A Letters Journal Exploring the Frontiers of Physics

May 2011

EPL, **94** (2011) 48005
doi: 10.1209/0295-5075/94/48005

www.epljournal.org

Asymmetric folding pathways and transient misfolding in a coarse-grained model of proteins

K. Wolff[1(a)], M. Vendruscolo[2] and M. Porto[3]

[1] *School of Physics, University of Edinburgh - JCMB Kings Buildings, Edinburgh EH9 3JZ, UK, EU*
[2] *Department of Chemistry, University of Cambridge - Lensfield Road, Cambridge CB2 1EW, UK, EU*
[3] *Institut für Theoretische Physik, Universität zu Köln - Zülpicher Str. 77, 50937 Köln, Germany, EU*

received 9 February 2011; accepted in final form 11 April 2011
published online 6 May 2011

PACS 87.14.E- – Proteins
PACS 87.15.A- – Theory, modeling, and computer simulation
PACS 87.15.Cc – Folding: thermodynamics, statistical mechanics, models, and pathways

Abstract – Coarse-grained approaches to study the protein folding process provide the possibility to explore timescales longer than those accessible to all-atom models and thus provide access, albeit in less detail, to larger regions of the conformational space. Here, we investigate the behaviour of a coarse-grained model whose two primary characteristics are a tube-like geometry to describe the self-avoidance effects of the polypeptide chain, and an energy function based on a one-dimensional structural representation that specifies the sequence's connectivity in a given conformation. Such an energy function, rather than favouring the formation of specific native pairwise contacts, promotes the establishment of a specific target connectivity for each amino acid. We illustrate the use of this model by showing that it enables to follow the complete process of folding and to efficiently determine the free energy landscapes of two small α-helical proteins, the villin headpiece domain and the ubiquitin associated domain, providing results that closely resemble those found in extensive molecular dynamics studies. These results support the idea that the use of coarse-grained models that capture the self-avoidance and the connectivity of a polypeptide chain represents a promising approach for obtaining effective descriptions of many aspects of the behaviour of proteins.

Introduction. – In recent years, great advances have been made in the use of atomistic models to study protein folding [1–5]. Despite these very significant improvements, however, simulations at atomistic detail still require great computational efforts, thus making a complete description of the protein folding process one of the open grand challenges in science. Coarse-grained models provide a range of opportunities to investigate the behaviour of proteins at a lower level of detail [6–8]. They may thus help answer more generic questions concerning protein folding and provide a qualitative understanding of the fundamental principles that govern this process. One such question is how the native structure restricts the possible folding pathways [9,10]. At the geometrical level, it is well established that it is sufficient to know which amino acid pairs are spatially close (in contact) to be able to recover protein native structures with relatively high accuracy [11]. These observations prompt the question whether the information provided in the native contacts is also sufficient to infer major features of the protein folding process. In this context, a prominent role has been played by Gō-models [9,12], a particularly well-studied and versatile class of coarse-grained protein models, which is characterised by an energy function that favours the formation of native contacts between amino acids. These models embody the principle of minimal frustration [13], which assumes that proteins have been evolutionarily optimised for fast and reliable folding, and have been applied to relatively large proteins to capture many features of their folding process [9,14,15].

Models based solely on native contact energy contributions are, however, not designed for reproducing phenomena such as misfolding, precisely because of their smooth energy landscape funnelled towards the native state [9]. To overcome these limitations such models are therefore often used in conjunction with additional energy terms to account for, *e.g.*, hydrogen bonds or physico-chemical details of amino acids [6–8,16]. In this letter we propose

(a) E-mail: katrin.wolff@ed.ac.uk

another strategy to address the question of whether the native topology determines the folding behaviour by investigating the properties of a model whose ground state is determined by the native state, but which does not rely on an energy function favouring the formation of specific native pairwise contacts.

Model. – Our model is based on the notion of effective connectivity of an amino acid, which is calculated from the contact map corresponding to a given conformation [17]. The effective connectivity profile (EC) is a vector quantity of dimension L specifying the effective connectivity for each of the L amino acids comprising a given polypeptide chain. Of particular relevance in this discussion is the effective connectivity in the native state. The native EC does not specify which individual contacts between amino acid pairs should be formed, but is related to the total number of contacts that each amino acid has in the native state, which is in turn closely associated to this amino acid's hydrophobicity [17,18]. The EC represents an analytic solution of the "statistical inverse folding problem", which consists in finding the statistical properties of sequences compatible with a given structure. As such it allows 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 [17].

The contact map is defined as

$$C_{ij} = \begin{cases} 1, & d_{i,j} < r_c \wedge |i-j| > 2, \\ 0, & d_{i,j} \geqslant r_c \vee |i-j| \leqslant 2, \end{cases} \tag{1}$$

where $d_{i,j}$ is the distance between C_α-atom i and C_α-atom j and r_c is a cut-off distance for contacts, which in the present case is set to 8.5 Å. Contacts that are present all the time, such as self-contacts and any contact with $|i-j| \leqslant 2$, are disregarded. From this the EC $\mathbf{c}$ is computed according to

$$\mathbf{c} = \frac{1}{A} \sum_{k=1}^{L} \frac{1}{\Lambda - \lambda^{(k)}} \mathbf{v}^{(k)} \langle v^{(k)} \rangle. \tag{2}$$

Here L is the number of amino acids in the protein, $\mathbf{v}^{(k)}$ $(k = 1, \ldots, L)$ are the L eigenvectors of the contact map 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 $\Lambda > \lambda^{(k)}$ $\forall k$ are used to set the average and variance of $\mathbf{c}$ following ref. [17]. The EC 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 EC and PE is high for fully folded globular proteins (see ref. [17]). We restrict the EC, for both the target (native) and the current conformation, to residues showing cooperative contact patterns reminiscent of secondary structure [19,20].

The total energy then consists of two terms

$$E_{\text{tot}} = E_{\text{steric}} + E_{\text{EC}}, \tag{3}$$

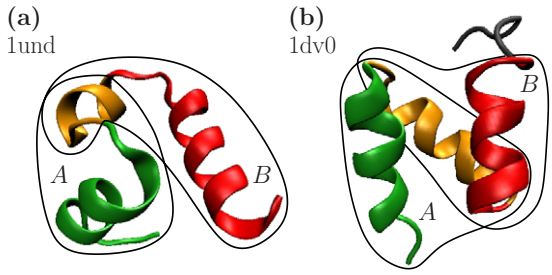

Fig. 1: (Colour online) Native structures of the two proteins studied in this work: (a) villin headpiece domain (PDB ID 1und); (b) ubiquitin associated domain (PDB ID 1dv0). In both cases, part A consists of the two N-terminal α-helices, and part B of the two C-terminal α-helices. In the ubiquitin associated domain, the disordered tail (grey) is disregarded. Images were created using VMD [21].

where E_{steric} refers to the underlying description of the protein's chain of C_α-atoms as a tube with a finite thickness [22,23]. This energy term accounts for excluded volume of the tube by prohibiting overlapping parts and bending rigidity by disallowing too tight angles by a very high energy penalty. In the implementation of the tube and in the moveset used in Monte Carlo simulations (pivot and local crankshaft moves with Gaussian distribution of angles) we strictly follow refs. [20,23]. The term E_{EC} in eq. (3) is the bias towards the EC of the native structure

$$E_{\text{EC}} = \epsilon \sum_{i=1}^{L} |t_i - c_i|, \tag{4}$$

and sums up absolute differences over all amino acids i between EC profiles in the native conformation $\mathbf{t}$ and the current conformation $\mathbf{c}$. The parameter ϵ sets the unit of energy used throughout this work. Apart from this parameter, which only sets the scale for temperatures used in simulations, and the cut-off radius r_c, to which results are largely insensitive, our model is free from any parameters to tune interaction strengths.

It has been shown that with the EC model it is possible to reliably fold small proteins from unfolded states to native conformations [19], and the usefulness of connectivity profiles was also demonstrated in the context of protein structure comparison [24]. Most importantly, the EC model was shown to help protein structure prediction [25]; in that case the EC profiles are not calculated from the native conformation but predicted from the amino acid sequence as a first step of the procedure. Those latter results indicate that structural profiles such as the EC constitute a "fingerprint" of the folded structure [26] being encoded in and reliably deducible from the amino acid sequence. Thus, using in eq. (4) not the EC profile derived from the native structure (as we do in this letter) but an EC profile predicted from the amino acid sequence might allow in the future to derive properties of the folding process from sequence without the need of atomistic models.

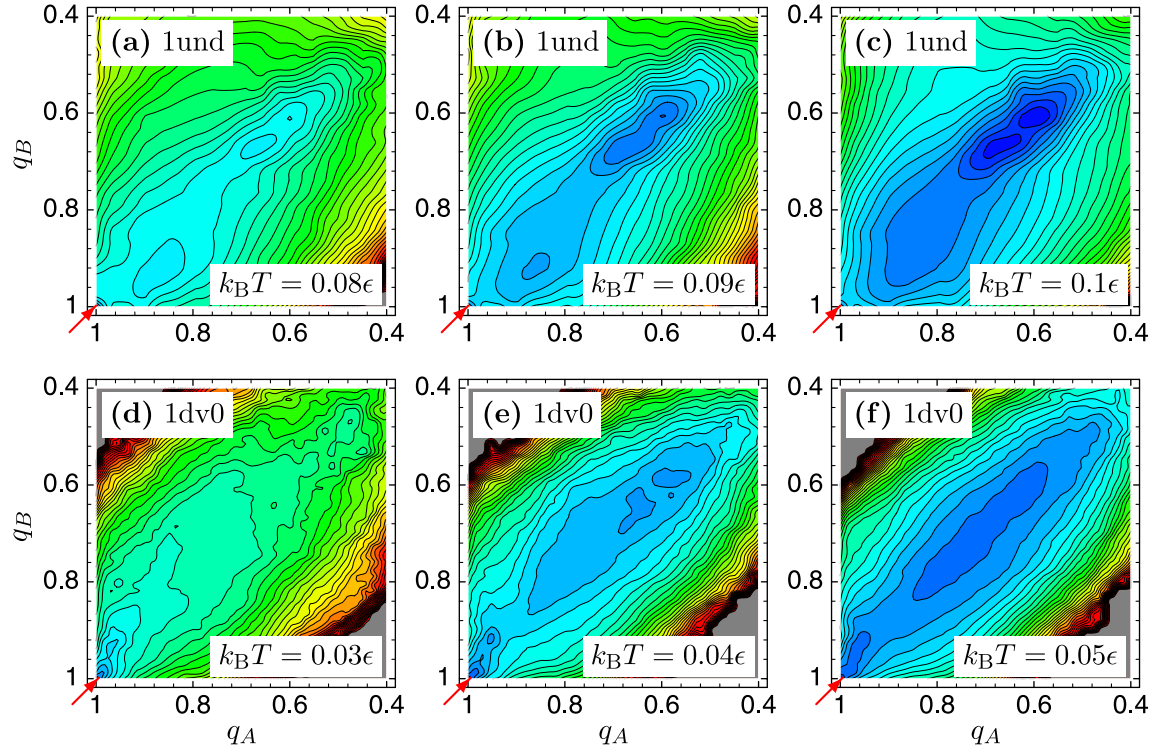

Fig. 2: (Colour online) Free energy landscapes at increasing temperatures in the reaction coordinates q_A and q_B: (a)–(c) villin headpiece domain; (d)–(f) ubiquitin associated domain. Regions of low free energy are shown in blue, regions of high free energy in red. Contour lines are spaced at 0.1ϵ. Grey regions have not been visited during the time of simulation. The red arrow points at the minimum representing the native state $(q_A, q_B) = (1, 1)$.

Results. – One of the most interesting features of the EC model is its ability to produce a cooperative folding transition without additional inputs [20], whereas pairwise additive contact-based models tend to underestimate the free energy barriers in protein folding [27]. As we show below, the EC model is able to reproduce features of the free energy landscape that have been found so far in much more computationally demanding fully atomistic simulations [1]. In the latter study, the complete folding process of the villin headpiece domain was calculated by carrying out several microsecond-long molecular-dynamics simulation runs, which revealed an asymmetric folding behaviour with one part of the protein structure typically forming before the other part. The villin headpiece is one of the shortest naturally occurring sequences which has been shown to autonomously fold, thereby displaying a surprisingly complex folding behaviour (see, for instance, ref. [28]). In a recent simulation study the villin headpiece domain was investigated using explicit water molecular dynamics simulations [4] which agree with ref. [1] on the asymmetry of folding pathways. In this study we consider the formidable challenge of finding similar results by using the coarse-grained model described here.

Following ref. [1], we divide the native structure of the villin headpiece domain (PDB ID 1und) into two overlapping parts (fig. 1(a)). Part A comprises the first two α-helices, from residue 1 to 21, and part B the last two α-helices, from residue 14 to 36. In a topology-based model such as ours, the fraction of native contacts q_A and q_B of part A and B are particularly well suited as reaction coordinates to measure the progress of folding of the respective parts. The protein is thus regarded as completely folded if all native contacts are present and $q_A = q_B = 1$.

To efficiently explore large regions of the free energy landscape including those with high free energy, we use the well-tempered metadynamics method [29] with parameters $\Delta T = \omega = 0.2\epsilon/k_B$. We thus calculated the free energy landscapes for the villin headpiece domain at different temperatures (fig. 2(a)–(c)), just below, roughly at, and just above the folding temperature. The native state at $(q_A, q_B) = (1, 1)$ (red arrow) is the global minimum for $k_BT = 0.08\epsilon$ whereas it is of equal depth as the unfolded minimum at $(q_A, q_B) \approx (0.6, 0.6)$ for $k_BT = 0.09\epsilon$. Below and at the folding temperature $k_BT = 0.09\epsilon$, the landscape displays an asymmetry with an intermediate basin of partially folded structures centred at $(q_A, q_B) \approx (0.86, 0.92)$. Because of this feature in the free energy landscape, 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 behaviour agrees with the general picture found in previous fully atomistic molecular-dynamics simulations [1].

In order to elucidate whether this may be a generic behaviour of proteins consisting of three α-helices, or one specific to the villin headpiece domain, we investigated another three-α-helix protein, the C-terminal ubiquitin associated (UBA) domain (PDB ID 1dv0), which has a different tertiary structure. This protein consists of 47 residues of which we discarded the last 7 as they are unstructured in the native state (see fig. 1(b)). Part A again comprises the first two helices, from residue 1 to 28, and part B the last two helices, from residue 14 to 40.

According to our results, the folding dynamics and the free energy landscape of the UBA domain (fig. 2(d)–(f)) differ from those of the villin headpiece domain in some important aspects. In contrast to the villin headpiece domain (fig. 2(a)), the UBA domain exhibits an essentially symmetric free energy landscape also at low temperatures, *i.e.* under folding conditions. Although the near-native region is not perfectly symmetric, the pathway leading from the unfolded ensemble to the free energy barrier, which represents the bottleneck of the folding process, is indeed almost completely symmetric (fig. 2(e)). Folding trajectories thus stay close to the diagonal for most of the time and both parts fold simultaneously with almost equal probability of straying to either side. Furthermore, there is no distinct off-diagonal minimum corresponding to intermediate states as for the villin headpiece domain.

What is clearly visible is that, as the temperature is increased, a new minimum at unfolded configurations emerges, increasing in weight and moving towards ever less folded structures. We thus find two-state behaviour and cooperative folding for both proteins but with details in the free energy landscapes that are specific to the protein in question. At the folding temperature of the villin headpiece domain, $k_{\mathrm{B}}T = 0.09\epsilon$, we find a barrier of 0.3ϵ corresponding to $3.3\,k_{\mathrm{B}}T$ which is towards the upper limit of the estimate from calorimetric data [30]. For the UBA domain at the folding temperature $k_{\mathrm{B}}T = 0.04\epsilon$ we find a barrier of 0.2ϵ corresponding to $5\,k_{\mathrm{B}}T$ thus predicting a higher free energy barrier than for the villin headpiece domain.

We also analysed the unfolded ensembles of both proteins in detail and found some distinctive features. These differences are more readily observed in a different projection of the conformation space, namely in free energy landscapes plotted as a function of the α-helical content and the total number of contacts as reaction coordinates (see fig. 3). While for the villin headpiece domain we observed misfolded conformations with significantly less α-helical content than the native state (fig. 3(a), yellow arrow), which in fact are mostly due to a dissolved N-terminal helix, this is not the case for the UBA domain, for which misfolded conformations are predominantly

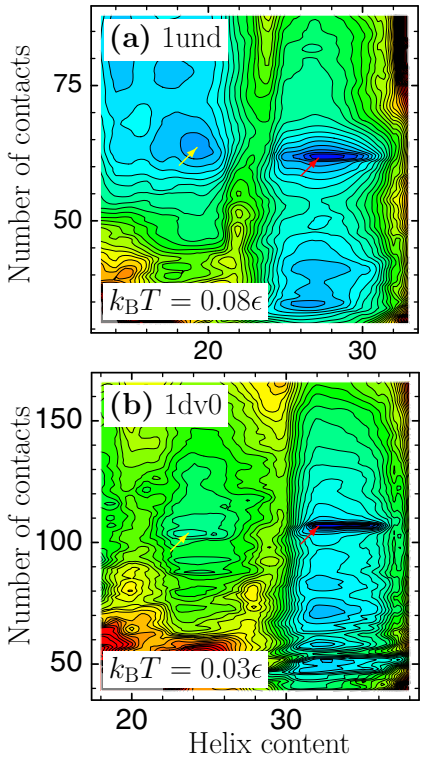

Fig. 3: (Colour online) Free energy landscapes as a function of two reaction coordinates alternative to those used in fig. 2 —the α-helical content and the total number of contacts: (a) villin headpiece domain at $k_{\mathrm{B}}T = 0.08\epsilon$, and (b) ubiquitin associated domain at $k_{\mathrm{B}}T = 0.03\epsilon$. Regions of low free energy are shown in blue, and regions of high free energy in red. Contour lines are spaced at 0.1ϵ. The red arrow points at the native state, which lies at (27,62) for the villin headpiece domain, and at (32,107) for the ubiquitin associated domain.

α-helical but have non-native topologies (fig. 3(b)) and regions of low α-helical content are rather high-lying.

Conclusions. – We have presented a coarse-grained model of proteins that enables the efficient study of free energy landscapes of folding. We have illustrated the model by comparing the free energy landscapes of two small α-helical proteins, the villin headpiece domain and the ubiquitin associated domain. The asymmetry in the free energy landscape of the villin headpiece domain agrees well with folding behaviour found in extensive fully atomistic molecular dynamics studies. By contrast, we found a symmetric free energy landscape and a higher free energy barrier for the ubiquitin associated domain. Although in the present study we have focused on α-helical proteins, we have previously shown [19,20] that our model enables to describe the folding process also of β-sheet proteins. Studies with larger proteins with more complex topologies are currently under way. Taken together with the results that we have presented here, we suggest that the use of coarse-grained models that combine tube-like geometries with connectivity-based energy functions represent an effective tool for the

efficient characterisation of the folding behaviour of globular proteins.

Our model offers the possibility to study misfolded structures of proteins, which is otherwise often problematic within Go-like models. Another promising direction for further research stems from the connection between sequence hydrophobicity and EC. In future studies it will be worthwhile to study which features of the free energy landscapes are robust enough to survive the use of noisy and eventually predicted EC profiles as targets.

* * *

We gratefully acknowledge funding by the *Deutscher Akademischer Austauschdienst* (KW), by the *Deutsche Forschungsgemeinschaft* via the Heisenberg program (PO 1025/6) (MP), and financial support under a travel grant by the *Deutscher Akademischer Austauschdienst*, grant No. D/08/08872, and *The British Council*, grant No. ARC 1319.

REFERENCES

[1] Lei H., Wu C., Liu H. and Duan Y., *Proc. Natl. Acad. Sci. U.S.A.*, **104** (2007) 4925.
[2] Ensign D. L. and Pande V. S., *Biophys. J.*, **96** (2009) L53.
[3] Shaw D. E., Maragakis P., Lindorff-Larsen K., Piana S., Dror R. O., Eastwood M. P., Bank J. A., Jumper J. M., Salmon J. K., Shan Y. and Wriggers W., *Science*, **330** (2010) 341.
[4] Yoda T., Sugita Y. and Okamoto Y., *Biophys. J.*, **99** (2010) 1637.
[5] Vendruscolo M. and Dobson C. M., *Curr. Biol.*, **21** (2011) R68.
[6] Hills R. D. jr. and Brooks C. L. III, *Int. J. Mol. Sci.*, **10** (2009) 445.
[7] Tozzini V., *Acc. Chem. Res.*, **43** (2010) 220.
[8] Hills R. D. jr., Lu L. and Voth G. A., *PLoS Comput. Biol.*, **6** (2010) e1000827.
[9] Clementi C., Nymeyer H. and Onuchic J. N., *J. Mol. Biol.*, **298** (2000) 937.
[10] Baker D., *Nature*, **405** (2000) 39.
[11] Vendruscolo M., Kussel E. and Domany E., *Fold. Des.*, **2** (1997) 295.
[12] Taketomi H., Ueda Y. and Gō N., *Int. J. Pept. Protein Res.*, **7** (1975) 445.
[13] Bryngelson J. D., Onuchic J. N., Socci N. D. and Wolynes P. G., *Proteins*, **21** (1995) 167.
[14] Das P., Wilson C. J., Fossati G., Wittung-Stafshede P., Matthews K. S. and Clementi C., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 14569.
[15] Andrews B. T., Gosavi S., Finke J. M., Onuchic J. N. and Jennings P. A., *Proc. Natl. Acad. Sci. U.S.A.*, **105** (2008) 12283.
[16] Kleiner A. and Shakhnovich E., *Biophys. J.*, **92** (2007) 2054.
[17] Bastolla U., Ortiz A. R., Porto M. and Teichert F., *Proteins*, **73** (2008) 872.
[18] Bastolla U., Porto M., Roman H. E. and Vendruscolo M., *Proteins*, **58** (2005) 22.
[19] Wolff K., Vendruscolo M. and Porto M., *PMC Biophys.*, **1** (2008) 5.
[20] Wolff K., *Protein structure prediction and folding dynamics* (Doctoral thesis) 2010, available on `http://tuprints.ulb.tu-darmstadt.de/2068/`.
[21] Humphrey W., Dalke A. and Schulten K., *J. Mol. Graph.*, **14** (1996) 33.
[22] Hoang T. X., Trovato A., Seno F., Banavar J. R. and Maritan A., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 7960.
[23] Auer S., Dobson C. M. and Vendruscolo M., *HFSP J.*, **1** (2007) 137.
[24] Teichert F., Bastolla U. and Porto M., *BMC Bioinformatics*, **8** (2007) 425.
[25] Wolff K., Vendruscolo M. and Porto M., *Proteins*, **78** (2010) 249.
[26] Porto M., Bastolla U., Roman H. E. and Vendruscolo M., *Phys. Rev. Lett.*, **92** (2004) 218101.
[27] Eastwood M. P. and Wolynes P. G., *J. Chem. Phys.*, **114** (2001) 4702.
[28] Tang Y., Rigotti D. J., Fairman R. and Raleigh D. P., *Biochemistry*, **43** (2004) 3264.
[29] Barducci A., Bussi G. and Parrinello M., *Phys. Rev. Lett.*, **100** (2008) 020603.
[30] Godoy-Ruiz R., Henry E. R., Kubelka J., Hofrichter J., Munoz V., Sanchez-Ruiz J. M. and Eaton W. A., *J. Phys. Chem. B*, **112** (2008) 5938.