residues participating in the nucleus: Y33-W43 (present in 25% of snapshots), T25-F30 (17%), Y45-K50 (16%), Y45-T51 (15%), Y3-W43 (15%). The frequent contacts of Y33 may indicate an important role of this residue in the tight packing of a slightly deformed helix against the sheet. This issue has also been discussed by others[23,29].

Recently, there has been an increasing number of reports suggesting a significant presence of the secondary structure and hydrophobic clustering, even under highly denaturing conditions[30,31]. Moreover, it was found that a highly denatured protein can exhibit long-range ordering loosely resembling the native-like topology[32]. Therefore, the folding process can be directed from the very beginning and the search for the conformational space could be more efficient when not starting from an accidental structure. Current assumptions are that the understanding of the folding process may be possible after more complete structural studies of the denatured state.

Clearly, from the very beginning of the folding process, the residual structure is initiated by hydrophobic interactions. The major role of the hydrophobic interactions in determining the folding route was also found in a reduced modeling study of protein G whereby the hydrophobic interactions excised from physical energy terms critically affected the folding route[5]. The MC simulations presented here confirm the crucial role of the hydrophobic interactions in the initiation and propagation of protein folding[33].

Structural characteristics of the native-like globule

The formation of the folding nucleus is followed by the assembly of a native-like loosely packed globular structure with the average cRMSD above 5 Å from the native at the Tt (see Fig. 1c). To investigate structural properties of this globular state, we extracted a large number of snapshots from the isothermal MC trajectory at Tt of 100,000 structures. Over 14,000 structures (the structures from the low-energy basin having CABS energy values between -240 and -320 and cRMSD from native between 4 and 8 Å) were extracted and clustered using a single-link clustering algorithm[34]. In the single-link clustering, the distance between the clusters is defined as the distance between their closest members. With the cut-off distance of 2.2 Å, the largest cluster (representing the lowest free-energy basin) consists of 2557 structures. The structures in this cluster have the correct native topology, although with highly variable structural details. The remaining clusters are by two orders of magnitude smaller and represent incorrect folds (incorrect arrangements of the stands in the β sheet or the C-terminal and the N-terminal strands shifted in the register by 2 residues). The schematic drawing of the centroid of the largest cluster and the comparison of the sizes of the top-ten clusters is given in Fig. 6. Clearly, relatively well-defined native-like structures dominate the simulation trajectory.

Is there an intermediate state?

GB1 was initially thought to fold by a two-state process, like many other small proteins[35]. However, continuous-flow fluorescence measurements of GB1 kinetics demonstrated clear deviations from the kinetics expected for a simple two-state process, and indicated presence of an intermediate[36]. The time course of the refolding revealed a prominent exponential phase with a time constant of 600-700 μs followed by a second, rate-limiting process with a time constant of 2 ms or longer, depending on denaturant concentration. According to Park et al.[36] the biphasic kinetics of the folding can be modeled quantitatively on the basis of a three state folding mechanism (folding through an intermediate). An

ensemble of intermediate states represents native-like fluorescence properties: W43 becomes largely buried during the initial phase of folding. Additionally, the denaturant-dependent rate constant studies provided insight into the solvent-accessible surface area associated with each transition[36]. According to this analysis, the initial barrier, the first transition state TS1, represents a well solvated ensemble of states with α=0.29 (the α values indicate a change in the solvent accessible surface area relative to the unfolded state, for the unfolded state α=0, for the native α=1), while both the intermediate and the second transition state, TS2, are nearly as desolvated as the native state (α=0.85). The high α value for the intermediate state, $\alpha_{Intermediate}$ = 0.85, implies that it represents a compact set of conformations with the solvent exposed surface area only slightly larger than that of the native state, which is consistent with its native-like fluorescence properties. A comparison with other proteins for which $\alpha_{Intermediate}$ values have been determined on the basis of stopped-flow data shows that the intermediate in the GB1 folding is unusually compact[37]. According to Park et al. this is probably a consequence of the fact that the hydrophobic core of GB1 is relatively large for a protein of its size.

The plots shown in Fig. 1 suggest a two-state behavior, although a closer inspection of the structures at the temperatures just below the region of the steepest changes of structural properties clearly indicates that this cooperative chain collapse leads to the molten globule state (see also Fig. 2), consistent with that described by Park et al.[36]

The molten globule is known as a stable, collapsed state with a partial native-like ordering that proteins can adopt under certain conditions[38,39]. The molten globule possesses a somewhat native secondary and tertiary structure, although with a high mobility of a little exposed side chains. Interestingly, in our earlier simulation studies (unpublished) of small single domain proteins significantly more compact and closer to the native structures were usually observed. This is another indication of the three-state folding of GB1. In some simulations of GB1 we observed that long relaxation of the molten globule-like structures, after the initial fast collapse to the proper topology, leads to more closely packed structures, with much smaller overall fluctuations near the native state. Due to the coarse-grained character of the CABS model these compact structures exhibit only partially native-like packing (the best structures were around 2.5 Å from the native). The backbone geometry and the main-chain hydrogen bond network were native-like, although only a fraction of the side chains became fully fixed (compare the work by Hubner, Shimada et al.[9]).

Interestingly, the presence of an intermediate is still under debate. Krantz et al.[40] questioned the validity of Roder and coworkers' analysis[36], suggesting a folding without accumulation of an intermediate, which was later refuted by Roder and coworkers[41].

Our GB1 simulations support the folding scenario starting from the formation of a loosely packed ensemble of relatively compact states with a native-like overall fold, followed by a rate limiting formation of the unique native structure with its tightly packed core. Whether the two folding stages can be observed experimentally depends on the stability of the compact intermediates[36]. Such accumulation of compact states with the native-like features during the GB1 folding was also observed in MD simulations[42].

Distinct mechanisms of GB1 and CI2 folding

Very interesting is the comparison of the GB1 folding mechanism with the folding mechanism of another small protein, CI2, being also a paradigm for kinetics studies. Results of CABS modeling of the CI2 folding pathway have been published recently. We observed that GB1 and CI2 fold by somewhat different mechanisms. This is in slight disagreement with the interpretation of the experimental data suggested by Daggett and Fersht[43]. According to them CI2 folds via nucleation-collapse around an extended nucleus, similarly to what has been observed in the present work for GB1. Indeed in the case of GB1 all nucleus residues take part in the nucleation event at very early stages of folding. On the contrary, CI2 folds via the assembly of distinct cooperative subunits[13]. At the folding transition the only native tertiary interactions are observed between the two central strands, β3β4. Consolidation of the helix and β3β4 takes place at lower temperatures. Interestingly, when looking at the CABS energy (or radius of gyration) as a function of temperature during the CI2 folding the stepwise formation of cooperative subunits doesn't affect the exponential thermal characteristics of these observables. It appears that the differences observed in simulations in the folding pathways of GB1 and CI2 are actually in agreement with the available experimental data; just the mechanistic picture the simulations provide is easier for comprehensive analysis and interpretation.

Conclusions

The theoretical model of protein G folding resulting from the present computer studies is consistent with the experimental observations[19,36] as well as with previous theoretical studies performed by the all-atom Go model, revealing the existence of a well-defined folding nucleus[4,9] and providing a comprehensive picture of the folding mechanism.

The present approach may be a very useful tool for qualitative studies of entire folding pathways of large proteins and macromolecular assemblies. Since a procedure exists for a fast and accurate protein chain rebuilding[44] with subsequent all-atom refinement and model assessment[45], the proposed method should be applicable to detailed computational studies of long-time dynamics of biomolecular systems. An example of such multiscale modeling is schematically illustrated in Fig. 5, where the rebuilding of the atomic details enabled precise identification of the interaction of the side chains forming the folding nucleus.

The physical folding mechanism observed in the CABS simulations strongly suggests that the interactions in the denatured state are very similar to those in the native structures. Consequently, the knowledge-based potentials derived from native structures are a good approximation of the interactions in the denatured state.

Simulations described here provide a detailed insight into the folding mechanism at the level of individual residues. The results consistent with the experimental and theoretical findings prove that the proposed MC dynamics and a sampling scheme mimic qualitative features of the continuous long-time protein dynamics. This opens up a possibility of efficient multiscale computational studies of protein dynamics, folding mechanisms and protein docking mechanisms.

Methods

CABS model description

The CABS protein representation and the model force field have been described in detail recently[10]. In the reduced representation of the CABS protein chains each residue is represented by four united groups: Cα, Cβ, the center of mass of the side group and the center of the peptide bond. Positions of the Cα atoms are restricted to a simple cubic lattice with the lattice grid equal to 0.61 Å. A large number (800) of possible orientations of the virtual Cα - Cα bonds ensure lack of lattice anisotropy effects. On the other hand, the lattice representation facilitates very fast computation of interactions and local conformational transitions. The alpha carbon trace provides a convenient reference frame for the definition of the position of

the remaining interaction centers which are located off-lattice. In the studies of protein dynamics the simulation process is controlled by the asymmetric Metropolis MC scheme with a long random sequence of local conformational updates. A single step of the MC algorithm consists of several attempts to various local conformational transitions for each residue of the model chain. The sequence of the attempts at particular transitions is generated in a random fashion. The MC process simulates a long-time stochastic dynamics of a polypeptide chain.

The force field consists of several potentials of the mean force derived from a statistical analysis of the structural correlations seen in the known protein structures. The short-range interactions include generic protein-like conformational biases and statistical potentials describing local conformational propensities. A model of the main chain network of hydrogen bonds controls mutual packing of β-strands and a proper cooperative assembly of helices. The sequence and geometric context-dependent statistical potentials describe side group interactions with a cooperative (multibody) component built-in in an implicit fashion. The CABS model has been extensively tested in numerous applications, including protein structure prediction, protein docking and studies of long time dynamics and thermodynamics of proteins and protein assemblies.

CABS simulations and data analysis

During the simulations analyzed in this work multiple 10,000,000 MC steps isothermal trajectories were collected at different temperatures. The BioShell package[46] for protein modeling computation was used for managing and analyzing the large volume of simulation data.

The side chain contact patterns were derived from the distributions of the distances between the centers of gravity of the side chains. The values of the contact cut-off distances depend on the identity and mutual orientation of the amino acids involved: two amino acids were assumed to be "in contact" when the distance between any pair of their heavy atoms was smaller than 4.5 Å.

The estimations of the density of states (see Fig. 1a,b,d and Fig. 2) were computed using T-pile[47] from 10,000,000 MC steps' isothermal trajectories. Simulations were performed independently twice at all temperatures from 1.90 to 2.10 with 0.01 increment. Each simulation produced 100,000 structures, and required approximately 2 days using a standard machine (3 GHz CPU, Linux box).

The software used for the analysis of the simulation data as well as the executable of CABS can be downloaded from our website (http://www.biocomp.chem.uw.edu.pl/). The

movie illustrating the evolution of the density of states (and other system properties) with changing temperature can also be viewed at our homepage (http://www.biocomp.chem.uw.edu.pl/movies.php).

Multiscale modeling

The multiscale modeling procedure (for the data presented in Fig. 5) consisted of the following steps. First, the missing backbone atoms were reconstructed using the BBQ algorithm[44]. Subsequently, the side-chain rotamers were added using SCWRL[48]. The resulting all-atom models were subjected to a short refinement procedure. We run the all-atom minimization with frozen alpha carbons using Amber7 ff99 force field, Amber charges, dielectric constant equal to 1.0, and Powell minimization method implemented in Sybyl (Tripos Inc. St. Louis, MO), without initial optimization. 1000 iterations of the refinement procedure were done to improve arrangement of the side chain rotamers. Recently a very similar procedure has been effectively used in a hierarchical approach to the model refinement and final structure selection after a coarse-grained modeling with CABS[45]. Such multiscale methodology leads to a very good approximation of a model's energy (and free energy). Thus it may be very useful in comprehensive studies of protein folding energetics, which is now being pursued in our laboratory.

Acknowledgements

The authors thank Dr. Dominik Gront for the calculations of the density of states[47], which enabled fine data visualization (used in Fig. 1, 2) and valuable assistance in the contact clusters statistics.

Some simulations described here were conducted at the Computer Center of the Faculty of Chemistry of the University of Warsaw.

References

1. Duan, Y. & Kollman, P.A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740-4 (1998).
2. Vila, J.A., Ripoll, D.R. & Scheraga, H.A. Atomically detailed folding simulation of the B domain of staphylococcal protein A from random structures. *Proc Natl Acad Sci U S A* **100**, 14812-6 (2003).
3. Prieto, L., de Sancho, D. & Rey, A. Thermodynamics of Go-type models for protein folding. *J Chem Phys* **123**, 154903 (2005).
4. Shimada, J. & Shakhnovich, E.I. The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation. *Proc Natl Acad Sci U S A* **99**, 11175-80 (2002).
5. Lee, S.Y., Fujitsuka, Y., Kim, D.H. & Takada, S. Roles of physical interactions in determining protein-folding mechanisms: molecular simulation of protein G and alpha spectrin SH3. *Proteins* **55**, 128-38 (2004).
6. Blanco, F.J., Ortiz, A.R. & Serrano, L. Role of a nonnative interaction in the folding of the protein G B1 domain as inferred from the conformational analysis of the alpha-helix fragment. *Fold Des* **2**, 123-33 (1997).
7. Liwo, A., Khalili, M. & Scheraga, H.A. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc Natl Acad Sci U S A* **102**, 2362-7 (2005).
8. Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253**, 694-8 (1975).
9. Hubner, I.A., Shimada, J. & Shakhnovich, E.I. Commitment and nucleation in the protein G transition state. *J Mol Biol* **336**, 745-61 (2004).
10. Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim Pol* **51**, 349-71 (2004).
11. Kolinski, A. & Bujnicki, J.M. Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins* **61**, 84-90 (2005).
12. Bradley, P. et al. Free modeling with Rosetta in CASP6. *Proteins* **61 Suppl 7**, 128-34 (2005).
13. Kmiecik, S. & Kolinski, A. The Characterization of Protein Folding Pathways by Reduced-space Modeling. *in the press*.
14. Gronenborn, A.M. et al. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253**, 657-61 (1991).
15. Blanco, F.J., Rivas, G. & Serrano, L. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat Struct Biol* **1**, 584-90 (1994).
16. Kuszewski, J., Clore, G.M. & Gronenborn, A.M. Fast folding of a prototypic polypeptide: the immunoglobulin binding domain of streptococcal protein G. *Protein Sci* **3**, 1945-52 (1994).
17. Munoz, V., Thompson, P.A., Hofrichter, J. & Eaton, W.A. Folding dynamics and mechanism of beta-hairpin formation. *Nature* **390**, 196-9 (1997).
18. Kolinski, A., Ilkowski, B. & Skolnick, J. Dynamics and thermodynamics of beta-hairpin assembly: insights from various simulation techniques. *Biophys J* **77**, 2942-52 (1999).
19. McCallister, E.L., Alm, E. & Baker, D. Critical role of beta-hairpin formation in protein G folding. *Nat Struct Biol* **7**, 669-73 (2000).
20. Derreumaux, P. Role of supersecondary structural elements in protein G folding. *Journal of Chemical Physics* **119**, 4940-4944 (2003).
21. Frank, M.K., Clore, G.M. & Gronenborn, A.M. Structural and dynamic characterization of the urea denatured state of the immunoglobulin binding domain of

streptococcal protein G by multidimensional heteronuclear NMR spectroscopy. *Protein Sci* **4**, 2605-15 (1995).
22. Brooks, C.L. Protein and peptide folding explored with molecular simulations. *Accounts of Chemical Research* **35**, 447-454 (2002).
23. Sheinerman, F.B. & Brooks, C.L., 3rd. A molecular dynamics simulation study of segment B1 of protein G. *Proteins* **29**, 193-202 (1997).
24. Kolinski, A., Klein, P., Romiszowski, P. & Skolnick, J. Unfolding of globular proteins: monte carlo dynamics of a realistic reduced model. *Biophys J* **85**, 3271-8 (2003).
25. Mirny, L. & Shakhnovich, E. Evolutionary conservation of the folding nucleus. *J Mol Biol* **308**, 123-9 (2001).
26. Alm, E. & Baker, D. Matching theory and experiment in protein folding. *Curr Opin Struct Biol* **9**, 189-96 (1999).
27. Babu, C.R., Hilser, V.J. & Wand, A.J. Direct access to the cooperative substructure of proteins and the protein ensemble via cold denaturation. *Nat Struct Mol Biol* **11**, 352-7 (2004).
28. Smith, C.K. et al. Surface point mutations that significantly alter the structure and stability of a protein's denatured state. *Protein Sci* **5**, 2009-19 (1996).
29. Clore, G.M. & Gronenborn, A.M. Localization of bound water in the solution structure of the immunoglobulin binding domain of streptococcal protein G. Evidence for solvent-induced helical distortion in solution. *J Mol Biol* **223**, 853-6 (1992).
30. Klein-Seetharaman, J. et al. Long-range interactions within a nonnative protein. *Science* **295**, 1719-22 (2002).
31. Kazmirski, S.L. et al. Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc Natl Acad Sci U S A* **98**, 4349-54 (2001).
32. Shortle, D. & Ackerman, M.S. Persistence of native-like topology in a denatured protein in 8 M urea. *Science* **293**, 487-9 (2001).
33. Dyson, H.J., Wright, P.E. & Scheraga, H.A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc Natl Acad Sci U S A* **103**, 13057-61 (2006).
34. Gront, D. & Kolinski, A. HCPM--program for hierarchical clustering of protein models. *Bioinformatics* **21**, 3179-80 (2005).
35. Jackson, S.E. How do small single-domain proteins fold? *Fold Des* **3**, R81-91 (1998).
36. Park, S.H., Shastry, M.C. & Roder, H. Folding dynamics of the B1 domain of protein G explored by ultrarapid mixing. *Nat Struct Biol* **6**, 943-7 (1999).
37. Roder, H. & Colon, W. Kinetic role of early intermediates in protein folding. *Curr Opin Struct Biol* **7**, 15-28 (1997).
38. Ptitsyn, O.B. Molten globule and protein folding. *Adv Protein Chem* **47**, 83-229 (1995).
39. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins* **6**, 87-103 (1989).
40. Krantz, B.A., Mayne, L., Rumbley, J., Englander, S.W. & Sosnick, T.R. Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J Mol Biol* **324**, 359-71 (2002).
41. Roder, H., Maki, K. & Cheng, H. Early events in protein folding explored by rapid mixing methods. *Chem Rev* **106**, 1836-61 (2006).
42. Sheinerman, F.B. & Brooks, C.L., 3rd. Calculations on folding of segment B1 of streptococcal protein G. *J Mol Biol* **278**, 439-56 (1998).

43. Daggett, V. & Fersht, A. The present view of the mechanism of protein folding. *Nat Rev Mol Cell Biol* **4**, 497-502 (2003).
44. Gront, D., Kmiecik, S. & Kolinski, A. Backbone building from quadrilaterals: A fast and accurate algorithm for protein backbone reconstruction from alpha carbon coordinates. *J Comput Chem* **28**, 1593-7 (2007).
45. Kmiecik, S., Gront, D. & Kolinski, A. Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field. *submitted* (2007).
46. Gront, D. & Kolinski, A. BioShell--a package of tools for structural biology computations. *Bioinformatics* **22**, 621-2 (2006).
47. Gront, D. & Kolinski, A. T-Pile - a package for thermodynamic calculations for biomolecules. *Bioinformatics*, Advance Access published on May 12, 2007, doi:10.1093/bioinformatics/btm256 (2007).
48. Canutescu, A.A., Shelenkov, A.A. & Dunbrack, R.L., Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* **12**, 2001-14 (2003).
49. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-637 (1983).
50. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-8, 27-8 (1996).

Figure captions

Table 1. Temperature-dependent evolution of the most persistent native long-range ([i-ii]≥5) side-chain contacts and their clusters. For each temperature five most frequently appearing native contacts (1), doublets (2), triplets (3) and quadruplets of native contacts (4) are shown. Their frequencies, expressed as the percentages of the snapshots in corresponding trajectories, are also presented. At T=1.96, near the folding Tt, the sixth contact (in the order of the frequency of occurrence) is additionally shown. Data were collected from 10,000,000 MC steps' isothermal trajectories.

T	1	%1	2	%2	3	%3	4	%4
2.10	45-52	25.7	43-52 45-52	15.0	43-52 43-54 45-52	6.9	30-43 30-52 43-52 45-52	3.1
	43-52	25.0	43-52 43-54	9.6	30-43 30-52 43-52	4.6	5-43 5-52 43-52 45-52	2.2
	30-43	20.7	43-54 45-52	9.2	30-43 43-52 45-52	4.5	30-43 43-52 43-52 43-54	2.1
	5-43	17.8	30-43 45-52	6.9	30-52 43-52 45-52	4.3	30-52 43-52 43-54 54-52	2.0
	30-52	14.3	30-52 43-52	6.9	30-43 30-52 45-52	4.0	30-43 30-52 43-52 43-54	1.8
2.04	45-52	32.0	43-52 45-52	19.8	43-52 43-54 45-52	9.9	30-43 30-52 43-52 45-52	5.8
	43-52	31.7	43-52 45-52	12.9	30-43 30-52 43-52	8.6	5-43 5-52 43-52 45-52	4.3
	30-43	28.3	43-54 45-52	12.7	30-43 43-52 45-52	8.0	30-43 43-52 43-54 45-52	4.1
	5-43	26.5	30-43 45-52	12.4	30-52 43-52 45-52	7.8	30-52 43-52 43-54 45-52	4.0
	30-52	22.4	30-52 43-52	11.9	30-43 30-52 45-52	7.3	26-52 30-43 30-52 43-52	3.6
2.01	43-52	38.5	43-52 45-52	24.4	43-52 43-54 45-52	13.3	30-43 30-52 43-52 45-52	8.4
	45-52	37.2	30-52 43-52	17.7	30-43 30-52 43-52	12.4	30-43 43-52 43-54 45-52	6.5
	30-43	33.2	43-54 45-52	17.1	30-43 43-52 45-52	11.4	3-52 5-52 5-54 7-54	6.4
	5-43	30.6	30-43 43-52	16.6	30-52 43-52 45-52	11.2	30-52 43-52 43-54 45-52	6.3
	30-52	29.4	43-52 43-54	16.3	30-43 30-52 45-52	10.4	5-43 5-52 43-52 45-52	6.1
1.98	43-52	46.9	43-52 45-52	28.9	30-43 30-52 43-52	19.3	30-43 30-52 43-52 45-52	12.2
	45-52	42.1	30-52 43-52	26.5	43-52 43-54 45-52	16.7	3-52 5-52 5-54 7-54	11.9
	30-43	41.3	30-43 43-52	24.8	30-43 43-52 45-52	16.3	5-52 30-43 30-52 43-52	11.3
	30-52	39.8	30-43 30-52	23.3	5-52 30-52 43-52	15.7	4-51 5-52 6-53 7-54	11.0
	5-43	37.2	30-43 45-52	21.7	30-52 43-52 45-52	15.6	3-52 5-52 7-54 30-52	10.7
1.96	43-52	52.4	30-52 43-52	32.6	30-43 30-52 43-52	20.8	3-52 5-52 5-54 7-54	15.8
	30-52	47.2	43-52 45-52	32.0	3-52 5-52 30-52	18.3	4-51 5-52 6-53 7-54	14.7
	30-43	46.4	30-43 43-52	30.6	3-52 5-52 5-54	18.2	3-52 5-52 7-54 30-52	14.4
	45-52	45.0	30-43 30-52	28.9	5-52 30-52 43-52	18.1	3-52 5-52 5-54 30-52	14.2
	5-52	44.0	30-43 45-52	25.6	3-52 5-52 7-54	18.0	3-52 5-52 6-53 7-54	14.0
	3-52	39.9						

Figure 1. Estimation of the density of states for the various observables: (a) CABS energy as a function of reduced temperature. (b) cRMSD as a function of temperature. (c) cRMSD as a function of CABS energy at the transition temperature (Tt) (T=1.955). (d) Radius of gyration (in Ångstroms) as a function of temperature. Color indicates density of states in arbitrary units, depending on the kinds of data presented in particular plots.

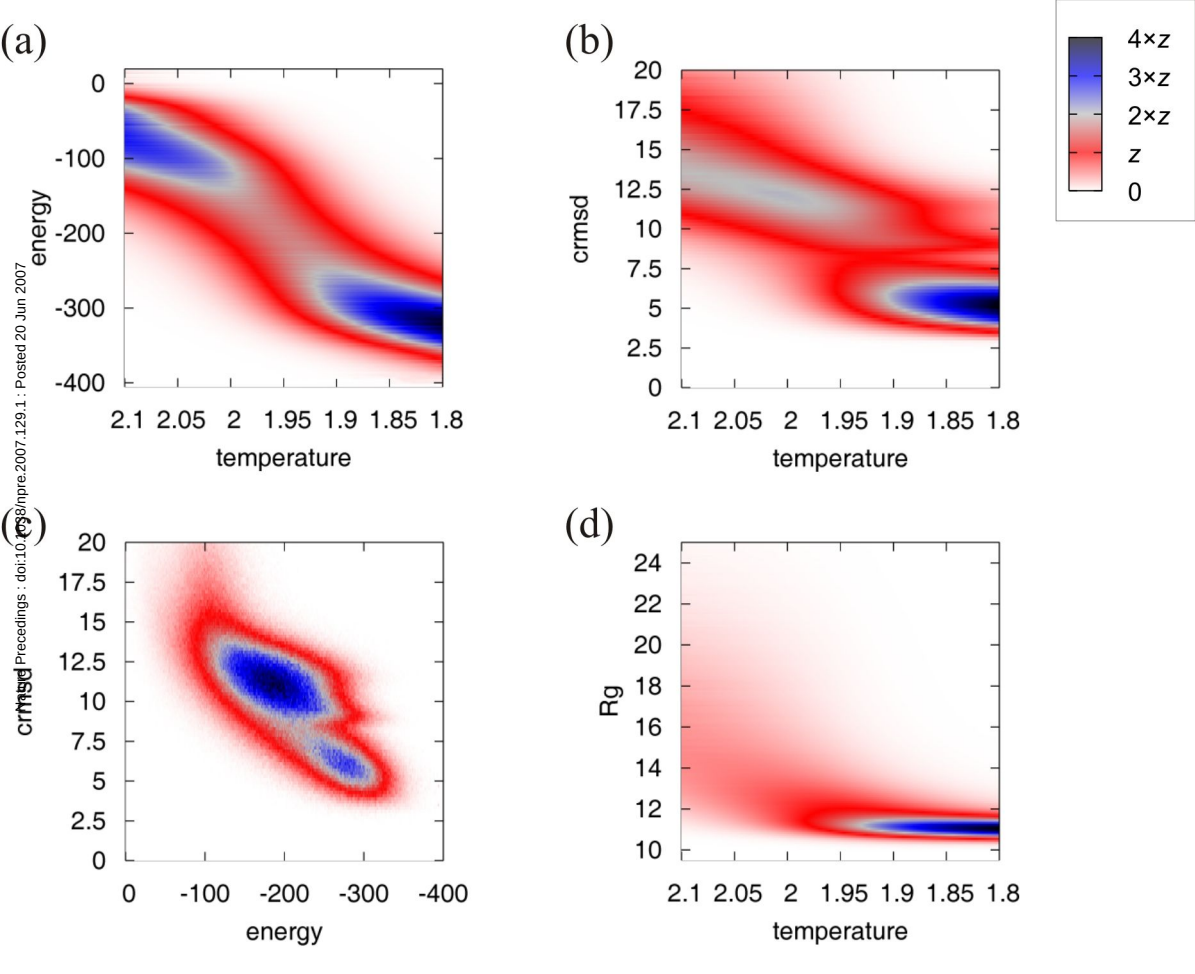
Figure 2. Temperature dependence of the content of native secondary structure elements measured in the simulations. Each panel represents temperature dependence of the number of native contacts within the particular secondary structure elements (depicted on the right side of the figure). Beginning from the top the data for the first hairpin, the second hairpin, the helix, and for the pair of the terminal strands are illustrated, respectively. The Tt is marked by a vertical line throughout all panels. The color scale denotes probability of the particular number of the contacts (the scale is given to the right of the drawing).

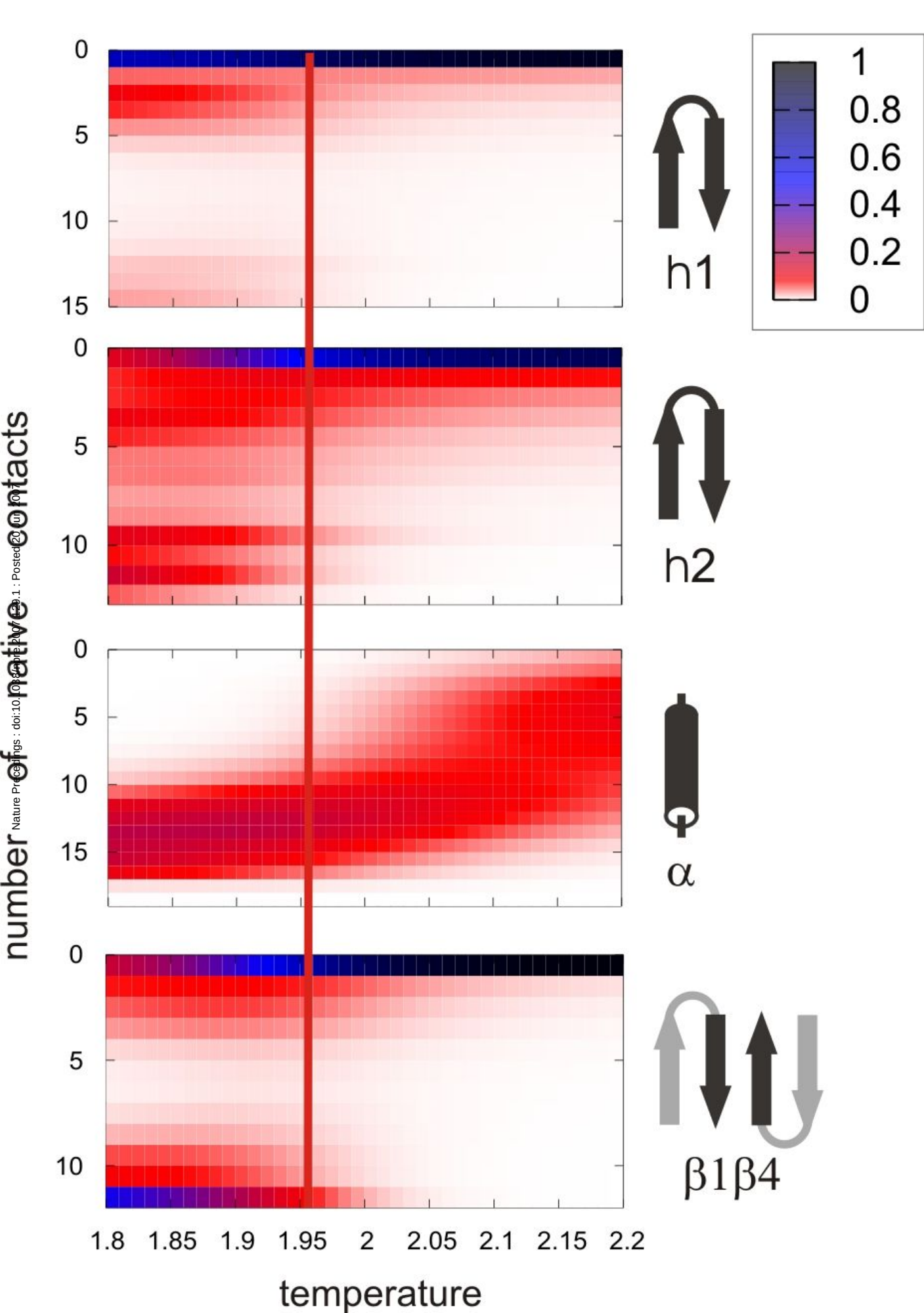
Figure 3. Acquisition of the side-chain contacts at the Tt. Average number of native (a) and all (c) side chain contacts formed by a single residue is shown. The numbers of contacts are calculated for all the structures from the trajectory at the Tt (red color) and separately for two classes of structures extracted from the trajectory: denatured structures located characterized by high cRMSD and high energy values (gray color) and transition structures (placed between the basin of native-like states and broad basin of denatured conformations, see also Fig. 1c) characterized by medium cRMSD and energy values (black color). For the transition structures the distributions of the average number of contacts (b) and the fraction of the native contacts (d) are illustrated in terms of the distances along the sequence. All i-ii≥3 interactions (gray color) and only non-local i-ii≥5 interactions (black color) are displayed in both panels.

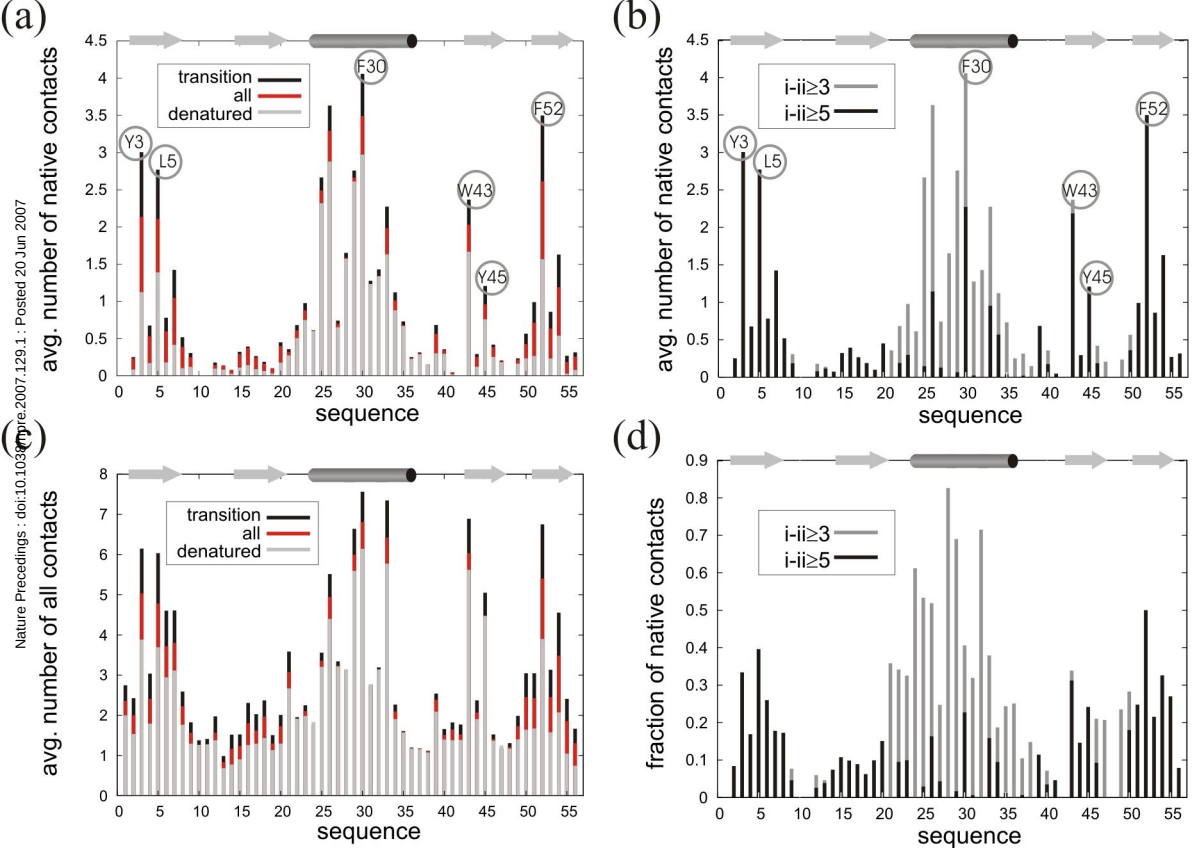
Figure 4. Average contact maps (above the diagonals) compared with the native contact maps (below the diagonals) for the denatured state (a) and the transition structures (b). Colors indicate frequency of contacts. The average contact map for the denatured state was calculated from the entire isothermal trajectory at T=2.1. The average contact map for the transition structures is the contact map from isothermal simulation at the Tt for the structures with the transient energy and cRMSD values (the same defined in the Fig. 3 caption). The most persistent contacts between the nucleus residues are circled.

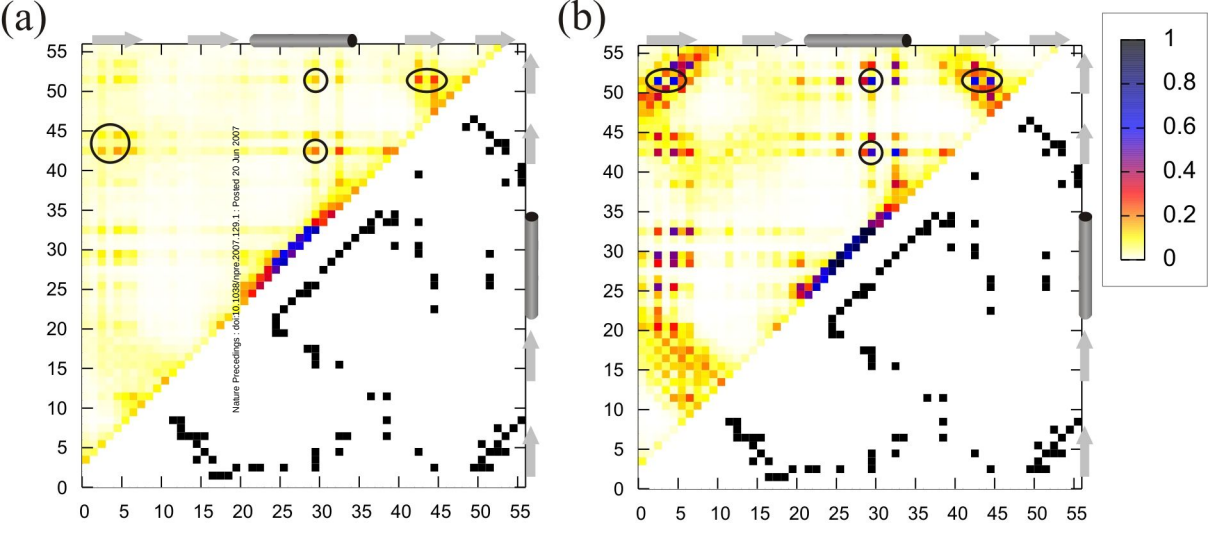
Figure 5. Multiscale modeling: example snapshots from a simulation presenting interactions of the nucleus side-chains at various folding stages. Heavy atom bonds of the nucleus residues are marked with colored sticks; the secondary structure is depicted using transparent ribbons. Models were reconstructed from the reduced representation and refined by the procedure described in the Methods section. Afterward, the secondary structure was assigned by the DSSP algorithm.[49] At the bottom, the experimental structure of protein G is drawn for comparison with the nucleus residues labeled by their position in the sequence. Pictures of models were drawn using PyMOL (DeLano, W. L., The PyMOL Molecular Graphics System, http://www.pymol.org).

Figure 6. Classification of the structurally related low-energy conformations at Tt. The left-side plot (a) displays in the logarithmic scale the relative sizes (number of members of a cluster) of structural classes obtained from the clustering of all the structures from the low-energy basin at Tt (structures having CABS energy value between -240 and -320 and cRMSD from native between 4 and 8 Å). The right side (b) shows the structure of the centroid of the largest cluster (in grey), exhibiting loosely defined conformation (albeit in the correct topology), 5 Å from the native structure, superimposed on the native structure (blue). The image was created using VMD.[50]








(a)
(b)
1
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35 40 45 50 55
Nature Precedings : doi:10.1038/npre.2007.125.1 : Posted 20 Jun 2007