Acta Biochimica Polonica

Vol. 53 No. 1/2006, 131–143

on-line at: www.actabp.pl

*Regular paper*

# Denatured proteins and early folding intermediates simulated in a reduced conformational space

Sebastian Kmiecik, Mateusz Kurcinski, Aleksandra Rutkowska, Dominik Gront and Andrzej Kolinski✉

*Faculty of Chemistry, Warsaw University, Warszawa, Poland; ✉e-mail: kolinski@chem.uw.edu.pl*

Received: 18 August, 2005; revised: 05 October, 2005; accepted: 06 November, 2005
available on-line: 19 December, 2005

**Conformations of globular proteins in the denatured state were studied using a high-resolution lattice model of proteins and Monte Carlo dynamics. The model assumes a united-atom and high-coordination lattice representation of the polypeptide conformational space. The force field of the model mimics the short-range protein-like conformational stiffness, hydrophobic interactions of the side chains and the main-chain hydrogen bonds. Two types of approximations for the short-range interactions were compared: simple statistical potentials and knowledge-based protein-specific potentials derived from the sequence–structure compatibility of short fragments of protein chains. Model proteins in the denatured state are relatively compact, although the majority of the sampled conformations are globally different from the native fold. At the same time short protein fragments are mostly native-like. Thus, the denatured state of the model proteins has several features of the molten globule state observed experimentally. Statistical potentials induce native-like conformational propensities in the denatured state, especially for the fragments located in the core of folded proteins. Knowledge-based protein-specific potentials increase only slightly the level of similarity to the native conformations, in spite of their qualitatively higher specificity in the native structures. For a few cases, where fairly accurate experimental data exist, the simulation results are in semiquantitative agreement with the physical picture revealed by the experiments. This shows that the model studied in this work could be used efficiently in computational studies of protein dynamics in the denatured state, and consequently for studies of protein folding pathways, i.e. not only for the modeling of folded structures, as it was shown in previous studies. The results of the present studies also provide a new insight into the explanation of the Levinthal's paradox.**

**Keywords**: protein folding, high resolution lattice protein models, statistical potentials, sequence profiles, Replica Exchange Monte Carlo, molten globule, protein folding intermediates

Protein structure determination is nowadays probably the most important task of computational biology (Baker *et al.*, 2003). Knowledge of protein's three-dimensional structure is extremely helpful (sometimes even critical) in issues such as protein function understanding (Bartlett *et al.*, 2003), recognition of disease mechanisms, or rational drug design (Skolnick *et al.*, 2000). We now know about 30 thousand of high-resolution protein structures, which is only about 0.1% of the number of known protein sequences (Whisstock & Lesk, 2003). This gap is rapidly spreading due to the huge difference between the costs of sequencing and experimental structure determination. Whereas protein sequencing is relatively cheap and performed by machines, experimental structure solving (X-ray diffraction or/and NMR) requires large resources and highly qualified staff. Theoretical structure prediction may

**Abbreviations**: CABS, Cα–Cβ-side group protein model; CASP, Critical Assessment of Techniques for Protein Structure Prediction; CD, circular dichroism; cRMSD, coordinate root mean square deviation; MCD, Monte Carlo dynamics; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; REFINER, a continuous space reduced model, similar to the CABS model; REMC, Replica Exchange Monte Carlo; SICHO, Side Chain Only protein model.

be a perfect complement to or even a substitute for the experimental techniques (Contreras-Moreira *et al.*, 2002; Chance *et al.*, 2004). Although the computing power of CPUs increases quite fast (it approximately doubles every 18 months — according to the "Moore's law"[1]), *ab initio* calculations perform well only for very small proteins (Bonneau *et al.*, 2002). The reason of the limited applicability of these methods is the enormous number of degrees of freedom of a molecular system consisting of a protein and the solvent surrounding it. Also the force fields used in such calculations must be precisely designed to cover all (or at least all significant) interactions between all atoms of the system. This makes the calculations extremely time-consuming. At appropriate conditions proteins fold to a unique three-dimensional structure. This process takes from milliseconds to minutes, which is much too long for the present-day computers to do modeling in a classical (molecular dynamics) way (Brooks *et al.*, 1998). At the same time the database of the 30 thousand already solved structures is a good starting point for knowledge-based modeling techniques such as classical comparative modeling or threading (Baker *et al.*, 2003). Programs following those methods build new models, using already solved structures as templates selected from PDB (Berman *et al.*, 2000), mostly by sequence alignment (Sali, 1998). They perform much faster than the molecular mechanics calculations and may be used to model even the biggest proteins. Additional progress in reducing the computation cost may be obtained by using reduced models of proteins. In our lab a few high-resolution reduced models of proteins (SICHO, CABS, REFINER) have been developed and applied to such tasks as comparative modeling and docking (Bujnicki *et al.*, 2001; Rotkiewicz *et al.*, 2001; Sicinski *et al.*, 2002; Bolesta *et al.*, 2005), loop-building, or modeling the process of protein folding (Kolinski *et al.*, 2001; Boniecki *et al.*, 2003; Kolinski, 2004; Kolinski & Skolnick, 2004; Malolepsza *et al.*, 2005; Plewczynska & Kolinski, 2005). Along with a reduced representation of the protein conformational space, specific knowledge-based force fields were also constructed (Gront & Kolinski, 2005; Pokarowski *et al.*, 2005).

In this work we compared the results of protein folding simulations for two sets of potentials (Gront & Kolinski, 2005) applied to the CABS model (Kolinski, 2004), controlled by the REMC sampling scheme (Swendsen & Wang, 1986). Both sets consisted of identical potentials reflecting the sequence-independent and the long-range sequence-dependent interactions. The difference between the two sets of the computational experiments was in the way the short-range sequence-dependent potentials were designed. In the first simulation we used potentials calculated for the whole non-redundant structure database with 35% sequence similarity cut-off (one set for all proteins), while in the second one the corresponding potentials were based on profiles obtained from multiple sequence alignments (Gront & Kolinski, 2005) (one set for each protein).

The main questions addressed in this work are the following. Are the knowledge-based potentials derived from a statistical analysis of structural regularities seen in known protein structures appropriate also for the denatured state? If so, what is the effect of the profile-based enhancement of the short-range sequence-specific potentials? Then, what is the role of the short-range conformational propensities in the denatured state and what are the implications for our understanding of the general physics of protein folding? Finally, we would like to show that the reduced model mimics semiquantitatively not only the average properties of polypeptide chains in denatured conditions but also some fine features of early folding intermediates.

In the first part of this work we address only the question related to the average conformational properties and distributions. Since the Replica Exchange Monte Carlo (Swendsen & Wang, 1986) sampling is used, any interpretations of the system dynamics would be uncertain. On the other hand, REMC is a very efficient algorithm for accurate estimation of average properties (Gront *et al.*, 2000; 2001), and this is exactly why we have chosen this method for the purpose of the present work. For four small globular proteins, for which experimental studies of early folding intermediates have been done, we performed also long Monte Carlo dynamics simulations (a single copy of the modeled system) to extract the average properties of the folding intermediates at the isothermal conditions near the folding temperature and to follow some key intermediates on the folding pathways. Very good agreement with the experimental data was observed.

## MATERIALS AND METHODS

### CABS MODEL

The model has been described recently in great detail (Kolinski, 2004). Here, for the reader's convenience, we provide an outline of its basic features.

[1]In 1965 Gordon Moore, one of the Intel founders, formulated an empirical law stating that every 18 months the number of transistors in CPUs doubles. For almost 40 years this "law" has been obeyed, but recently the progress in computing technology seems to be slowing down, due to the limited potential of silicon technology. That is why *ab initio*, all-atom calculations will not probably play a significant role in large-scale protein modeling, until some new computing technology is invented.

**CABS representation of protein conformational space.** CABS is a high-resolution lattice model. It assumes the representation of every amino acid as four interaction centers — Cα (CA), Cβ (B), the center of mass of the side chain (S) and the center of the peptide bond (pb). Positions of alpha carbons are restricted to the simple cubic lattice with the lattice spacing equal to 0.61 Å. The lengths of the virtual Cα–Cα bonds vary from $29^{1/2}$ to $49^{1/2}$ lattice units, which translates to a 3.27–4.28 Å distance in absolute units, with the mean value equal to 3.78 Å (as in real proteins). There are 800 possible Cα–Cα vectors which ensures that here are no adverse lattice effects such as anisotropy. The beta carbons are located off-lattice and their positions are defined by the position of the alpha carbons (three consecutive alpha carbons determine the location of the central beta carbon). The position of the center of mass of the side chain depends on the local secondary structure (condensed or expanded). In the center of the virtual Cα–Cα bond one additional "atom" is defined for each residue. It takes no part in most interactions, but is helpful in defining the model's hydrogen bonds.

**CABS generic force field**. The force field for the CABS model consists of several elements. Generic terms impose spatial barriers which make the model chain behave like a real protein and reduce the number of the conformational space dimensions. Besides that, the force field contains sequence-dependent potentials, which were constructed using statistical analysis of real protein structures. Here we give only a brief account of the essence of the force field components, with details presented elsewhere (Kolinski, 2004).

The sequence-independent short-range interactions contain three components. The first one controls the planar angle between two neighboring peptide bonds within a range of 70–150°, which reflects real protein-like short-range interactions between the atoms located close to each other in the peptide chain. The second potential is responsible for the secondary structure induced chain stiffness observed in real proteins. The distribution of the end-to-end distance for four-bond chain fragments needs to be bimodal with the first peak corresponding to a compact and the second — to an expanded secondary structure. This potential ensures such a distribution. The third potential controls such properties of secondary structure elements as the distance between respective residues in helices and strands. The last sequence-independent short-range potential prevents the chain from crumpling. It sets the minimal length of fragments between turns to be equal to 5 residues (including the number of residues in a turn), which reflects the protein tendency to avoid creating turn-next-to-turn structures.

Due to the reduced representation, the main-chain hydrogen bonding is included in a specially designed interaction between alpha carbons. There is a set of conditions which have to be fulfilled for such an interaction to occur.

For all Cαs and Cβs a typical hard-core repulsion (infinite energy) was defined with the cut-off distance equal to 3.05 Å for Cα–Cα and Cβ–Cβ, and 3.65 Å for the Cα–Cβ interaction. Besides that, for the Cα–Cα, Cα–pb (pb, center of the peptide bond) and Cα–SG (SG, side chain group) interactions a soft-core potential was defined as a function regularizing the distance distribution within a protein-like geometry.

The sequence-dependent long-range pairwise interactions between the side-groups are context dependent and take into account the identity of interacting groups, their spatial separation, mutual orientation and the geometry of corresponding fragments of the main chain. Thus, complex multibody effects are accounted for in an implicit way.

Short-range interactions are discussed separately in the section below.

**Sequence dependent short-range interactions.** The CABS model distinguishes three types of short-range sequence-specific interactions: between the i-th and i+2nd (R13), between i-th and i+3rd (R14) and between i-th and i+4th (R15) residues. In the present work we used two sets of sequence-dependent short-range interaction potentials. Both were designed using statistical analysis of known protein structures extracted from the PDB database. The first set (statistical potentials) bases on a non-redundant database containing three-dimensional structures of proteins with sequence similarity below 35%. It contains three subsets of potentials for three kinds of interactions (R13, R14 and R15). All the potentials depend on the identity of the two amino acids in the corresponding fragment and on the predicted secondary structure. The R14 term is chiral — it distinguishes between the right-handed and left-handed conformations of three consecutive Cα–Cα virtual bonds. Thus, it can be considered as a pseudo-torsional potential for the Cα-trace.

The second set of potentials (profile-based potentials) was constructed on the basis of a specific database including not only the structures of proteins, but also (Altschul *et al.*, 1997) sequence profiles generated by PSIBLAST. For each protein, in the non-redundant database of protein structures with sequence similarity below 30% (PDB30), a profile was created using PSIBLAST (7 iterations, 1e-7 cutoff e-value for including a hit into a profile). The test set of profile-based potentials was created according to the procedure described in our previous work (Gront & Kolinski, 2005). This set contains 966 subsets — one for each protein. Every subset consists of three potentials for the three kinds of short-range sequence-specific interactions — r13, r14 and

r15 — as described above. Obviously, such potentials are different for each protein and can be easily derived for new proteins.

Both types of potentials have the forms of histograms. The histograms are either specific for a given pair of amino acids (simple statistical potential) or for a given position in the sequence of a given protein. Type R13 potentials contain 8 bins of $\mathbf{r}_{13}$ distance, from 0 Å to 8 Å. R14 potentials have 24 bins, from –12 Å to 12 Å, where the negative sign of the $\mathbf{r}_{14}$ distance corresponds to left-handed conformations, and the positive one to right-handed conformations. R15 potentials contain 16 bins. In all the potentials the bins corresponding to distances from 0 Å to 3 Å are set to an arbitrarily high value, in order to cover the real protein-like distance restrictions. The entire test set of all the described potentials can be viewed and downloaded from our homepage: www.biocomp.chem.uw.edu.pl.

It is worth to mention that the CABS model and its force field are now mature tools for protein modeling. It was tested in several applications, from a simple comparative modeling to *de novo* structure prediction from the amino-acid sequence alone. Recently, the CABS model was used by the Kolinski–Bujnicki group during the 6th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction — CASP6, which was conducted in the summer of 2004. The idea of CASP is to test blind predictions of protein three-dimensional structures before the experimental data become avaiable. The Kolinski–Bujnicki group was scored second best among about 200 groups participating in the experiment. The summary of the CASP results can be found on our homepage http://biocomp.chem.uw.edu.pl or on CASP homepage http://predictioncenter.org/casp6/.

## SIMULATION METHODS

The model systems were sampled *via* the Replica Exchange Monte Carlo method, and *via* the isothermal Monte Carlo dynamics near the folding temperature, with the set of local conformational updates described previously (Kolinski, 2004). In the REMC simulations we used ten replicas and the temperature of the lowest-temperature replica was set above the folding temperature, however low enough to keep the average density of the system close to a density expected for a partly collapsed, molten globule-like state. In the single copy MCD simulations the temperature was set as close as possible to the folding temperature, estimated from preliminary REMC simulations.

Two sets of short-range sequence-dependent potentials (statistical and profile-based) were tested in the framework of the CABS model using the REMC sampling strategy. Twenty high-resolution protein structures from the PDB30 database were selected for the present studies. To each of these proteins the following procedure was applied. In the first step we conducted a short high-temperature Monte Carlo dynamics, using the statistical potentials, in order to relax the structure. Afterwards, the relaxed structure was used in two different Monte Carlo simulations. Both used the CABS representation and identical set of potentials except for the short-range sequence-specific ones. While the first one based on the statistical potentials — identical for all twenty proteins, in the second we used profile-based "local homology" potentials separate for each protein. Both simulations were run with the same parameters (the number of time steps and of replicas, the initial and final temperatures). The collected trajectories were used for further analysis.

Since the final temperatures of the simulations were close to the folding temperatures, we sometimes obtained structures similar to those taken from the PDB database. For each snapshot of a pseudo-trajectory (we use pseudo- since in the REMC the data at a given temperature contain strings of snapshots for different copies of the modeled molecule) the resulting structures and C$\alpha$ chains from PDB files were superimposed in the best possible way (so that the cRMSD over the whole molecule was minimal). To score the details of the modeled structures we calculated cRMSD for each snapshot of the pseudo-trajectory, for fragments of four, six and eight residues, respectively, along the whole chain. The results were presented in the form of graphs along with the predicted secondary structure. This allowed us to estimate which secondary structure elements were modeled with the highest precision.

Additionally, isothermal MCD simulations were performed for four small globular proteins for which properties of the denatured state and early folding intermediates have been well characterized in physical experiments. Here, we analyzed not only the short-range conformational correlations with the native structure but also the frequency of the native side chains' contacts. These were compared with the experimental measurements of the hydrogen-exchange protection factors, which are considered to be very accurate descriptors of the native-like arrangements of protein fragments. Other available experimental data were also compared with the simulation results.

## RESULTS AND DISCUSSION

### Conformations of denatured proteins observed in REMC simulations

The results of the REMC calculations are compared in Table 1. The first two columns provide

the basic data for the 20 proteins of the test systems. Included are proteins (from the PDB30 database) of various lengths (50–150 amino acids) and of various structural classes. The second block of two columns provides the average cRMSD (coordinate root-mean square deviation) data for protein fragments of four amino acids. The third and the fourth blocks provide the results for fragments of six and eight amino acids, respectively. The first column in each block provides the average cRMSD values for models generated using the simple statistical potentials. The second column provides the average cRMSD values for models generated using the protein-dependent homology profile-based potentials. All cRMSD values were calculated for the C$\alpha$ atoms and averaged over the entire trajectory.

It is easy to note that although the average global structures of the modeled proteins are frequently random (high values of the average cRMSD collected in Table 2.), the fragments are not random and resemble the native-like local conformations, although the magnitude of fluctuations increases with the increasing length of the fragment under consideration. On average, the profile-based, protein-specific potentials perform slightly better than the simple statistical potentials. Although the differences for the fragments in the denatured state are small, the difference in the predictive power for the folded state could be significant, as it was demonstrated in the past (Gront & Kolinski, 2005). This is probably the effect of the cooperative character of the folding process — even small local differences can be enhanced on a more global scale.

The simulated structures are relatively compact — their average radii of gyration are by a factor of only about 1.3 larger than the corresponding values for the PDB structures. The data are compiled in Table 3. The fluctuations of the radii of gyration along the trajectories are large. Rarely, values close to those for the native state are observed. Such swollen, highly mobile structures with loosely defined local native-like conformations have a number of characteristic features of the molten globule state, which is believed to be the universal folding transition state for the majority of globular proteins (Ohgushi & Wada, 1983; Kuwajima, 1989; Privalov

**Table 1. Comparison of the cRMSD values (in Å) in respect to the native structure obtained with two types of short-range potentials for 4-residue, 6-residue, and 8-residue fragments.**

| Protein | | | 4 Amino acids long | | 6 Amino acids long | | 8 Amino acids long | |
|---|---|---|---|---|---|---|---|---|
| Name | Structure | N | Statistical | Homology | Statistical | Homology | Statistical | Homology |
| 1aba | $\alpha/\beta$ | 87 | 1.98 | 1.88 | 2.79 | 2.66 | 3.46 | 3.30 |
| 1aho | $\alpha+\beta$ | 64 | 1.80 | 1.70 | 2.61 | 2.46 | 3.34 | 3.13 |
| 1aly | $\beta$ | 146 | 1.65 | 1.60 | 2.42 | 2.32 | 3.11 | 2.98 |
| 1bdo | $\beta$ | 80 | 1.66 | 1.61 | 2.41 | 2.34 | 3.13 | 3.05 |
| 1ctj | $\alpha$ | 89 | 2.20 | 2.04 | 3.12 | 2.91 | 3.84 | 3.59 |
| 1dhn | $\alpha+\beta$ | 121 | 1.87 | 1.75 | 2.66 | 2.49 | 3.35 | 3.13 |
| 1eca | $\alpha$ | 136 | 2.34 | 2.08 | 3.14 | 2.84 | 3.73 | 3.37 |
| 1erv | $\alpha/\beta$ | 105 | 1.99 | 1.84 | 2.80 | 2.61 | 3.46 | 3.24 |
| 1fna | $\beta$ | 91 | 1.63 | 1.54 | 2.41 | 2.25 | 3.13 | 2.91 |
| 1gvp | $\beta$ | 87 | 1.66 | 1.61 | 2.45 | 2.37 | 3.18 | 3.06 |
| 1hoe | $\beta$ | 74 | 1.71 | 1.67 | 2.54 | 2.46 | 3.30 | 3.18 |
| 1hyp | $\alpha$ | 75 | 2.13 | 2.00 | 2.93 | 2.77 | 3.55 | 3.36 |
| 1jer | $\beta$ | 110 | 1.80 | 1.75 | 2.60 | 2.51 | 3.28 | 3.15 |
| 1nxb | $\beta$ | 62 | 1.72 | 1.64 | 2.56 | 2.41 | 3.35 | 3.12 |
| 1onc | $\alpha+\beta$ | 102 | 1.81 | 1.75 | 2.59 | 2.52 | 3.28 | 3.19 |
| 1ptq | $\alpha+\beta$ | 50 | 1.76 | 1.78 | 2.58 | 2.59 | 3.32 | 3.30 |
| 1tif | $\alpha+\beta$ | 76 | 1.96 | 1.86 | 2.73 | 2.58 | 3.35 | 3.17 |
| 1utg | $\alpha$ | 70 | 2.32 | 2.11 | 3.16 | 2.91 | 3.74 | 3.44 |
| 1vcc | $\alpha+\beta$ | 77 | 1.80 | 1.74 | 2.58 | 2.50 | 3.26 | 3.16 |
| 1vie | $\beta$ | 60 | 1.72 | 1.68 | 2.53 | 2.47 | 3.25 | 3.17 |
| average | | | 1.88 | 1.78 | 2.68 | 2.55 | 3.37 | 3.20 |

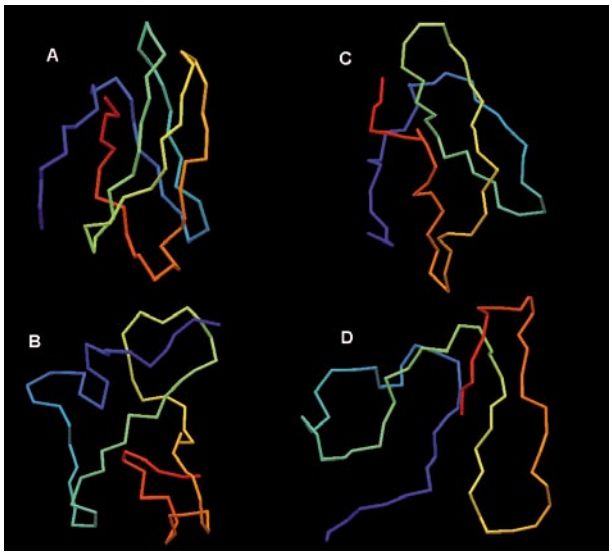
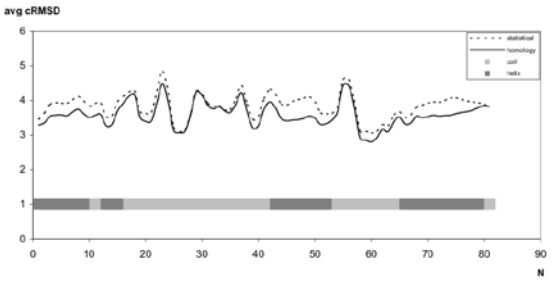



**Figure 1. Example conformations of the 1hoe protein.**
A, Native structure; B–D, selected snapshots from the REMC simulations. Structure C has a native-like, loosely packed conformation, characteristic for the molten globule transition state.

*et al.*, 1989; Ptitsyn *et al.*, 1990). The local structure (the data for the short, 4–8 residue fragments), even on the supersecondary level (see the data for the 16-residue fragments) is on average native-like. Frequently, the secondary structure elements (Table 2) adopt structures very close to the native one, although the mutual arrangement of at least some of these elements is usually incorrect. The size of globule corresponds to a loosely packed semi-compact, yet mobile structure. Sometimes, distorted native-like conformations could be observed, as it could be seen after inspection of the data collected in Table 2. Example snapshots given in Fig. 1 provide an additional illustration to the above statements. Consequently, one may postulate that the CABS model, in spite of the way its force field has been developed, is suitable not only for structure prediction but also for the study of the static and dynamic properties of proteins and polypeptides in the denatured state. This opens up a possibility to study folding mechanisms, as it has been shown in another work from this series, devoted to biophysical applications of the CABS modeling tool (Ekonomiuk *et al.*, 2005).

An interesting insight could be gained from the analysis of the conformational properties of the putative (seen in the native state) secondary elements of various types. In Figs. 2–4 we plotted the average cRMSD along the polypeptide chain for the three distinct types of proteins, all-alpha, beta and alpha/beta, measured for 8-residue fragments. Clearly, the average conformations of the fragments are not random; they have some native-like features, albeit with significant distortions (Dolgikh *et al.*, 1981). Random conformations would have the average values of the cRMSD about two times bigger. Interestingly, the putative extended fragments are reproduced with a higher fidelity than are the coil and helix fragments. The reason is most likely of entropic nature. At a temperature above the folding transition the hydrogen bonds and side group interactions are probably too weak to compact helical structures which have much lower entropy than the expanded ones. The loosely defined expanded states can be obtained in many ways. Additionally, steric clashes are more frequent in helix-like structures, and thereby the excluded volume favors also the expanded conformations. Nevertheless, fluctuating helices are observed frequently, and their locations along the sequence correlate well with the native helices.




**Figuer 2. Average cRMSD for 8-residue fragments plotted against the position in the 1ctj protein chain.**
The dashed line corresponds to simple statistical potential, the solid one is for homology profile-based short-range potentials. The bar at the bottom of the figure depicts the protein secondary structure, with black sections corresponding to the extended state, the gray ones to helices, and the light-gray ones to coil fragments.

The picture of the molten globule-like denatured state of the protein emerging from the present studies provides an interesting explanation of the so called Levinthal's paradox (Levinthal, 1968; Zwanzig *et al.*, 1992). The Levinthal's paradox may be formulated as follows. In a polypeptide chain the average number of rotameric isomers per one residue

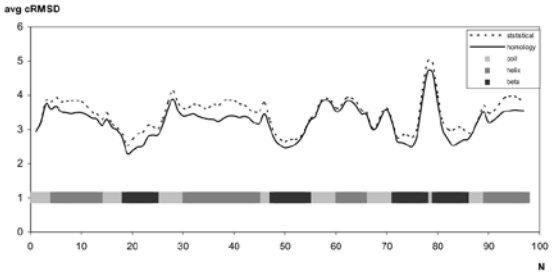



**Figure 3. Average cRMSD for 8-residue fragments plotted against the position in the 1erv protein chain.**
The dashed line corresponds to simple statistical potential, the solid one is for homology profile-based short-range potentials. The bar at the bottom of the figure depicts the protein secondary structure, with black sections corresponding to the extended state, the gray ones to helices, and the light-gray ones to coil fragments.

**Table 2. Comparison of the average and minimal cRMSD values (in Å) in respect to the native structure obtained with two types of short-range potentials for 16-residue fragments and entire proteins**

| Protein | | 16 Amino acids long | | | | Entire protein | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Statistical | | Homology | | Statistical | | Homology | |
| Name | N | Minimal | Average | Minimal | Average | Minimal | Average | Minimal | Average |
| 1aba | 87 | 2.27 | 5.82 | 1.41 | 5.53 | 9.68 | 14.84 | 9.21 | 14.41 |
| 1aho | 64 | 2.44 | 5.86 | 1.88 | 5.46 | 8.00 | 13.30 | 8.16 | 12.98 |
| 1aly | 146 | 1.79 | 5.80 | 1.81 | 5.58 | 13.32 | 19.84 | 13.70 | 19.35 |
| 1bdo | 80 | 1.99 | 5.68 | 1.83 | 5.65 | 9.61 | 15.08 | 11.11 | 15.20 |
| 1ctj | 89 | 2.83 | 6.11 | 1.85 | 5.69 | 9.57 | 14.72 | 8.37 | 14.06 |
| 1dhn | 121 | 2.01 | 5.89 | 1.44 | 5.43 | 10.84 | 17.84 | 11.78 | 17.34 |
| 1eca | 136 | 2.59 | 5.62 | 0.89 | 5.01 | 12.36 | 18.00 | 13.08 | 17.07 |
| 1erv | 105 | 2.13 | 5.68 | 1.58 | 5.34 | 10.93 | 15.60 | 9.63 | 15.63 |
| 1fna | 91 | 1.78 | 5.88 | 1.67 | 5.49 | 10.31 | 16.08 | 9.90 | 15.95 |
| 1gvp | 87 | 1.78 | 5.96 | 1.74 | 5.68 | 8.07 | 15.64 | 8.58 | 15.07 |
| 1hoe | 74 | 1.94 | 6.22 | 1.98 | 6.02 | 10.61 | 14.78 | 10.16 | 14.58 |
| 1hyp | 75 | 2.43 | 5.49 | 2.23 | 5.16 | 7.00 | 13.42 | 7.51 | 13.02 |
| 1jer | 110 | 1.67 | 5.73 | 1.89 | 5.47 | 12.62 | 17.69 | 13.15 | 17.44 |
| 1nxb | 62 | 1.44 | 6.26 | 1.99 | 5.86 | 8.82 | 13.96 | 8.70 | 13.53 |
| 1onc | 102 | 2.07 | 5.87 | 1.90 | 5.77 | 10.66 | 16.16 | 11.59 | 16.10 |
| 1ptq | 50 | 2.39 | 6.08 | 2.33 | 5.89 | 7.40 | 12.02 | 7.04 | 11.63 |
| 1tif | 76 | 2.22 | 5.54 | 1.41 | 5.30 | 7.92 | 13.77 | 9.11 | 13.46 |
| 1utg | 70 | 2.90 | 5.64 | 1.91 | 5.13 | 7.34 | 12.96 | 7.49 | 12.51 |
| 1vcc | 77 | 2.09 | 5.54 | 1.76 | 5.43 | 8.93 | 13.30 | 8.84 | 13.52 |
| 1vie | 60 | 2.11 | 6.14 | 2.04 | 6.03 | 7.22 | 13.43 | 9.05 | 13.76 |
| average | | 2.14 | 5.84 | 1.78 | 5.54 | 9.56 | 15.12 | 9.81 | 14.83 |

is in the range of 10. Thus, the number of possible conformations of a small protein (say, composed of 100 amino acids) is equal to $10^{100}$. A random search of this enormous space for a unique native structure would lead to the folding time that is longer than the estimated age of the Universe. Consequently, the folding routes can not be completely random. A number of explanations of the Levinthal's paradox have been proposed (Skolnick & Kolinski, 1990; Zwanzig *et al.*, 1992). The results of the present work provide a rather clear picture. The denatured state is far from being completely random. Short fragments are native-like, therefore folding should be to some extent viewed as a process of assembly from a much smaller number of partially "prefabricated" fragments than the number of residues. Moreover, due to excluded volume effects, the number of possible mutual orientations of two such fragments is relatively small. An additional factor compressing the conformational space is the relatively compact structure of the denatured state – it is far from the statistical characteristics of random coil polymers. Recently, a number of experiments have shown that the denatured state of proteins is more compact, at least locally, than it has been assumed (Xie & Freire, 1994; Navon *et al.*, 2001). Due to the above, the relatively fast folding, employing rapid narrowing of the number of accessible conformations, is not surprising.

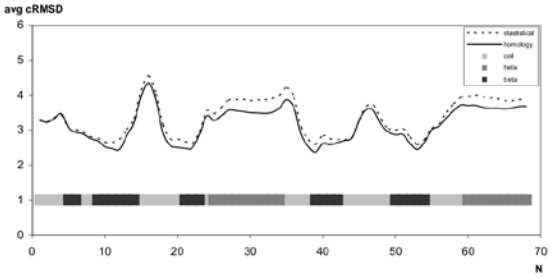



**Figure 4. Average cRMSD for 8-residue fragments plotted against the position in the 1tif protein chain.**
The dashed line corresponds to simple statistical potential, the solid one is for homology profile-based short-range potentials. The bar at the bottom of the figure depicts the protein secondary structure, with black sections corresponding to the extended state, the gray ones to helices, and the light-gray ones to coil fragments.

### Early folding events observed in MCD simulations – comparison with experiment

The MCD simulations were performed for cytochrome *c*, barnase, chymotrypsin inhibitor CI2 and Src SH3 domain. These proteins exhibit very

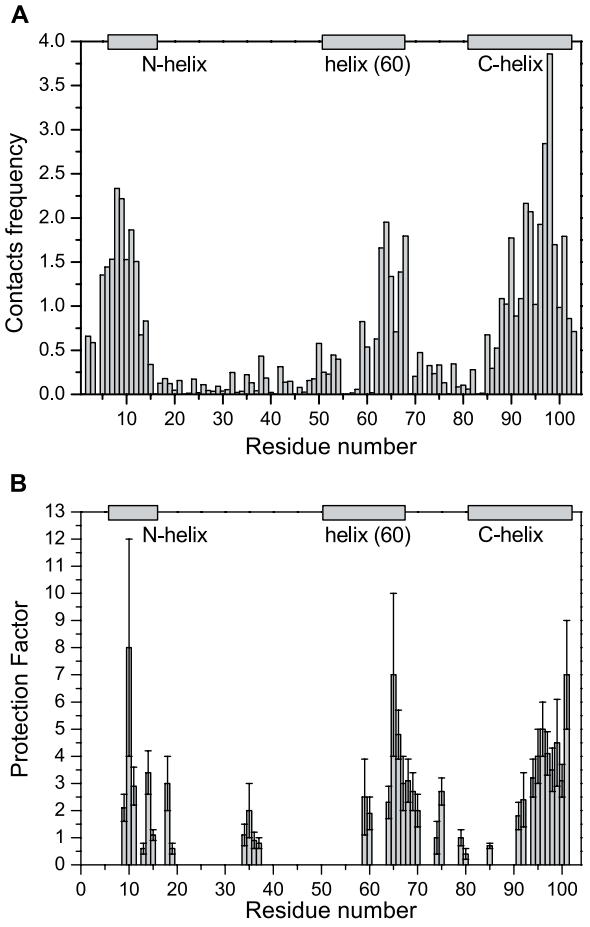



**Figure 5. (A) Frequency of native contacts along the sequence of cytochrome c observed during isothermal MCD simulation slightly above the transition temperature.**
Short-range (|i-j| < 3) contacts were ignored for clarity.
**(B) Protection factors and estimated errors for the burst intermediate of cytochrome c, corrected for any residual protection in the unfolded state (Sauder & Roder, 1998).**
Significant protection during the first 2 ms of folding is seen in the three helical regions, indicated at the top.

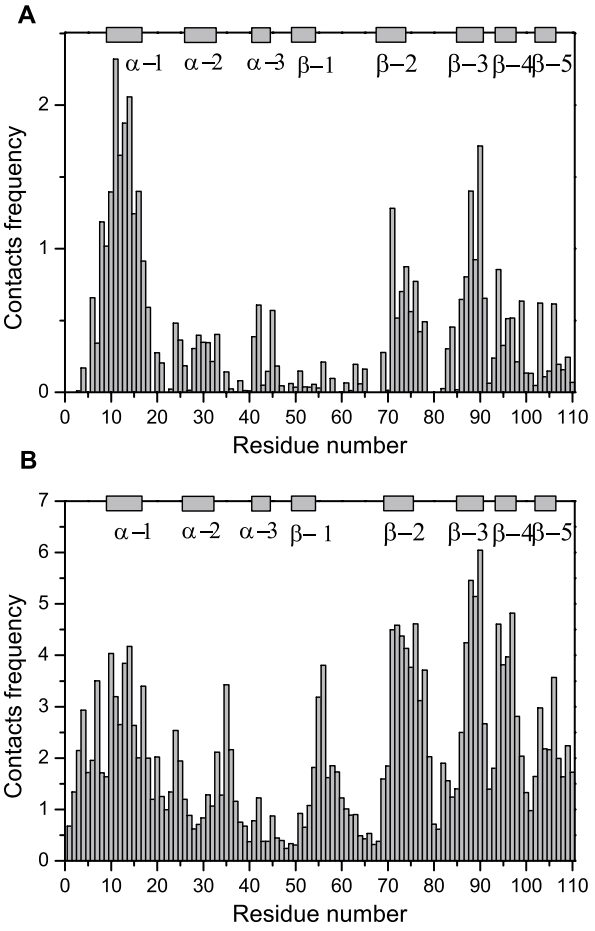



**Figure 6. Frequency of native contacts (A) and all contacts (B) observed during isothermal simulation of barnase near the transition temperature.**
Short-range (|i-j| < 3) contacts were ignored for clarity.

different folding mechanisms and therefore provide a good difficult test of the proposed reduced-space simulation technique.

#### Cytochrome *c*

Cytochrome *c* is a single domain protein of 104 amino acids. It is built of three helices (Bushnell *et al.*, 1990): N-terminal (residues 6–14), C-terminal (residues 87–102) and the 60's helix (residues 60–69). CD measurements on the folding process of cytochrome *c* evidenced stepwise formation of these helices through two folding intermediates (Akiyama *et al.*, 2000). First, the unfolded protein condenses into the compact intermediate I, which has an $\alpha$-helical content that is about 20% of the native level (Akiyama *et al.*, 2000). Intermediate I has several characteristics of denatured protein but exists along the pathway as an individual intermediate (Sosnick *et al.*, 1997; Shastry & Roder, 1998). Subsequent changes are characterized by a further collapse of the polypeptide chain and the development of about 70% of the native secondary structure. This considerable amount of $\alpha$-helical content in intermediate II is consistent with the general definition of the molten globule state (Ptitsyn *et al.*, 1990; Ptitsyn, 1995). Furthermore, hydrogen-exchange labeling coupled with NMR (used to monitor the formation of a stable hydrogen-bonded and solvent-excluded structure) demonstrated significant protection from exchange in all three helices, whereas the other regions showed no significant protection (Houry *et al.*, 1998; Sauder & Roder, 1998) (Fig. 5). However, the protection factors were several orders of magnitude smaller than those in the native state, which suggested that the helical structures were loosely packed.

A histogram of frequency of the native contacts (Fig. 5A) obtained as an average from the MCD simulations is consistent with the molten globule state illustrated by the histogram of the protection factors (Fig. 5B). We can observe the presence of a considerable amount of secondary structure and an absence of the precise tertiary

**Table 3. Average and minimal radii of gyration in respect to native values for the set of test proteins**

| Protein | # aa | Potential | | | |
|---|---|---|---|---|---|
| | | Statistical | | Homology | |
| | | $<R_{gyr}>$ | $R_{gyr}min$ | $<R_{gyr}>$ | $R_{gyr}min$ |
| 1aba | 89 | 1.258 | 0.989 | 1.248 | 0.976 |
| 1aho | 66 | 1.378 | 1.104 | 1.381 | 1.034 |
| 1aly | 148 | 1.206 | 1.007 | 1.220 | 0.989 |
| 1bdo | 82 | 1.340 | 1.051 | 1.336 | 1.048 |
| 1ctj | 91 | 1.430 | 1.133 | 1.376 | 1.097 |
| 1dhn | 123 | 1.125 | 0.914 | 1.117 | 0.902 |
| 1eca | 138 | 1.257 | 1.032 | 1.213 | 1.025 |
| 1erv | 107 | 1.391 | 1.133 | 1.380 | 1.092 |
| 1fna | 93 | 1.236 | 0.991 | 1.234 | 1.019 |
| 1gvp | 89 | 1.106 | 0.887 | 1.097 | 0.884 |
| 1hoe | 76 | 1.365 | 1.107 | 1.347 | 1.089 |
| 1hyp | 77 | 1.283 | 1.059 | 1.261 | 1.008 |
| 1jer | 112 | 1.301 | 1.070 | 1.298 | 1.065 |
| 1nxb | 64 | 1.318 | 1.016 | 1.312 | 1.029 |
| 1onc | 105 | 1.235 | 0.992 | 1.231 | 0.988 |
| 1ptq | 52 | 1.317 | 1.001 | 1.287 | 0.996 |
| 1tif | 74 | 1.259 | 0.955 | 1.252 | 0.999 |
| 1utg | 72 | 1.196 | 0.921 | 1.143 | 0.868 |
| 1vcc | 79 | 1.234 | 0.992 | 1.229 | 0.979 |
| 1vie | 62 | 1.338 | 1.027 | 1.359 | 1.099 |
| avg. | | 1.279 | 1.019 | 1.266 | 1.009 |

structure, although on average the native contacts in the helices form very frequently. The number of all the observed (native and non-native) contacts is by about 50% larger, although it follows the same pattern along the sequence.

Mutagenesis studies on the static molten globule state of cytochrome *c* indicate that close tertiary contacts are established between the N-terminal and C-terminal helices (Marmorino & Pielak, 1995; Colon *et al.*, 1996). Our simulations show very frequent contacts of Phe-10 and Val-11 with Leu-94, Ile-97 and Val-98, which occupy central positions at the interface between the terminal helices and are among the most frequent long-range (well separated along the sequence) tertiary contacts.

In our observation the C-terminal helix is much more stable and the interactions between the N-terminal and the C-terminal helix play an important role in the N-terminal helix formation. It has also been shown experimentally that association of the terminal helices is a critical early event in the folding process and is responsible for the formation of a partially folded intermediate (Colon *et al.*, 1996). Folding of cytochrome *c* is a good example of the collision-diffusion (Karplus & Weaver, 1979) model of protein folding.

**Barnase**

Barnase is a 110-residue protein of $\alpha$+$\beta$ structure (Bycroft *et al.*, 1991). Barnase is an example of a small multidomain protein that folds in a three-state fashion through an experimentally detectable intermediate and its folding has been described in detail (Bycroft *et al.*, 1990; Fersht, 1993; 1997). Barnase contains a considerable amount of residual structure in its denatured state, as probed by NMR and other experimental techniques (Wong *et al.*, 2000).

In its major hydrophobic core the folding is nucleated by formation of a native-like helical structure and hydrophobic clusters located around $\beta$4 and $\beta$3. It has been shown that this hairpin participates in the helix formation which has been called "contact assisted secondary structure formation" (Bond *et al.*, 1997). Definitely, in our simulations, the most frequent non-local contacts of the $\alpha$1 helix are with the $\beta$3 $\beta$4 hairpin (especially with the $\beta$4 strand). However, the $\beta$3 strand interacts more frequently with the $\beta$2 strand than with $\beta$4, which all together form the central part of the $\beta$-sheet. This picture of the $\beta$-sheet and the hydrophobic core formed by the packing of $\alpha$1 against the $\beta$ sheet is characteristic for the major intermediate state observed in experiments as well as in molecular dynamics simulations (Li & Daggett, 1998).

The structure of core 2, the smallest of the three cores formed by residues contained within $\alpha$2, $\alpha$3, loop1, loop2 and $\beta$1 tends to be disrupted, it has definitely the smallest number of the native contacts. Core 3 is formed by packing of loop3 and loop5 against the $\beta$-sheet. Like core 1, core 3 seems to be partially formed. Noteworthy is the large number of nonnative interactions of Phe-56 (belonging to core 3) and neighboring residues against the $\beta$ sheet. The patterns of the native contacts and all the contacts are quite similar (Fig. 6), and they are in agreement with the above experimental findings (Fersht, 1997). Thus the emerging picture of the denatured/intermediate state indicates a loosely defined native topology, which provides a scaffold for later folding events.

**Chymotrypsin inhibitor**

The 64-residue CI2 protein is built of a single $\alpha$ helix running from residues 12 to 24, and a mixed parallel/antiparallel $\beta$ sheet. The strands and the helix form a single hydrophobic core. CI2 is a single-module protein — the interatomic interactions are quite uniform over the structure and there are no regions that make significantly more tertiary interactions within themselves than they do with neighboring regions.

The results from NMR experiments and MD simulations (Ferguson & Fersht, 2003) indicate that CI2 has a highly unfolded denatured state with only

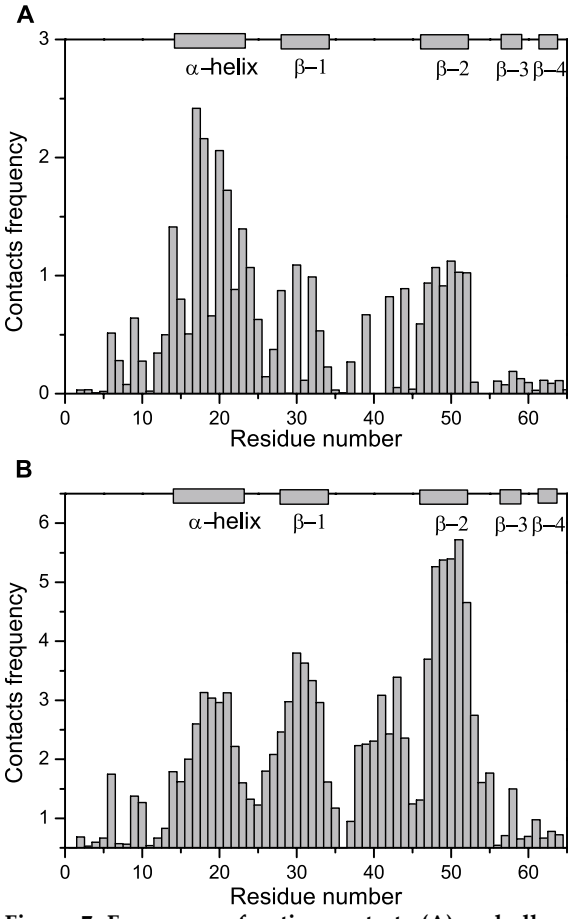



**Figure 7. Frequency of native contacts (A) and all contacts (B) observed during isothermal simulation of the CI2 protein near the transition temperature.**
Short-range ($|i-j| < 3$) contacts were ignored for clarity.

two small regions of a dynamic residual structure: the native-like helix along with hydrophobic clustering in the center of the chain (Kazmirski *et al.*, 2001). This is in good agreement with our observations. The $\alpha$–helix and the $\beta$-sheet between strands $\beta$1 and $\beta$2 possess the most stable native residual structure (Fig. 7). Additionally, a major role in the nonnative interactions is played by the $\beta$2 strand contacting most frequently with $\beta$1, less frequently with the region between both strands, and exceptionally with the $\alpha$-helix (Fig. 7). The nonnative interactions have a hydrophobic character — a loose, extended hydrophobic core is formed. The folding of CI2 starts from a collapse, or condensation, around an extended nucleus that consists of portions of the helix and the $\beta$-sheet. This is followed by subsequent consolidation of the tertiary and secondary structure. This is a classical example of the nucleation-condensation (Fersht, 1997) mechanism of protein folding.

**SH3 domain**

The SH3 domain is a 57-residue globular protein that consist of two $\beta$-sheets packed orthogonally to form a single hydrophobic core (Xu *et al.*, 1997). Extensive experimental characterization of the ther-

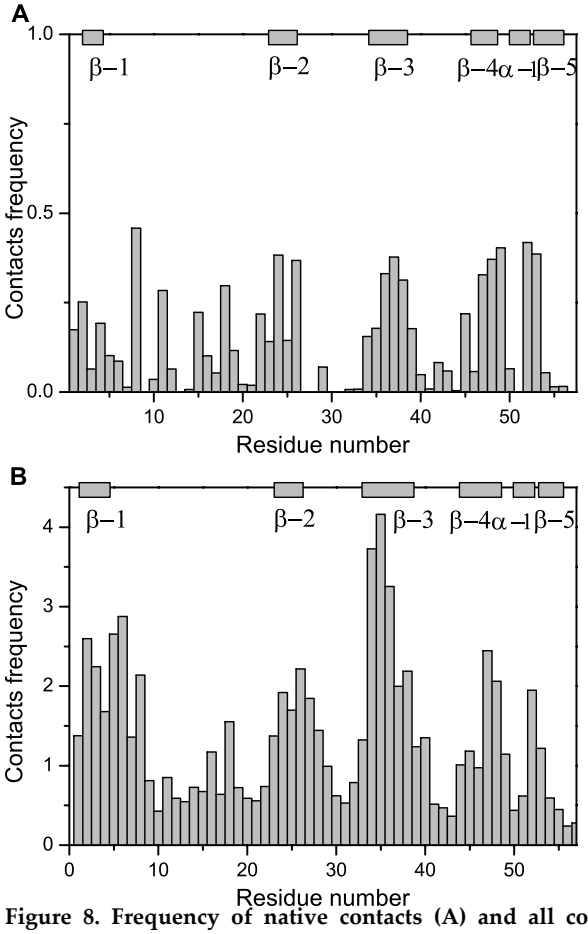



**Figure 8. Frequency of native contacts (A) and all contacts (B) observed during isothermal simulation of the SH3 domain near the transition temperature.**
Short-range ($|i-j| < 3$) contacts were ignored for clarity.

modynamics and kinetics of the Src SH3 domain exhibited a two-state folding process without detectable populations of partially folded intermediates (Grantcharova & Baker, 1997). Our simulations are consistent with the experimental data. Indeed, the distribution of the native contacts along the sequence is more or less flat. The local fluctuations reflect the hydrophobic/hydrophilic pattern, and the overall number of observed native contacts is a relatively small fraction of all the contacts seen near the folding transition.

Interestingly, the local geometry (secondary propensities) seen for all the tested proteins (including the SH3 domain) is highly correlated with the secondary structure of the native state (Grantcharova *et al.*, 2000).

## CONCLUSIONS

In this work we used a reduced representation model in studies of conformational properties of proteins in the denatured state. Analysis of the simulation results shows that protein chains in these conditions are locally native-like, although with significant fluctuations of their geometry. This is a quite common property of some globular proteins (Dolgikh *et al.*, 1981). The model chains above the folding temperature are relatively compact and exhibit characteristic features of the molten globule state, postulated as a universal transition state in protein folding (Dolgikh *et al.*, 1981; Ohgushi & Wada, 1983; Kuwajima, 1989; Privalov *et al.*, 1989; Ptitsyn *et al.*, 1990). The simulation results provide a convincing explanation of the Levinthal's paradox (Levinthal, 1968) and are in agreement with known experimental facts. Moreover, the experimental protection factors and other experimental findings are consistent with the properties of the early folding intermediates observed in the simulations. Thus, it may be postulated that reduced models employing knowledge-based force field, such as the CABS model explored here, could be used not only for protein structure prediction but also for the study of the denatured state structure, dynamics and general aspects of the protein folding mechanisms. The model reproduces the qualitative differences between the folding pathways of various proteins. Recently, other reduced models (Klimov & Thirumalai, 2000), employing physics-based (instead of knowledge-based) potentials (Oldziej *et al.*, 2005), have been successfully used in simulations of protein folding pathways (Liwo *et al.*, 2005). Finally, it should be noted that the structures in the reduced representation could be subject to an all-atom rebuilding procedure (Feig *et al.*, 2000). Since the side chain contacts and the secondary structure geometry are very well defined in the reduced representation the rebuilding procedure was not necessary in the present studies.

**Acknowledgements**

This work was partially supported by grant No. PBZ-KBN-088/P04/2003.

## REFERENCES

Akiyama S, Takahashi S, Ishimori K, Morishima I (2000) Stepwise formation of alpha-helices during cytochrome *c* folding. *Nat Struct Biol* **7**: 514–520.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Baker EN, Arcus VL, Lott JS (2003) Protein structure prediction and analysis as a tool for functional genomics. *Appl Bioinformatics* **2**: S3–10.

Bartlett GJ, Todd AE, Thornton JM (2003) Inferring protein function from structure. *Methods Biochem Anal* **44**: 387–407.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242.

Bolesta E, Kowalczyk A, Wierzbicki A, Rotkiewicz P, Bambach B, Tsao CY, Horwacik I, Kolinski A, Rokita H, Brecher M, Wang X, Ferrone S, Kozbor D (2005) DNA vaccine expressing the mimotope of GD2 ganglioside induces protective GD2 cross-reactive antibody responses. *Cancer Res* **65**: 3410–3418.

Bond CJ, Wong KB, Clarke J, Fersht AR, Daggett V (1997) Characterization of residual structure in the thermally denatured state of barnase by simulation and experiment: description of the folding pathway. *Proc Natl Acad Sci USA* **94**: 13409–13413.

Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A (2003) Protein fragment reconstruction using various modeling techniques. *J Comput Aided Mol Des* **17**: 725–738.

Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D (2002) De novo prediction of three-dimensional structures for major protein families. *J Mol Biol* **322**: 65–78.

Brooks CL 3rd, Gruebele M, Onuchic JN, Wolynes PG (1998) Chemical physics of protein folding. *Proc Natl Acad Sci USA* **95**: 11037–11038.

Bujnicki JM, Rotkiewicz P, Kolinski A, Rychlewski L (2001) Three-dimensional modeling of the I-TevI homing endonuclease catalytic domain, a GIY-YIG superfamily member, using NMR restraints and Monte Carlo dynamics. *Protein Eng* **14**: 717–721.

Bushnell GW, Louie GV, Brayer GD (1990) High-resolution three-dimensional structure of horse heart cytochrome *c*. *J Mol Biol* **214**: 585–595.

Bycroft M, Matouschek A, Kellis JT Jr, Serrano L, Fersht AR (1990) Detection and characterization of a folding intermediate in barnase by NMR. *Nature* **346**: 488–490.

Bycroft M, Ludvigsen S, Fersht AR, Poulsen FM (1991) Determination of the three-dimensional solution structure of barnase using nuclear magnetic resonance spectroscopy. *Biochemistry* **30**: 8697–8701.

Chance MR, Fiser A, Sali A, Pieper U, Eswar N, Xu G, Fajardo JE, Radhakannan T, Marinkovic N (2004) High-throughput computational and experimental techniques in structural genomics. *Genome Res* **14**: 2145–2154.

Colon W, Elove GA, Wakem LP, Sherman F, Roder H (1996) Side chain packing of the N- and C-terminal helices plays a critical role in the kinetics of cytochrome c folding. *Biochemistry* **35**: 5538–5549.

Contreras-Moreira B, Fitzjohn PW, Bates PA (2002) Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. *Appl Bioinformatics* **1**: 177–190.

Dolgikh DA, Gilmanshin RI, Brazhnikov EV, Bychkova VE, Semisotnov GV, Venyaminov SY, Ptitsyn OB (1981) $\alpha$-Lactalbumin: compact state with fluctuating tertiary structure? *FEBS Lett* **136**: 311–313.

Ekonomiuk D, Kielbasinski M, Kolinski A (2005) Protein modeling with a reduced representation: statistical potentials and protein folding mechanism. *Acta Biochim Polon* **52**: 741–758.

Feig M, Rotkiewicz P, Kolinski A, Skolnick J, Brooks CL 3rd (2000) Accurate reconstruction of all-atom protein representations from side-chain-based low-resolution models. *Proteins* **41**: 86–97.

Ferguson N, Fersht AR (2003) Early events in protein folding. *Curr Opin Struct Biol* **13**: 75–81.

Fersht AR (1993) The sixth Datta Lecture. Protein folding and stability: the pathway of folding of barnase. *FEBS Lett* **325**: 5–16.

Fersht AR (1997) Nucleation mechanisms in protein folding. *Curr Opin Struct Biol* **7**: 3–9.

Grantcharova VP, Baker D (1997) Folding dynamics of the src SH3 domain. *Biochemistry* **36**: 15685–15692.

Grantcharova VP, Riddle DS, Baker D (2000) Long-range order in the src SH3 folding transition state. *Proc Natl Acad Sci USA* **97**: 7084–7089.

Gront D, Kolinski A (2005) A new approach to prediction of short range conformational propensities in proteins. *Boinformatics* **21**: 981–987.

Gront D, Kolinski A, Skolnick J (2000) Comparison of three Monte Carlo search strategies for a protein-like homopolymer model: folding thermodynamics and identification of low-energy structures. *J Chem Phys* **113**: 5065–5071.

Gront D, Kolinski A, Skolnick J (2001) A new combination of Replica Exchange Monte Carlo and histogram analysis for protein folding and thermodynamics. *J Chem Phys* **115**: 1569–1574.

Houry WA, Sauder JM, Roder H, Scheraga HA (1998) Definition of amide protection factors for early kinetic intermediates in protein folding. *Proc Natl Acad Sci USA* **95**: 4299–4302.

Karplus M, Weaver DL (1979) Diffusion-collision model for protein folding. *Biopolymers* **18**: 1421–1437.

Kazmirski SL, Wong KB, Freund SM, Tan YJ, Fersht AR, Daggett V (2001) Protein folding from a highly disordered denatured state: the folding pathway of chymotrypsin inhibitor 2 at atomic resolution. *Proc Natl Acad Sci USA* **98**: 4349–4354.

Klimov DK, Thirumalai D (2000) Mechanisms and kinetics of beta-hairpin formation. *Proc Natl Acad Sci USA* **97**: 2544–2549.

Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. *Acta Biochim Polon* **51**: 349–371.

Kolinski A, Skolnick J (2004) Reduced models of proteins and their applications. *Polymer* **45**: 511–524.

Kolinski A, Betancourt M, Kihara D, Rotkiewicz P, Skolnick J (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, lattice and off-lattice modeling for protein structure prediction and refinement. *Proteins* **44**: 133–149.

Kuwajima K (1989) The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. *Proteins* **6**: 87–103.

Levinthal C (1968) Are there pathways for protein folding? *Chem Phys* **65**: 44–45.

Li A, Daggett V (1998) Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. *J Mol Biol* **275**: 677–694.

Liwo A, Khalili M, Scheraga HA (2005) Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc Natl Acad Sci USA* **102**: 2362–2367.

Malolepsza E, Boniecki M, Kolinski A, Piela L (2005) Theoretical model of prion propagation: A misfolded protein induces misfolding. *Proc Natl Acad Sci USA* **102**: 7835–7840.

Marmorino JL, Pielak GJ (1995) A native tertiary interaction stabilizes the A state of cytochrome *c*. *Biochemistry* **34**: 3140–3143.

Navon A, Ittah V, Landsman P, Scheraga HA, Haas E (2001) Distributions of intramolecular distances in the reduced and denatured states of bovine pancreatic ribonuclease A. Folding initiation structures in the C-terminal portions of the reduced protein. *Biochemistry* **40**: 105–118.

Ohgushi M, Wada A (1983) Molten globule state: a compact form of globular proteins with mobile side-chains. *FEBS Lett* **164**: 21–24.

Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Ma-

kowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA (2005) Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proc Natl Acad Sci USA* **102**: 7547–7552.

Plewczynska D, Kolinski A (2005) Protein folding with a reduced model and inacurate short-range restraints. *Macromol Theory Simul* **14**: 444–451.

Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A (2005) Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins* **59**: 49–57.

Privalov PL, Tiktopulo EI, Venyaminov S, Griko Yu V, Makhatadze GI, Khechinashvili NN (1989) Heat capacity and conformation of proteins in the denatured state. *J Mol Biol* **205**: 737–750.

Ptitsyn OB (1995a) How the molten globule became. *Trends Biochem Sci* **20**: 376–379.

Ptitsyn OB (1995b) Molten globule and protein folding. *Adv Protein Chem* **47**: 83–229.

Ptitsyn OB, Pain RH, Semisotnov GV, Zerovnik E, Razgulyaev OI (1990) Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett* **262**: 20–24.

Rotkiewicz P, Sicinska W, Kolinski A, DeLuca HF (2001) Model of three-dimensional structure of vitamin D receptor and its binding mechanism with 1alpha,25-dihydroxyvitamin D(3). *Proteins* **44**: 188–199.

Sali A (1998) 100,000 protein structures for the biologist. *Nat Struct Biol* **5**: 1029–1032.

Sauder JM, Roder H (1998) Amide protection in an early folding intermediate of cytochrome *c*. *Fold Des* **3**: 293–301.

Shastry MC, Roder H (1998) Evidence for barrier-limited protein folding kinetics on the microsecond time scale. *Nat Struct Biol* **5**: 385–392.

Sicinski RR, Rotkiewicz P, Kolinski A, Sicinska W, Prahl JM, Smith CM, DeLuca HF (2002) 2-Ethyl and 2-ethylidene analogues of 1alpha,25-dihydroxy-19-norvitamin D(3): synthesis, conformational analysis, biological activities, and docking to the modeled rVDR ligand binding domain. *J Med Chem* **45**: 3366–3380.

Skolnick J, Fetrow JS, Kolinski A (2000) Structural genomics and its importance for gene function analysis. *Nat Biotechnol* **18**: 283–287.

Skolnick J, Kolinski A (1990) Simulations of the folding of a globular protein. *Science* **250**: 1121–1125.

Sosnick TR, Shtilerman MD, Mayne L, Englander SW (1997) Ultrafast signals in protein folding and the polypeptide contracted state. *Proc Natl Acad Sci USA* **94**: 8545–8550.

Swendsen RH, Wang JS (1986) Relica Monte Carlo simulations. *Phys Rev Lett* **57**: 2607–2609.

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**: 307–340.

Wong KB, Clarke J, Bond CJ, Neira JL, Freund SM, Fersht AR, Daggett V (2000) Towards a complete description of the structural and dynamic properties of the denatured state of barnase and the role of residual structure in folding. *J Mol Biol* **296**: 1257–1282.

Xie D, Freire E (1994) Molecular basis of cooperativity in protein folding. V. Thermodynamic and structural conditions for the stabilization of compact denatured states. *Proteins* **19**: 291–301.

Xu W, Harrison SC, Eck MJ (1997) Three-dimensional structure of the tyrosine kinase c-Src. *Nature* **385**: 595–602.

Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's paradox. *Proc Natl Acad Sci USA* **89**: 20–22.