Coarse-grained model for protein folding based on structural profiles

Katrin Wolff,[1] Michele Vendruscolo,[2] and Markus Porto[3]
[1]*School of Physics, University of Edinburgh, JCMB Kings Buildings, Edinburgh EH9 3JZ, United Kingdom*
[2]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom*
[3]*Institut für Theoretische Physik, Universität zu Köln, Zülpicher Strasse 77, D-50937 Köln, Germany*
(Received 21 July 2011; published 28 October 2011)

We study a coarse-grained protein model whose primary characteristics are (i) a tubelike geometry to describe the self-avoidance effects of the polypeptide chain and (ii) an energy function based on a one-dimensional structural representation. The latter specifies the connectivity of a sequence in a given conformation, so that the energy function, rather than favoring the formation of specific native pairwise contacts, promotes the establishment of a specific target connectivity for each amino acid. We show that the resulting dynamics is in good agreement with both experimental observations and the results of all-atoms simulations. In contrast to the latter, our coarse-grained approach provides the possibility to explore longer time scales and thus enables one to access, albeit in less detail, larger regions of the conformational space. We illustrate our approach by its application to the villin headpiece domain, a three-helix protein, by studying its folding behavior and determining heat capacities and free-energy landscapes in various reaction coordinates.

DOI: 10.1103/PhysRevE.84.041934 PACS number(s): 87.14.E−, 87.15.A−, 87.15.Cc, 87.15.hm

I. INTRODUCTION

According to our current understanding of protein folding, the energy function governing the folding process provides a bias toward the native state to avoid an extensive sampling of the extremely large number of configurations available to the polypeptide chain [1–4]. Even though detailed atomistic models are able to reproduce this overall feature, the underlying force fields yield energy surfaces that are relatively rough on smaller scales. As a consequence, atomistic models require very significant computational efforts, despite great advances in recent years [5–9], and hence coarse-grained models provide a range of opportunities to investigate the behavior of proteins at a lower level of detail [10–12].

Over the years, a variety of coarse-grained models of proteins have been developed. Many of these models can be defined as native-centric, i.e., in these models a knowledge about the native structure is a crucial ingredient. But even if the native structure is known, it is still not obvious how the native conformation is attained during folding, and one fundamental question is whether and how the native structure determines the possible folding pathways. At the geometrical level, it is sufficient to know which amino acid pairs are spatially close (in contact) in the native structure to construct it with relatively high accuracy [13]. However, whether knowledge of the native contacts is also sufficient to infer major features of the folding process is still unclear.

A particularly well studied and versatile class of coarse-grained models are based on the Gō approximation [14,15], which defines an energy function favoring specifically the formation of native contacts between amino acids. These models are based on the principle of minimal frustration [16], which assumes that the energy landscape displays minimal roughness, as proteins are under evolutionary selection for fast and reliable folding. Gō models have been applied in various contexts and have been shown to capture many important features of the protein folding process [15,17,18]. However, models solely based on additive native contact energy contributions have difficulties in reproducing phenomena such as cooperative folding behavior or misfolding, precisely because of their smooth energy landscape, so additional energy contributions are usually included [10–12,19].

An alternative approach that we proposed recently is a native-centric model whose energy function is based on a one-dimensional structural representation (a structural profile) [20–22]. The latter specifies the sequence's connectivity in a given conformation, so that the energy function, rather than favoring the formation of specific native pairwise contacts as in standard Gō models, promotes the establishment of a specific target connectivity for each amino acid. In comparison with other coarse-grained approaches of comparable complexity, the resulting dynamics is in better agreement with both experimental observations and the results of all-atoms simulations, as we show in the following using the villin headpiece domain as an example. This work is organized as follows: the model and its properties are discussed in detail in Sec. II; results concerning free-energy landscapes, heat capacities, and folding transitions are presented in Sec. III; and Sec. IV reports our conclusions.

II. MODEL

Our protein model uses a coarse-grained representation of a polypeptide chain and defines it as a tube with a finite thickness [23,24], which accounts for both (i) excluded volume by prohibiting overlapping parts and (ii) finite bending rigidity by disallowing too tight angles. The tube can be formalized by a steric energy E_{steric}, which is 0 for allowed conformations and infinite otherwise. In the implementation of the tube and in the move set used in Monte Carlo simulations, we follow Refs. [21,22,24]: The tube consists of spherocylinders (cylinders capped with semispheres) of diameter 3.3 Å, whose axes coincide with the links joining consecutive C_α atoms being 3.8 Å apart. Spherocylinders that do not share a C_α atom are not allowed to intersect. Pivot and crankshaft moves are tried in 10% and in 90% of the cases, respectively. In a pivot move, a C_α atom and an axis of rotation are picked randomly

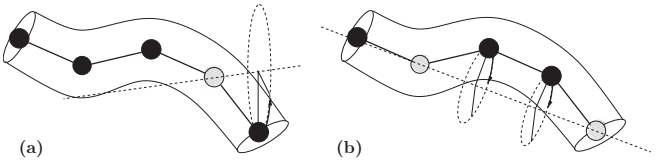

FIG. 1. Monte Carlo move set of the tube (the individual spherocylinders are left out for clarity) exemplified using (a) a pivot and (b) a crankshaft move.

and all C_α atoms with indices larger than the selected one are rotated by an angle ϕ [see Fig. 1(a)]. In a crankshaft move, two C_α atoms are selected randomly such that there are one to four C_α atoms between them, and the intermediate C_α atoms are rotated around the axis determined by their connecting vector by an angle ϕ [see Fig. 1(b)]. In both cases, the angle ϕ is drawn from a Gaussian distribution of mean 0 and standard deviation $\pi/25$.

The tube is combined with an energy term biased toward the (native) target conformation. We consider two variants of such an energy term in the following, one for our effective-connectivity (EC) model and one for a Gō model with which we compare our model.

In the case of the Gō model, the establishment of native contacts is energetically rewarded, whereas the establishment of non-native contacts is energetically neutral or even penalized. This energy can hence be defined using the contact map $\mathbf{C}$ of a given conformation, which is given by

$$C_{ij} = \begin{cases} 1 & |i-j| > 2 \wedge d_{i,j} < r_c \wedge \mathrm{p}(i) \wedge \mathrm{p}(j) \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $d_{i,j}$ is the distance between C_α atom i and C_α atom j and r_c is a cutoff distance for contacts (to which results are largely insensitive), which in the present case is set to 8.5 Å [see Fig. 2(b)]. Contacts with $|i-j| \leqslant 2$ that are present all the time (including self-contacts) are not taken into account.

We restrict the contact map further, for both the target (native) and the current conformation, to contacts between amino acids i and j showing both cooperative contact patterns reminiscent of α helices or β sheets [formally denoted by boolean functions $\mathrm{p}(i)$ and $\mathrm{p}(j)$ in Eq. (1)] [20–22], see Fig. 2(b). For an α helix, contacts between i and $i+k$, $i+1$ and $i+1+k$, up to $i+\ell-1$ and $i+\ell-1+k$ are required. The number of residues per helix turn is 3.6 and hydrogen bonds exist between residues i and $i+4$ or, for 3_{10} helices, between residues i and $i+3$. Therefore, $k=4$ or $k=3$ are allowed. A minimum size of a helix of $\ell_{\min}=4$ consecutive contacts and a chirality $\chi = (\vec{e}_{i,i+1} \times \vec{e}_{i+1,i+2}) \cdot \vec{e}_{i+2,i+3} > 0.2$ is required, where $\vec{e}_{i,i+1}$ is the unit vector pointing from C_α atom i to C_α atom $i+1$. A single residue following or preceding an α helix, and itself having the contact pattern of an α helix but not the correct chirality, is included into the α helix. The contact pattern of parallel β sheets is very much the same except that $k>4$ and there is no constraint on chirality. For antiparallel β sheets, the pattern that has to be followed is contacts between position i and j, $i+1$ and $j-1$, up to $i+\ell-1$ and $j-\ell+1$, without constraint on chirality. The minimum size of β sheets (both parallel and antiparallel) is also set to $\ell_{\min}=4$ consecutive contacts.

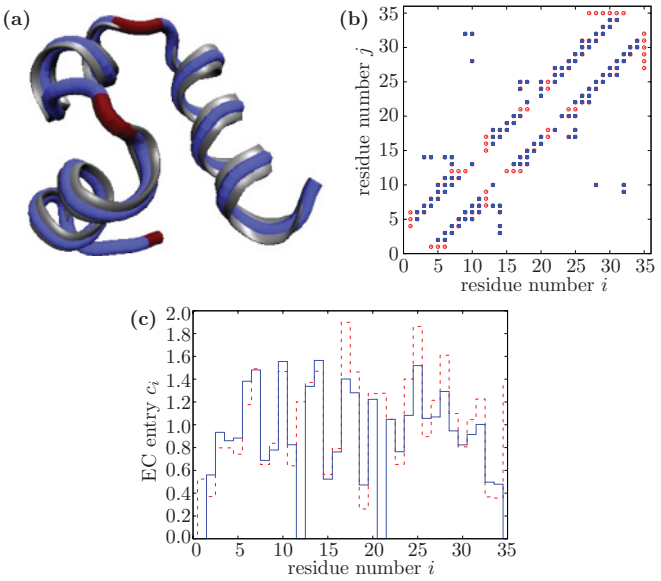

FIG. 2. (Color online) (a) Protein structure with amino acids that are part of cooperative contacts in blue (light gray) and those that are not part in red (dark gray), (b) contact map $\mathbf{C}^{(t)}$, and (c) effective-connectivity (EC) profile $\mathbf{c}^{(t)}$ of the villin headpiece domain (PDB ID 1UND). Restricting the contact map to pairs of amino acids that are both part of secondary structure elements (blue filled squares, here only helices) and disregarding all others (red open circles, shown for comparison), as we do in our simulations, results in a slightly modified contact map and, hence, EC profile [solid blue line for blue (filled squares) contacts only, dashed red line for both blue (filled squares) and red (open circles) contacts shown for comparison].

This restriction to amino acids which show cooperative contacts favors the formation of secondary structure and in effect reduces the conformation space that is sampled. It has been introduced for more efficient sampling in the EC model but its effect on the Gō model is minimal. Nevertheless, the restriction is applied to the computation of both energies to avoid introducing additional differences between the two models. As the energy for the Gō model, we define

$$E_{\mathrm{G\bar{o}}} = -\tilde{\epsilon}\, L \frac{\sum_{i=1}^{L-3}\sum_{j=i+3}^{L} C_{ij}\, C_{ij}^{(t)}}{\max\left\{\sum_{i=1}^{L-3}\sum_{j=i+3}^{L} C_{ij}, \sum_{i=1}^{L-3}\sum_{j=i+3}^{L} C_{ij}^{(t)}\right\}}, \quad (2)$$

which counts the number of contacts that the contact maps of the current conformation $\mathbf{C}$ and of the native conformation $\mathbf{C}^{(t)}$ have in common and normalizes it by the maximum of the two conformations' total number of contacts. Here, L is the number of amino acids in the protein, and the parameter $\tilde{\epsilon}$ sets the scale for temperatures used in simulations, and there are no other parameters to tune interaction strengths. The definition in Eq. (2) has the advantage that non-native contacts are energetically penalized. The total energy then consists of two terms

$$E_{\mathrm{G\bar{o}}}^{\mathrm{tot}} = E_{\mathrm{steric}} + E_{\mathrm{G\bar{o}}}. \quad (3)$$

The EC model, which is the main focus here, is based on the notion of the effective connectivity of an amino acid. The EC profile is a vector quantity whose entries are

the effective-connectivity values of the L amino acids in the protein and which is calculated from the contact map corresponding to a given protein conformation. The EC profile does not specify which individual contacts between amino acid pairs are formed; instead, a given vector component is related to the total number of contacts the corresponding amino acid has. An important property of the EC profile of a protein's native structure (the native EC profile) is that it represents an analytic solution of the "statistical inverse folding problem", which consists in finding the statistical properties of sequences compatible with this structure [25,26]. As such, the EC profile allows us to predict the average hydrophobicity of amino acids at each site for families of proteins sharing the same structure, although its correlation with an individual sequence's hydrophobicity profile may be small [25]. The EC profile $\mathbf{c}$ is computed according to

$$\mathbf{c} = \frac{1}{A}\sum_{k=1}^{L}\frac{\langle v^{(k)}\rangle}{\Lambda-\lambda^{(k)}}\mathbf{v}^{(k)}. \tag{4}$$

Here $\mathbf{v}^{(k)}$ $(k=1,\ldots,L)$ are the L eigenvectors of a given conformation's contact map, Eq. (1), with their eigenvalues $\lambda^{(k)}$, and $\langle v^{(k)}\rangle = L^{-1}\sum_{i=1}^{L} v_i^{(k)}$ is the average of entries $v_i^{(k)}$ of eigenvector k. The parameters $A>0$ and $\Lambda > \lambda^{(k)}$ are used to set the average and variance of $\mathbf{c}$ following Ref. [25]. The EC profile thus contains contributions from all the contact map's eigenvectors, while the contribution from the eigenvector belonging to the largest eigenvalue (principal eigenvector, PE) is largest; and in fact the correlation between the EC profile and PE is high for fully folded globular proteins (see Ref. [25]). While for compact structures with connected contact maps the PE contains full information on the connectivity of amino acids and is sufficient to reconstruct the contact map [27], for noncompact structures (as will be encountered during folding) the PE would only give information on the largest (or most well-connected) block within the contact map. This is the reason why here we employ the EC profile for which the contributions from further eigenvectors serve to include information on less well-connected blocks in the contact map. Note that due to the constraints applied to the contact map (see above), EC components are nonzero only for amino acids that display cooperative contact patterns. The energy of the EC model is defined as

$$E_{\mathrm{EC}} = \epsilon\sum_{i=1}^{L}\left|c_i - c_i^{(\mathrm{t})}\right|, \tag{5}$$

and sums up absolute differences over all amino acids i between the EC profile of the current conformation $\mathbf{c}$ and of the native conformation $\mathbf{c}^{(\mathrm{t})}$. Thus, in contrast to the Gō model, Eq. (2), the formation of native contacts is not necessarily energetically rewarded, nor is the formation of non-native contacts necessarily penalized. It is rather that changes in the contact map making the resulting EC profile more (less) similar to the target one are energetically rewarded (penalized). The parameter ϵ sets the scale for temperatures used in simulations, and there are no other parameters to tune interaction strengths. By changing the term within the sum of Eq. (5) to $|c_i - c_i^{(\mathrm{t})}|^{\alpha}$ with $\alpha > 1$, larger deviations from the target connectivity of single amino acids would be emphasized more strongly. Here we work with $\alpha = 1$ to ensure that no energy contribution from a single amino acid dominates the sum. The total energy then consists of two terms

$$E_{\mathrm{EC}}^{\mathrm{tot}} = E_{\mathrm{steric}} + E_{\mathrm{EC}}. \tag{6}$$

The energies $E_{\mathrm{G\bar{o}}}$ and E_{EC}, Eqs. (2) and (5), can be interpreted as two alternative metrics in conformation space, measuring the distance between the current and the native conformation (in the Gō model, one needs to add a constant, $E_{\mathrm{G\bar{o}}} + \tilde{\epsilon}L$, to make it a metric in the mathematical sense). The Gō model and the EC model are very similar in the sense that they are both native-centric, based on contact maps, Eq. (1), and are of comparable complexity. However, the two different energies derived from the contact maps, Eqs. (2) and (5), cause the two models to display very different dynamical behaviors, as we will show in the next section. Put simply, this difference stems from the way the energies are constructed as (mainly) additive in the case of the Gō model where each contact contributes toward the total energy in the same way [except where additional contacts change the denominator in Eq. (2)] and inherently cooperative as in the EC model where many contacts have to be correct to give the correct connectivity which then leads to a cooperative folding process.

Before discussing results of the two models, we note that the usefulness of structural profiles such as the EC was also demonstrated in other contexts, for instance, for protein structure comparison [28], protein sequence comparison [29], and protein structure prediction [30]. All these applications indicate that structural profiles such as the EC constitute a "fingerprint" of the folded structure [27]. In the latter two examples mentioned, the structural profiles are not calculated from the native conformation but predicted from the amino acid sequence as a first step of the respective procedure, suggesting that structural profiles such as the EC are furthermore encoded in and reliably deducible from the amino acid sequence. Thus, using in Eq. (5) not the EC profile derived from the native structure (as we do in the following) but an EC profile predicted from the amino acid sequence might allow us in the future to derive some properties of the folding process from the sequence without the need of atomistic models or even knowledge of the folded structure.

III. RESULTS

Villin headpiece domains are among the shortest naturally occurring sequences which have been shown to autonomously fold, thereby displaying a surprisingly complex folding behavior (see, for instance, Ref. [31]). It is hence not surprising that these proteins have already been the focus of numerous experimental and theoretical investigations, which motivated us to focus here on this fold [more precisely on the variant Protein Data Bank (PDB) ID 1UND, chain A, shown in Fig. 2(a)]. Particularly relevant to the following discussion is that the complete folding process of a villin headpiece domain was recently studied by carrying out several microsecond-long molecular dynamics simulation runs with implicit water [5], which revealed an asymmetric folding behavior with one part of the protein structure typically forming before the other part. In a subsequent simulation study, a villin headpiece domain was investigated by molecular dynamics simulations

with explicit water [8] which agree with Ref. [5] on the asymmetry of folding pathways. Those simulations required vast computing resources, and it has hence been a formidable challenge to find similar results by using a coarse-grained model.

A. Free-energy landscapes and folding simulations

Following Ref. [5], we divide the native structure of the villin headpiece domain into two overlapping parts: Part A comprises the first two of the three α helices, from residue 1 to 21, and part B the last two of the three α helices, from residue 14 to 36, thus overlapping in the middle α helix. In topology-based models such as the ones studied here, the fraction of native contacts q_A and q_B of parts A and B are particularly well suited as reaction coordinates to measure the progress of folding of the respective parts. The villin headpiece domain is thus regarded as completely folded if all native contacts are present and $(q_A,q_B) = (1,1)$. To efficiently explore large regions of the free-energy landscape including those with high free energy, we use the well-tempered metadynamics method [32] and add the potential

$$V(s,t) = k_{\mathrm{B}}\Delta T \ln\left[1 + N(s,t)\right] \tag{7}$$

to the total energy of Eq. (3) or (6). The histogram $N(s,t)$ counts the number of times state s, here defined by the reaction coordinates (q_A,q_B), is visited. This histogram dependent potential energy avoids trapping in local minima (or indeed in the global minimum) and allows the simulated protein to visit large portions of conformation space. The parameter $\Delta T = 0.2\tilde{\epsilon}/k_{\mathrm{B}}$ for the Gō model and $\Delta T = 0.2\epsilon/k_{\mathrm{B}}$ for the EC model, respectively, is chosen such that the simulation does not spend too much time trapped in a minimum but also samples said minimum sufficiently before leaving. The equilibrium free energy of a state s is then determined as $F(s) = -k_{\mathrm{B}}(T + \Delta T)\ln N(s,t)\big|_{t\to\infty}$. The resulting free-energy landscapes, for both the Gō and the EC models, are shown in Fig. 3 for five different temperatures (six for the Gō model). Simulations were started from the folded protein structure and run for 2×10^8 Monte Carlo (MC) steps during which the protein unfolds and refolds multiple times. For the free-energy landscapes we averaged the final histograms of nine independent simulations. The Gō and EC models display qualitatively different behavior. In the Gō model, Figs. 3(a)–3(f), the free-energy landscape is almost featureless. It is largely funnelled toward the native state $(q_A,q_B) = (1,1)$ for low temperatures, and the minimum wanders continuously to larger q_A and q_B and grows ever deeper for increasing temperatures. Thus, no clear folding temperature can be defined in the Gō model. The villin headpiece has been experimentally observed to display two-state folding between the native (folded) and the denaturated (unfolded) state, so that one (or possibly more) non-native minimum should be visible in the free-energy landscape together with the native minimum at intermediate and larger temperatures. In the Gō model, a downhill folding occurs instead that is not limited by any free-energy barrier, which is consistent with the general observation that pairwise additive contact-based models tend to underestimate the free-energy barriers in protein folding [33].

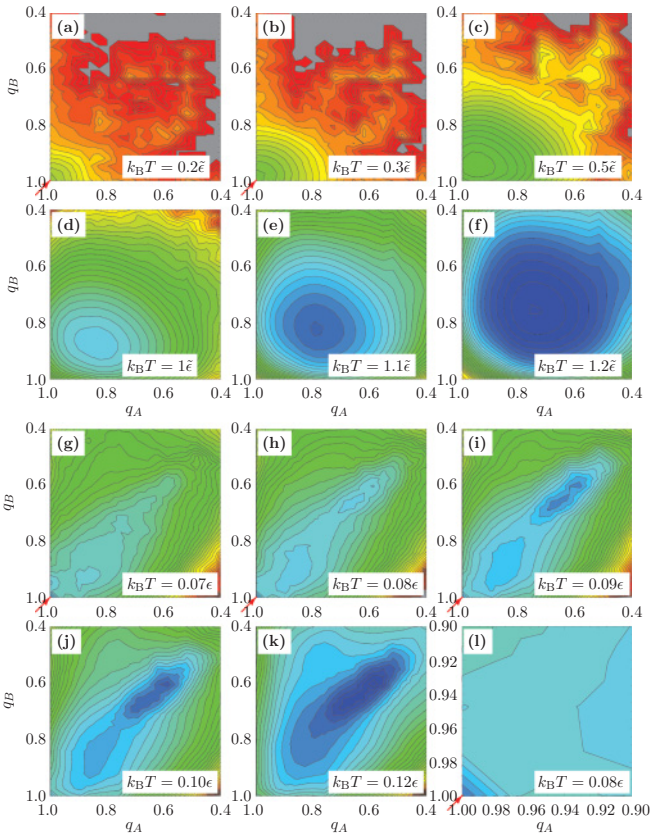

FIG. 3. (Color online) Comparison of the free-energy landscapes of the villin headpiece domain (PDB ID 1UND) in the reaction coordinates q_A and q_B, (a) to (f) in the Gō model and (g) to (l) in the EC model, at several temperatures. Colors range from blue (low free energy) via green and yellow to red (high free energy). (In gray scales this corresponds to dark, light, and dark again. Note that all local extrema are local minima, dark gray thus meaning low free energy.) A given color (gray value) indicates the same free energy for (a)–(f) and (g)–(l), respectively, and contour lines are drawn every (a)–(f) $0.5\tilde{\epsilon}$ and (g)–(l) 0.1ϵ. In the Gō model, the minimum wanders to larger q_A and q_B for increasing temperature and grows ever deeper, whereas in the EC model, a new free-energy minimum appears at larger q and drains the old native state minimum for increasing temperatures. (l) Detail of (h) showing the free-energy minimum. Red (gray) arrows indicate the native state where it is the global minimum of the free-energy landscape.

One of the most interesting features of the EC model is its ability to produce a cooperative folding transition without additional inputs [21,22], see Figs. 3(g)–3(l). The structural ensemble is mainly populating the native state at $(q_A,q_B) = (1,1)$ for temperatures below the folding temperature $k_{\mathrm{B}}T = 0.09\,\epsilon$ and a new free-energy minimum at $(q_A,q_B) \approx (0.6,0.6)$ appears that drains the native state minimum for increasing temperatures, becoming dominant for temperatures above the folding temperature. Thus, we observe a two-state folding process in the EC model. The equilibrium population of the native state at $(q_A,q_B) = (1,1)$ is 97% for $k_{\mathrm{B}}T = 0.07\,\epsilon$, 73% for $k_{\mathrm{B}}T = 0.08\,\epsilon$, 5% for $k_{\mathrm{B}}T = 0.09\,\epsilon$, 0.3% for $k_{\mathrm{B}}T = 0.1\,\epsilon$, and 0.0005% for $k_{\mathrm{B}}T = 0.12\,\epsilon$. For $k_{\mathrm{B}}T \leqslant 0.09\,\epsilon$ the native state is the free-energy minimum as indicated by the red (gray) arrows in Figs. 3(g)–3(l). Because the minimum

is very sharp and hard to discern, we magnify the region containing the native state for $k_B T = 0.08\,\epsilon$ in Fig. 3(l). At the folding temperature, $k_B T = 0.09\,\epsilon$, we find a barrier of $0.3\,\epsilon$ corresponding to $3.3\,k_B T$ which is toward the upper limit of the estimate from calorimetric data [34]. Furthermore, the free-energy landscape displays an asymmetry with an intermediate basin of partially folded structures centered at $(q_A,q_B) \approx (0.85,0.9)$. Because of this feature, the villin headpiece domain is approximately twice as likely to follow a folding route where part B becomes structured faster than part A (these trajectories would all lie below a diagonal line connecting the unfolded ensemble and the last barrier before the folded state). This type of folding behavior agrees with the one found in previous fully atomistic molecular dynamics simulations [5,8].

As alternative reaction coordinates, we use the helix content (measuring the number of amino acids identified as α helical according to the criterion described in Sec. II) and the total number of contacts. Figure 4 displays the free-energy landscape in these reaction coordinates for the Gō and EC models for one temperature each. In the Gō model, Fig. 4(a), one observes only one clear minimum centered at the native helix content 27 and the native number of contacts 62. There is a second very shallow minimum centered at a helix content of around 19 and the native number of contacts, which is, however, rather an inflection point than a minimum. An inspection of the configurations displaying these specific reaction coordinates reveals that they do not display alternative secondary structure; instead, the lower helix content is due only to the dissolution of the helices at their edges. A similar picture is observed for lower and higher temperatures. Folding simulations, started from an extended chain and run with standard Metropolis Monte Carlo for 4.5×10^7 MC steps, move directly into the native minimum in all cases and never leave it [see black trajectory projected onto the free-energy landscapes of Fig. 4(a)]. In the EC model, Fig. 4(b), one observes three pronounced minima at the temperature shown (which is slightly below the folding temperature): one centered at the native helix content 27 and the native number of contacts 62, and two others centered at the native helix content 27 but smaller than native number of contacts of around 40 and centered at the native number of contacts 62 but smaller than native helix content of around 19. Interestingly, an inspection of the configurations displaying reaction coordinates in the vicinity of the two non-native minima shows that the first non-native minimum hosts configurations displaying more or less the correct secondary structure but otherwise being rather extended, whereas the second minimum hosts configurations which are collapsed (and have approximately the correct number of contacts) some of which are clearly misfolded and display secondary structure elements reminiscent of β sheets. Folding simulations started from an extended chain conformation and run with standard Metropolis Monte Carlo for 4.5×10^7 MC steps visit all three free-energy minima, crossing the barriers in between, which means that the chain folds, unfolds, and misfolds continuously over time [see black trajectory projected onto the free-energy landscapes of Fig. 4(a)]. We ran nine such folding simulations unaided by the metadynamics method for the EC model and the temperature $k_B T = 0.08\,\epsilon$ displayed here out of which one simulation

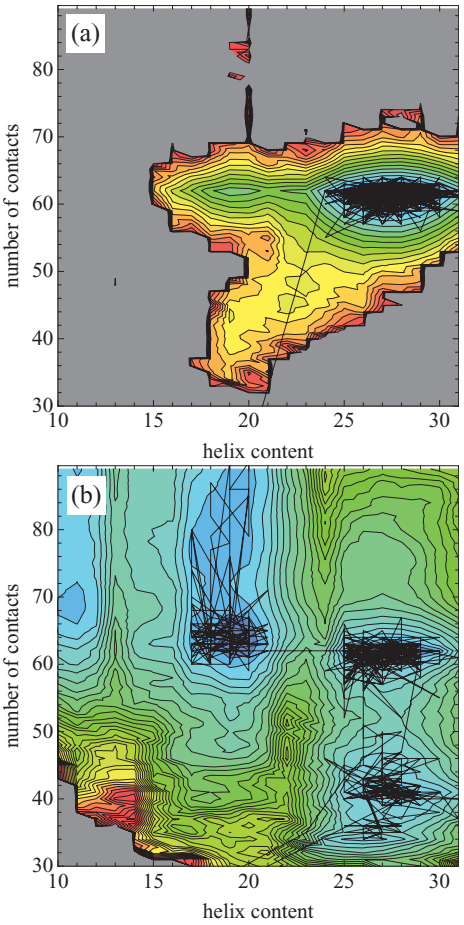

FIG. 4. (Color online) Free-energy landscape of villin headpiece domain (PDB ID 1UND) in the reaction coordinates helix content and total number of contacts, (a) in the Gō model at temperature $k_B T = 0.5\,\tilde{\epsilon}$ and (b) in the EC model at temperature $k_B T = 0.08\,\epsilon$. Colors range from blue (low free energy) via green and yellow to red (high free energy), and contour lines are drawn every (a) $0.5\tilde{\epsilon}$ and (b) 0.1ϵ. [In gray scales this corresponds to dark (low free energy), light (intermediate free energy), and dark again (high free energy). Note that all local extrema are local minima, dark gray thus meaning low free energy.] Several simulated folding trajectories are included for comparison.

was successful in reaching the exact native state, as defined by $(q_A,q_B) = (1,1)$, within the simulation time, and four simulations reached the free-energy minimum in the reaction coordinates of Fig. 4. Out of the nine folding simulations for the Gō model at $k_B T = 0.5\,\tilde{\epsilon}$, none reached $(q_A,q_B) = (1,1)$ but all visited structures of the same helix content and number of contacts as the native structure, which is, however, easily explained by the fact that the global free-energy minimum has already moved away from $(q_A,q_B) = (1,1)$ for this temperature [see Fig. 3(c)]. However, we chose to show and scrutinize this particular instance of a free-energy landscape for the Gō model because it displays the closest thing to a second minimum observed in any of the Gō model free-energy landscapes.

A more detailed analysis of an individual folding event is shown in Fig. 5, displaying the energy, the total number of contacts, and the helix content as a function of time. One notices that the helix content fluctuates around the correct

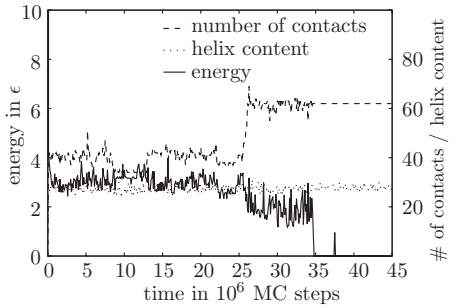

FIG. 5. Time evolution of number of contacts (dashed line), helix content (dotted line), and energy (solid line) for the villin headpiece domain (PDB ID 1UND) in the EC model at temperature $k_BT = 0.08\,\epsilon$.

value right from the start of the simulations with fluctuations becoming smaller for more compact structures. The first drop in energy (at about 25×10^6 MC steps) appears when the chain collapses to approximatively the correct number of contacts, and then another drop (at about 35×10^6 MC steps) when the chain finally folds into the correct configuration.

B. Heat capacities and folding transitions

To elucidate the two-state behavior of the EC model suggested by Figs. 3(g)–3(l) and the associated folding transition, the heat capacity at constant volume $C_V = (\langle E^2\rangle - \langle E\rangle^2)/T^2$ obtained by standard Metropolis Monte Carlo simulations is shown in Fig. 6 for the Gō and EC models. There is no peak visible for the Gō model, and energy fluctuations due to partial and noncooperative unfolding occur even at very low temperatures, in accordance with the free-energy landscapes in Figs. 3(a)–3(f). This behavior contrasts that of the EC model, for which a pronounced peak at $k_BT \approx 0.1\,\epsilon$ is observed, a value which agrees well with the folding temperature derived from free-energy landscapes (two coexisting minima) in Figs. 3(g)–3(l).

The energy distributions from which the heat capacity has been calculated are exemplified for three different temperatures for the Gō and EC models in Fig. 7 (the bin size for the Gō model is larger because of the discreteness of $E_{G\bar{o}}$). In the Gō model, the distribution is unimodal at all temperatures [we observe the closest thing to a second maximum at $k_BT = 0.09\,\tilde{\epsilon}$, Fig. 7(a), but only for a single bin at a single temperature], and the maximum shifts to

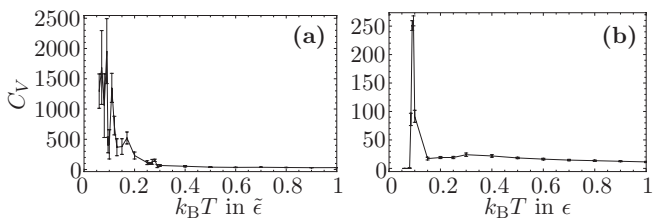

FIG. 6. Heat capacity curves for the villin headpiece domain (PDB ID 1UND) in (a) the Gō model and (b) the EC model. In the Gō model, the heat capacity simply rises for decreasing temperature while error bars increase as well, whereas in the EC model, a sharp heat capacity peak at a temperature of $k_BT \approx 0.1\,\epsilon$ is observed.

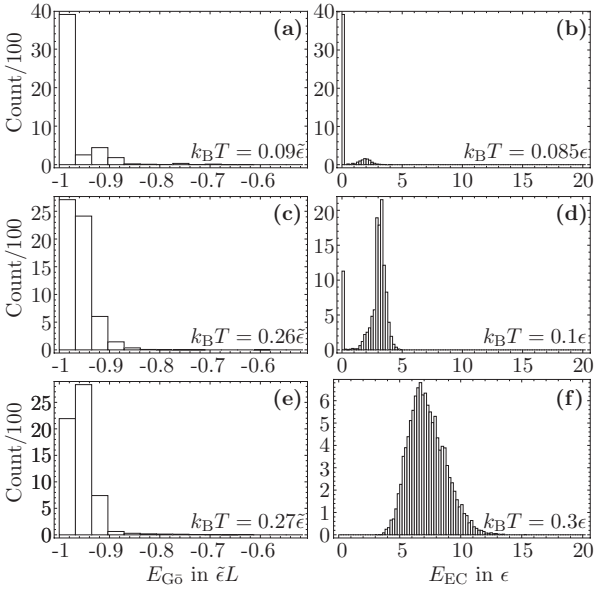

FIG. 7. Energy distribution for the villin headpiece domain (PDB ID 1UND) in (a),(c), and(e) the Gō model and (b),(d), and (f) the EC model at several temperatures (the bin size for the Gō model is larger because of the discreteness of $E_{G\bar{o}}$). In the Gō model, the distribution is unimodal at all temperatures, a weak signal corresponding to a second maximum appears at a single temperature $k_BT = 0.09\,\tilde{\epsilon}$ (and only for a single bin), and the maximum shifts to higher energies for higher temperatures. In the EC model, (b) a competing local maximum to the native energy first appears at $k_BT = 0.085\,\epsilon$ and (d) overtakes the first energy peak at $k_BT = 0.1\,\epsilon$. (f) At still higher temperatures the energy distribution shifts to higher average energies and broadens.

higher energies for higher temperatures. This is in contrast to the EC model, for which two clear and separated peaks are visible, one for the native energy and one for higher energies, and the second peak grows at the expense of the first (native) peak, causing the first peak to shrink and eventually to disappear.

Summing up the results for the villin headpiece domain in the two different models investigated here, we have found free-energy landscapes displaying two different states in the EC model as compared to single-state landscapes for the Gō model. Our choice of reaction coordinates in Fig. 3 allows a direct comparison with the results of Ref. [5] where the authors employed the same coordinates. Whereas the Gō model free-energy landscapes only show a single minimum at all temperatures, the EC model shows several minima (namely, native, unfolded, and one intermediate). This is strongly reminiscent of the results in [5] with minima corresponding to native, unfolded, and two intermediate states of which one intermediate state, however, was much less populated and also off-pathway in the folding process. We presume that this intermediate state cannot be seen in our coarse-grained simulations. The position of our intermediate state also agrees with their on-pathway intermediate which corresponds to part *B* of the protein becoming structured before part *A*. The asymmetry of our free-energy landscapes is thus in good agreement with the detailed molecular dynamics (MD) simulations of Refs. [5] and [8]. Reference [5] also reports

results for the heat capacity of the villin headpiece. However, without calibration of our energy scale, the width and height of the heat capacity peak are difficult to compare. The existence of a clear peak in the EC model though agrees with those findings (and the established two-state folding of the villin headpiece domain) whereas the Gō model does not exhibit such a peak. The local free-energy minimum of correct helix content but too few contacts in Fig. 4(b) agrees with the experimental observation of the secondary structure in the denatured state of the villin headpiece domain [31], and the height of the free-energy barrier in the EC model [Fig. 3(h)] is compatible with the estimates made in Ref. [34] from calorimetric data. Taken together, our results indicate that the EC model shows very good agreement with experimental and MD numerical results for a coarse-grained and native-centric protein model.

IV. CONCLUSIONS

We have recently proposed a coarse-grained model of proteins that enables the efficient study of folding trajectories and free-energy landscapes of folding [20–22]. We have compared this model here with a more standard Gō model of comparable complexity, using the villin headpiece domain as an example. For both models, we have analyzed free-energy landscapes in two different sets of reaction coordinates, properties of folding trajectories, and heat capacities. The asymmetry in the free-energy landscape of the villin headpiece domain observed in the EC model but not in the Gō model agrees well with the folding behavior found in extensive fully atomistic molecular dynamics studies. Furthermore, the two-state folding behavior observed experimentally for the villin headpiece is reproduced by the EC model but not by the Gō model. We suggest that the use of coarse-grained models that combine tubelike geometries with connectivity-based energy functions represents an effective tool for the efficient characterization of the folding behavior of globular proteins.

ACKNOWLEDGMENTS

K.W. gratefully acknowledges funding by the Deutscher Akademischer Austauschdienst, M.P. acknowledges funding from the Deutsche Forschungsgemeinschaft via the Heisenberg program (PO 1025/6), and all three authors acknowledge financial support under a travel grant by the Deutscher Akademischer Austauschdienst, Grant No. D/08/08872, and the British Council, Grant No. ARC 1319, in early stages of this work.

[1] R. Zwanzig, A. Szabo, and B. Bagchi, Proc. Natl. Acad. Sci. USA **89**, 20 (1992).
[2] P. Wolynes, J. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[3] K. A. Dill and H. S. Chan, Nat. Struct. Mol. Biol. **4**, 10 (1997).
[4] C. M. Dobson, A. Sali, and M. Karplus, Angew. Chem., Int. Ed. **37**, 868 (1998).
[5] H. Lei, C. Wu, H. Liu, and Y. Duan, Proc. Natl. Acad. Sci. USA **104**, 4925 (2007).
[6] D. L. Ensign and V. S. Pande, Biophys. J. **96**, L53 (2009).
[7] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Science **330**, 341 (2010).
[8] T. Yoda, Y. Sugita, and Y. Okamoto, Biophys. J. **99**, 1637 (2010).
[9] M. Vendruscolo and C. M. Dobson, Curr. Biol. **21**, R68 (2011).
[10] R. D. Hills, Jr., and C. L. Brooks III, Int. J. Mol. Sci. **10**, 889 (2009).
[11] V. Tozzini, Acc. Chem. Res. **43**, 220 (2010).
[12] R. D. Hills, Jr., L. Lu, and G. A. Voth, PLoS Comput. Biol. **6**, e1000827 (2010).
[13] M. Vendruscolo, E. Kussel, and E. Domany, Fold. Des. **2**, 295 (1997).
[14] H. Taketomi, Y. Ueda, and N. Gō, Int. J. Pept. Prot. Res. **7**, 445 (1975).
[15] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[16] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[17] P. Das, C. J. Wilson, G. Fossati, P. Wittung-Stafshede, K. S. Matthews, and C. Clementi, Proc. Natl. Acad. Sci. USA **102**, 14569 (2005).
[18] B. T. Andrews, S. Gosavi, J. M. Finke, J. N. Onuchic, and P. A. Jennings, Proc. Natl. Acad. Sci. USA **105**, 12283 (2008).
[19] A. Kleiner and E. Shakhnovich, Biophys. J. **92**, 2054 (2007).
[20] K. Wolff, M. Vendruscolo, and M. Porto, PMC Biophys. **1**, 5 (2008).
[21] K. Wolff, Ph.D. thesis, Technischen Universität Darmstadt, 2010, available on [http://tuprints.ulb.tu-darmstadt.de/2068/].
[22] K. Wolff, M. Vendruscolo, and M. Porto, EPL **94**, 48005 (2011).
[23] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. USA **101**, 7960 (2004).
[24] S. Auer, C. M. Dobson, and M. Vendruscolo, HFSP J. **1**, 137 (2007).
[25] U. Bastolla, A. R. Ortiz, M. Porto, and F. Teichert, Proteins **73**, 872 (2008).
[26] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo, Proteins **58**, 22 (2005).
[27] M. Porto, U. Bastolla, H. E. Roman, and M. Vendruscolo, Phys. Rev. Lett. **92**, 218101 (2004).
[28] F. Teichert, U. Bastolla, and M. Porto, BMC Bioinformatics **8**, 425 (2007).
[29] F. Teichert, J. Minning, U. Bastolla, and M. Porto, BMC Bioinformatics **11**, 251 (2010).
[30] K. Wolff, M. Vendruscolo, and M. Porto, Proteins **78**, 249 (2010).
[31] Y. Tang, D. J. Rigotti, R. Fairman, and D. P. Raleigh, Biochemistry **43**, 3264 (2004).
[32] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).
[33] M. P. Eastwood and P. G. Wolynes, J. Chem. Phys. **114**, 4702 (2001).
[34] R. Godoy-Ruiz, E. R. Henry, J. Kubelka, J. Hofrichter, V. Munoz, J. M. Sanchez-Ruiz, and W. A. Eaton, J. Phys. Chem. B **112**, 5938 (2008).