Two-State Folding, Folding through Intermediates, and Metastability in a Minimalistic Hydrophobic-Polar Model for Proteins

Stefan Schnabel,* Michael Bachmann,† and Wolfhard Janke‡

Institut für Theoretische Physik and Centre for Theoretical Sciences (NTZ), Universität Leipzig,
Augustusplatz 10/11, D-04109 Leipzig, Germany

(Received 23 March 2006; published 25 January 2007)

Within the frame of an effective, coarse-grained hydrophobic-polar protein model, we employ multicanonical Monte Carlo simulations to investigate free-energy landscapes and folding channels of exemplified heteropolymer sequences, which are permutations of each other. Despite the simplicity of the model, the knowledge of the free-energy landscape in dependence of a suitable system order parameter enables us to reveal complex folding characteristics known from real bioproteins and synthetic peptides, such as two-state folding, folding through weakly stable intermediates, and glassy metastability.

DOI: 10.1103/PhysRevLett.98.048103 PACS numbers: 87.15.Cc, 05.10.−a, 87.15.Aa

Folding of linear chains of amino acids, i.e., bioproteins and synthetic peptides, is, for single-domain macromolecules, accompanied by the formation of secondary structures (helices, sheets, turns) and the tertiary hydrophobic-core collapse. While secondary structures are typically localized to segments of the peptide, the effective hydrophobic interaction between nonbonded, nonpolar amino acid side chains results in a global, cooperative arrangement favoring folds with compact hydrophobic core and surrounding polar shell screening the core from the polar solvent. Systematic analyses for unraveling general folding principles are extremely difficult in microscopic all-atom approaches, since the folding process is strongly dependent on the "disordered" sequence of amino acids—twenty different types can typically occur in bioproteins—and the native-fold formation is inevitably connected with, at least, significant parts of the sequence. Moreover, for most proteins, the folding process is relatively slow (microseconds to seconds), which is due to a complex, rugged shape of the free-energy landscape [1–3] with "hidden" barriers, depending on sequence properties. Although there is no obvious system parameter that allows for a general description of the accompanying conformational transitions in folding processes (as, for example, the reaction coordinate in chemical reactions), it is known that there are only a few classes of characteristic folding behaviors, mainly single-exponential folding, two-state folding, folding through intermediates, and glasslike folding into metastable conformations [4–9].

An important step forward towards a better theoretical understanding of the basic mechanisms underlying these different classes could be the design and analysis of suitably designed coarse-grained models focusing on mesoscopic scales. The idea to use a strongly simplified model is twofold: first, it is believed that tertiary folding is mainly based on effective hydrophobic interactions such that atomic details play a minor role. Second, systematic comparative folding studies for mutated or permuted sequences are computationally extremely demanding at the atomic level and are to date virtually impossible for realistic proteins. In this Letter, we show that by employing a coarse-grained hydrophobic-polar heteropolymer model [10] and monitoring a simple angular "order" parameter it is indeed possible to identify different complex folding characteristics. This is comparable to studies of phase transitions based on effective order parameters in other disordered systems such as, e.g., spin glasses, where simplified models are successfully employed [11]. The individual folding trajectories as discussed in this work will be characterized by a similarity parameter which is related to the replica overlap parameter used in spin-glass analyses. This is useful as the amino acid sequence induces intrinsic disorder and frustration into the system and therefore a peptide behaves similar to a spin system with a quenched disorder configuration of couplings.

The simplified model [10] used incorporates only two types of amino acids, hydrophobic and polar residues [12], and focuses on qualitative aspects of tertiary heteropolymer folding, such as hydrophobic-core formation [13–17]. This physical, effective-potential approach has to be distinguished from knowledge-based models—typically of Gō type—where the contact map of the final fold already enters as input into the model. The latter models have proven to be useful in understanding two-state folding of selected proteins [18–23]. On the other hand, the kinetics of physics-based models is not biased towards a given structure, and a variety of folding behaviors can be studied. This has particular implications for non-two-state folding and metastability, the latter primary concerning designed synthetic peptides or mutated biopolymers.

Our results are obtained by employing the standard hydrophobic-polar off-lattice AB model [10] in three dimensions for the three sequences listed in Table I. The sequences were chosen from the set of deliberately designed sequences in Ref. [24] and have the same content of hydrophobic A (14 each) and polar B (6 each) residues. In

TABLE I. The three AB 20-mers studied in this Letter and the values of the associated (putative) global energy minima in natural units. Note that the given values for sequence $S3$ belong to two different, almost degenerate folds.

Label	Sequence	Global Energy Minimum
$S1$	$BA_6BA_4BA_2BA_2B_2$	-33.8236
$S2$	$A_4BA_2BABA_2B_2A_3BA_2$	-34.4892
$S3$	$A_4B_2A_4BA_2BA_3B_2A$	-33.5838, -33.5116

the following, we denote by $\mathbf{r}_i$ the spatial position of the ith monomer in the chain $\mathbf{X} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ of N residues. Covalent bonds have unit length. The bending angle between monomers k, $k+1$, and $k+2$ is ϑ_k $(0 \leq \vartheta_k \leq \pi)$ and $\sigma_i = A, B$ symbolizes the type of the monomer. The energy of a conformation is given by $E = E_{\mathrm{bend}} + E_{\mathrm{LJ}}$, where

$$E_{\mathrm{bend}} = \frac{1}{4}\sum_k (1 - \cos\vartheta_k) \tag{1}$$

is the bending energy and

$$E_{\mathrm{LJ}} = 4 \sum_{j>i+1} [r_{ij}^{-12} - C(\sigma_i, \sigma_j) r_{ij}^{-6}] \tag{2}$$

is the contribution of the residue specific Lennard-Jones potential, which depends on the distance r_{ij} of all pairs of nonbonded monomers i and j, being long-range attractive for AA and BB pairs $[C(A, A) = 1,\ C(B, B) = 0.5]$ and repulsive for AB pairs of monomers $[C(A, B) = C(B, A) = -0.5]$. Simulations of this model were performed using standard multicanonical Monte Carlo techniques [25] with spherical updates [17]. For each sequence, 10 independent simulations were performed and a total statistics of 2×10^9 conformations entered into the data analysis.

Since the number of degrees of freedom (virtual bond and torsion angles) in the coarse-grained model is comparable with the number of dihedral angles in all-atom protein models, AB heteropolymer folding is of similar complexity. The main advantage is the drastically reduced computational effort for calculating the interactions, which allows more comprehensive and systematic analyses of free-energy landscapes and folding channels in comparative studies for different sequences. In the following, we perform such an analysis of characteristic folding behaviors based on a suitably defined generalized angular overlap parameter, as introduced in Ref. [17] in analogy to all-atom studies [6]. It is a computationally low-cost measure for the similarity of two conformations, where the differences of the angular degrees of freedom are calculated. In order to consider this parameter as kind of order parameter, it is useful to compare conformations $\mathbf{X}$ of the actual ensemble with a suitable reference structure $\mathbf{X}^{(0)}$, which is preferably chosen to be the global-energy minimum conformation. The overlap parameter is defined as [17]

$$Q(\mathbf{X}) = 1 - d(\mathbf{X}). \tag{3}$$

Denoting by $N_b = N - 2$ and $N_t = N - 3$ the numbers of bending angles ϑ_i and torsional angles φ_i, respectively, the angular deviation between the conformations is calculated according to $d(\mathbf{X}) = [\sum_{i=1}^{N_b} d_b(\vartheta_i) + \max(\sum_{i=1}^{N_t} d_t^{(+)}(\varphi_i), \sum_{i=1}^{N_t} d_t^{(-)}(\varphi_i))]/\pi(N_b + N_t)$, where $d_b(\vartheta_i) = |\vartheta_i - \vartheta_i^{(0)}|$ and $d_t^{(\pm)}(\varphi_i) = \min(|\varphi_i \pm \varphi_i^{(0)}|, 2\pi - |\varphi_i \pm \varphi_i^{(0)}|)$. Note that this expression takes into account the reflection symmetry $\varphi_i \rightarrow -\varphi_i$ of the AB model. Reflection-symmetric conformations are not distinguished and therefore only the larger overlap is considered. The overlap is unity, if all angles coincide, else $0 \leq Q < 1$. The average overlap of a random conformation with the reference state is for the three sequences close to $\langle Q \rangle = 0.66 \pm 0.02$. Significant similarity is typically found if $Q > 0.8$.

For the qualitative discussion of the folding characteristics, we consider the multicanonical histograms of energy E and angular overlap Q, $H_{\mathrm{muca}}(E, Q) = \sum_t \delta_{E,E(\mathbf{X}_t)} \delta_{Q,Q(\mathbf{X}_t)}$, where the sum runs over all Monte Carlo sweeps t in the multicanonical simulation, which yields a constant energy distribution $h_{\mathrm{muca}}(E) = \int_0^1 dQ H_{\mathrm{muca}}(E, Q) \approx \mathrm{const}$. In consequence, $H_{\mathrm{muca}}(E, Q)$ is useful for identifying the folding channels, independently of temperature. Restricting the canonical partition function at temperature T to the "microoverlap" ensemble with overlap Q, $Z(Q) = \int \mathcal{D}\mathbf{X} \delta(Q - Q(\mathbf{X})) \times \exp\{-E(\mathbf{X})/k_B T\}$, where the integral is over all possible conformations $\mathbf{X}$, we define the overlap free energy as $F(Q) = -k_B T \ln Z(Q)$.

Figures 1(a)–1(c) show the thus obtained multicanonical histograms $H_{\mathrm{muca}}(E, Q)$ (left) and the overlap free-energy landscapes $F(Q)$ (right) at different temperatures for the three sequences listed in Table I. The different branches of $H_{\mathrm{muca}}(E, Q)$ indicate the channels the heteropolymer can follow in the folding process towards the reference structure. The heteropolymers, whose sequences differ only by permutations, exhibit noticeable differences in the folding behavior towards the native conformations. The first interesting observation is that the minimalistic model used is capable of revealing the different folding behaviors of the wild-type and permuted sequences. The second remarkable result is that the angular overlap parameter Q is a surprisingly manifest measure for the peptide macrostate.

From Fig. 1(a) we conclude that folding of sequence $S1$ exhibits a typical two-state characteristics. Above the transition temperature, conformations possess a random-coil-like overlap $Q \approx 0.7$, i.e., there is no significant similarity with the reference structure. Close to $T \approx 0.1$ the global minimum of the corresponding overlap free energy $F(Q)$ changes discontinuously towards larger Q values, and at the transition state the denatured (D) and the folded native (N) macrostate are equally probable. The existence of this pronounced transition state is a characteristic indication for first-order-like two-state folding. Decreasing the temperature further, the native-fold-like conformations $(Q > 0.95)$ dominate and fold smoothly towards the $Q = 1$ reference structure, i.e., the lowest-energy conformation (N) found for sequence $S1$, which is also depicted in Fig. 1(a).

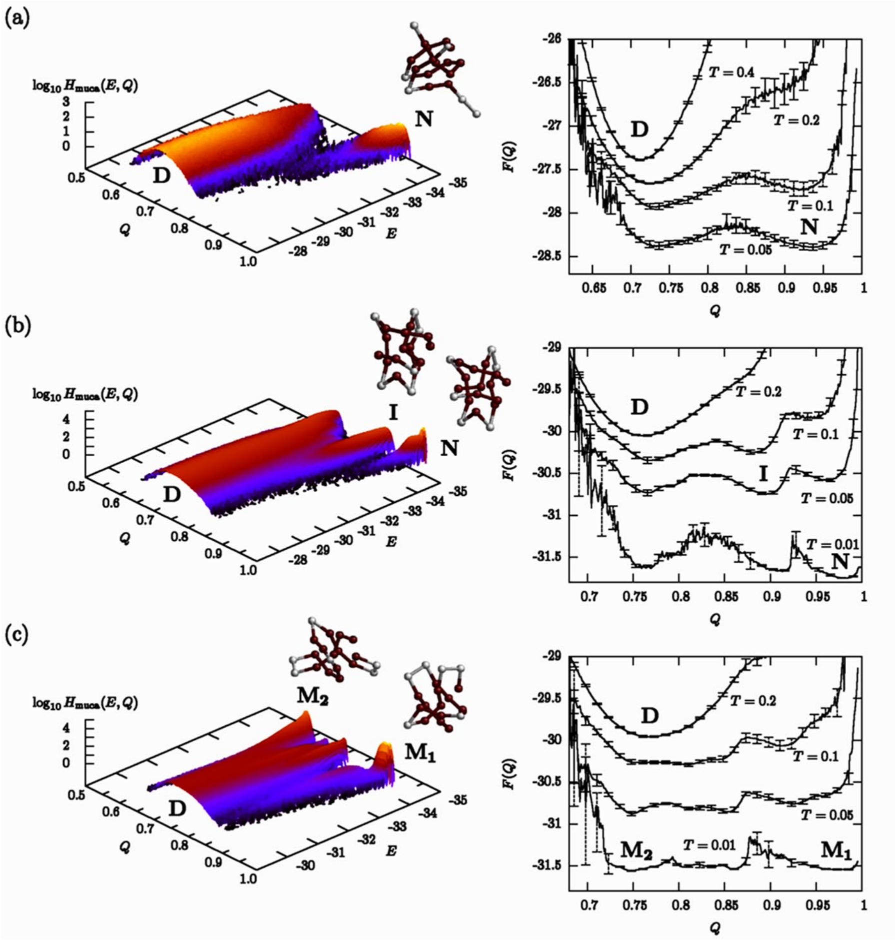

FIG. 1 (color online). Multicanonical histograms $H_{\mathrm{muca}}(E, Q)$ of energy E and angular overlap parameter Q and free-energy landscapes $F(Q)$ at different temperatures for the three sequences (a) $S1$, (b) $S2$, and (c) $S3$. The reference folds reside at $Q = 1$ and $E = E_{\mathrm{min}}$. Pseudophases are symbolized by D (denatured states), N (native folds), I (intermediates), and M (metastable states). Representative conformations in intermediate and folded phases are also shown.

The folding behavior of sequence $S2$ is significantly different, as Fig. 1(b) shows, and is a typical example for a folding event through an intermediate (I) macrostate. The main channel (D) bifurcates and a side channel (I) branches off continuously. For smaller energies (or lower temperatures), this branching is followed by the formation of a third channel, which ends in the native fold (N). The characteristics of folding through intermediates are also reflected by the free-energy landscapes. Starting at high temperatures in the pseudophase D of denatured conformations ($Q \approx 0.76$), the intermediary phase I with $Q \approx 0.9$ is reached close to the temperature $T \approx 0.05$. Decreasing the temperature further below the native-folding threshold close to $T = 0.01$, the hydrophobic-core formation is finished and stable native-fold-like conformations N with $Q > 0.97$ dominate.

The most extreme behavior of the three exemplified sequences is found for sequence $S3$, where the main channel (D) does not decay in favor of a native-fold channel. In fact, in Fig. 1(c) we observe both *two* separate native-fold channels (M_1 and M_2) and the main channel. Above the folding transition ($T \approx 0.2$), the typical sequence-independent denatured (D) conformations ($Q \approx 0.77$) dominate. Annealing below the glass-transition threshold, several channels form and coexist. The two most prominent channels (to which the lowest-energy conformations belong that we found in the simulations) eventually lead for $T \approx 0.01$ to ensembles of states M_1 with $Q > 0.97$, which are similar to the reference structure shown, and conformations M_2 with $Q \approx 0.75$. The lowest-energy conformation found in M_2 is also shown in Fig. 1(c). It is structurally different but energetically almost degenerate compared with the reference structure. It should also be noted that the lowest-energy main-channel conformations have only slightly larger energies than the two native folds. Thus, the folding of this heteropolymer is accompanied by a very complex, amorphous folding characteristics. In fact, the multiple-peaked distribution $H_{\mathrm{muca}}(E, Q)$ near minimum energies is a strong indication for metastability and bears similarities with spin-glass characteristics. A native fold in the natural sense does not exist; the $Q = 1$ conformation is only a reference structure but the folding towards this structure is not distinguished as it is in the folding characteristics of sequences $S1$ and $S2$.

We have confirmed our results of the angular overlap analysis for the folding behaviors by a corresponding study of the root mean square deviation (rmsd) which is frequently used to characterize folding trajectories in free-energy landscapes. The main advantage of using our angular overlap parameter is its efficient calculation which leads to a speed-up of computing time by a factor of about 10 compared with the efforts required for analyzing the folding channels based on the rmsd [26].

To summarize, we have demonstrated in this study that within a minimalistic heteropolymer frame it is possible to find clear indications for three different folding characteristics known from real proteins by analyzing macrostates based on an angular overlap parameter. Our primary physical objective is a more comprehensive, qualitative understanding of universal aspects of tertiary protein folding, where microscopic details are expected to be of less relevance and which are, therefore, averaged out at a mesoscopic scale in a coarse-grained model. For selected hydrophobic-polar heteropolymer sequences—not being explicitly designed for this study—we have shown that characteristic folding behaviors such as two-state folding, folding through intermediates, and metastability can be identified which are qualitatively comparable with real folding events in nature. Beyond the general interest in a theoretical understanding of the basic mechanisms of protein folding, the preparation of synthetic peptide macrostates in future applications, e.g., the successful design of substrate- or pattern-selective polymers [27–31], is strongly connected with the complex aspects of conformational folding transitions as investigated in this study.

This work is partially supported by the DFG (German Science Foundation) under Grant No. JA 483/24-1. Some simulations were performed on the supercomputer JUMP of the John von Neumann Institute for Computing (NIC), Forschungszentrum Jülich under Grant No. hlz11.

*Email address: Stefan.Schnabel@itp.uni-leipzig.de
†Email address: Michael.Bachmann@itp.uni-leipzig.de
‡Email address: Wolfhard.Janke@itp.uni-leipzig.de
Electronic address: http://www.physik.uni-leipzig.de/CQT.html

[1] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[2] C. Clementi, A. Maritan, and J.R. Banavar, Phys. Rev. Lett. **81**, 3287 (1998).

[3] J.N. Onuchic and P.G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).

[4] R. Du, V.S. Pande, A.Yu. Grosberg, T. Tanaka, and E.S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).

[5] V.S. Pande and D.S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 1273 (1999).

[6] U.H.E. Hansmann, M. Masuya, and Y. Okamoto, Proc. Natl. Acad. Sci. U.S.A. **94**, 10 652 (1997); B.A. Berg, H. Noguchi, and Y. Okamoto, Phys. Rev. E **68**, 036126 (2003).

[7] P.G. Wolynes, in *Directions in Condensed Matter Physics*, edited by D.L. Stein, Vol. 6: Spin Glasses and Biology (World Scientific, Singapore, 1992), p. 225.

[8] V.S. Pande, A.Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).

[9] E. Pitard and E.I. Shakhnovich, Phys. Rev. E **63**, 041501 (2001).

[10] F.H. Stillinger, T. Head-Gordon, and C.L. Hirshfeld, Phys. Rev. E **48**, 1469 (1993); F.H. Stillinger and T. Head-Gordon, Phys. Rev. E **52**, 2872 (1995).

[11] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975); S.F. Edwards and P.W. Anderson, J. Phys. F **5**, 965 (1975); G. Parisi, Phys. Rev. Lett. **43**, 1754 (1979).

[12] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986 (1989).

[13] J.M. Sorenson and T. Head-Gordon, Proteins: Struct., Funct., Genet. **37**, 582 (1999).

[14] M. Bachmann and W. Janke, Phys. Rev. Lett. **91**, 208105 (2003); J. Chem. Phys. **120**, 6779 (2004); Comput. Phys. Commun. **169**, 111 (2005).

[15] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, Phys. Rev. E **68**, 021113 (2003).

[16] F. Liang, J. Chem. Phys. **120**, 6756 (2004).

[17] M. Bachmann, H. Arkın, and W. Janke, Phys. Rev. E **71**, 031906 (2005).

[18] C. Clementi, H. Nymeyer, and J.N. Onuchic, J. Mol. Biol. **298**, 937 (2000).

[19] L. Li and E.I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **98**, 13 014 (2001).

[20] N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).

[21] H. Kaya and H.S. Chan, Phys. Rev. Lett. **90**, 258104 (2003); J. Mol. Biol. **326**, 911 (2003).

[22] J. Schonbrun and K.A. Dill, Proc. Natl. Acad. Sci. U.S.A. **100**, 12 678 (2003).

[23] T. Head-Gordon and S. Brown, Curr. Opin. Struct. Biol. **13**, 160 (2003).

[24] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, J. Chem. Phys. **107**, 273 (1997).

[25] B.A. Berg and T. Neuhaus, Phys. Lett. B **267**, 249 (1991); Phys. Rev. Lett. **68**, 9 (1992).

[26] S. Schnabel, M. Bachmann, and W. Janke (to be published).

[27] S.R. Whaley, D.S. English, E.L. Hu, P.F. Barbara, and A.M. Belcher, Nature (London) **405**, 665 (2000).

[28] M. Sarikaya, C. Tamerler, A.K.-Y. Jen, K. Schulten, and F. Baneyx, Nat. Mater. **2**, 577 (2003).

[29] K. Goede, P. Busch, and M. Grundmann, Nano Lett. **4**, 2115 (2004).

[30] T. Bogner, A. Degenhard, and F. Schmid, Phys. Rev. Lett. **93**, 268108 (2004).

[31] M. Bachmann and W. Janke, Phys. Rev. Lett. **95**, 058102 (2005); Phys. Rev. E **73**, 020901(R) (2006); **73**, 041802 (2006).