A simple hydrophobicity-based score for profiling protein structures

2005 J. Phys.: Condens. Matter 17 S1595

(http://iopscience.iop.org/0953-8984/17/18/015)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **17** (2005) S1595–S1606 doi:10.1088/0953-8984/17/18/015

A simple hydrophobicity-based score for profiling protein structures

Nelson A Alves[1], Vasyl Aleksenko[2] and Ulrich H E Hansmann[2]

[1] Departamento de Física e Matemática, FFCLRP Universidade de São Paulo, Avenida Bandeirantes 3900. CEP 14040-901 Ribeirão Preto, SP, Brazil
[2] Department of Physics, Michigan Technological University, Houghton, MI 49931-1295, USA

E-mail: alves@ffclrp.usp.br, valeksen@mtu.edu and hansmann@mtu.edu

Received 21 September 2004, in final form 21 September 2004
Published 22 April 2005
Online at stacks.iop.org/JPhysCM/17/S1595

Abstract

We propose a simple measure that allows the profiling of protein configurations. It is based on calculation of a restricted radius of gyration evaluated only between the centroids of hydrophobic residues and measures the formation and compactness of the hydrophobic core. Some preliminary results for applications of the new score in generalized-ensemble simulations are presented.

1. Introduction

One aspect of the protein folding problem is the prediction of the biologically active structure of a protein given only its chemical composition (the sequence of amino acids as encoded in the genome). The majority of globular proteins is thermodynamically stable and the folded state is the global minimum of the free energy landscape [1–5]. As its entropy is comparatively small, it is also the global minimum in the potential energy. However, present energy functions are not always accurate enough to ensure that the native structure indeed corresponds to the global minimum [6–8]. In addition, all-atom models of proteins are often plagued by spurious minima that complicate the search process. The latter problem can be circumvented by slightly changing the question. Instead of predicting the biologically active structure the task becomes now to identify it out of an ensemble of competing low-energy structures (so-called decoys).

The question arises of whether one can define a score other than the energy that discriminates between the native structure and other conformations. Search for such scoring functions is an active field of research [8–13]. Some are built from first principles; other are knowledge based, taking into account statistics from the set of known three-dimensional structures in the Protein Data Bank. In most cases these scores attempt to model some fundamental characteristics of folded proteins. For instance, it is believed that the hydrophobic effect is the main driving force towards the final shape of globular proteins [14, 15]. In the process of folding, amino acids that have hydrophobic side chains will aggregate in order to minimize their exposure to water. On the other hand, hydrophilic residues will prefer the

exterior of the proteins as it increases their chances of forming hydrogen bonds with the water molecules. Single-amino-acid replacement experiments suggest that the stability of the native structure is due to the so-formed hydrophobic core [16] with only marginal contributions by hydrophilic residues ([17] and references therein). Hence, the spatial distribution of hydrophobic and hydrophilic residues in a protein structure, its hydrophobic profile [18, 19], may offer a way to construct scoring functions.

One example is the score by Silverman *et al* [19–21] that allows a spatial profiling of the transition from the hydrophobic core to the hydrophilic exterior of globular proteins and was used to detect native protein foldings among decoy structures [11]. Here, we propose and test a score that also traces the formation of a hydrophobic core but is simpler to evaluate. It is based on calculation of a restricted radius of gyration evaluated only between the atoms of hydrophobic residues. Focusing on a set of small proteins we demonstrate that the new score indeed allows identification of the native structure out of a set of decoys.

2. Methods

In globular proteins, the native state is compact and exhibits a hydrophobic core. The compactness of protein configurations is often described by the radius of gyration

$$r_{\mathrm{g}}^2 = \frac{1}{2n^2}\sum_{i,j}^{n} r_{ij}^2, \qquad r_{ij} = |\vec{r}_i - \vec{r}_j|, \tag{1}$$

where the sum goes over all n amino acids of the protein, and $\vec{r}_i$ is the position vector of the centre of geometry for heavy atoms of the ith residue. Restricting the evaluation of this expression to hydrophobic residues leads to a 'hydrophobic radius of gyration'

$$r_{\mathrm{h}}^2 = \frac{1}{2n_{\mathrm{h}}^2}\sum_{i,j}^{n_{\mathrm{h}}} r_{ij}^2 \qquad i,\ j \text{ hydrophobic residues}, \tag{2}$$

that describes in a simple way the clustering of the n_{h} hydrophobic residues in a protein of n amino acids. In a similar way, we can define r_{p} as the corresponding 'hydrophilic radius of gyration' if we restrict our sum only to the hydrophilic residues. For a protein in its biologically active state, $r_{\mathrm{p}} > r_{\mathrm{h}}$; and the smaller r_{h} is compared to r_{p} (or r_{g}), the more pronounced is the hydrophobic core formed.

Note that the above definitions of r_{h} and r_{p} are not unique but depend on the choice of hydrophobicity scale. As there are various competing scales based on different theoretical considerations and measurement techniques [22–24], the numerical values of r_{h} and r_{p} will differ depending on the choice of scale. However, comparing various scales [18, 25, 26, 23] we found no change in the qualitative behaviour of hydrophobic and hydrophilic radius. This can be seen in figures 1(a) and (b), where we plot on a log–log scale the hydrophobic radius r_{h} as a function of the number of hydrophobic residues n_{h} for two different hydrophobicity scales: the consensus scale of [18] and the OONS scale of the Cornell group [26]. Our data points are obtained from an ensemble of 50 proteins listed in table 1 and resemble the set studied earlier by Silverman [19]. The linear correlation indicates that r_{h} as a function of protein size can be described by a power law. A fit

$$r_{\mathrm{h}} = A n_{\mathrm{h}}^{B} \tag{3}$$

has for the case of the OONS scale [26] a correlation coefficient 0.93 and leads to $A = 3.38(27)$ and $B = 0.354(20)$; and $A = 3.01(21)$ and $B = 0.353(16)$ (correlation coefficient 0.95) for the consensus scale [18]. OONS- and consensus-scale classifications are presented in table 2. Note that the exponents are within the error bars compatible with each other but slightly larger

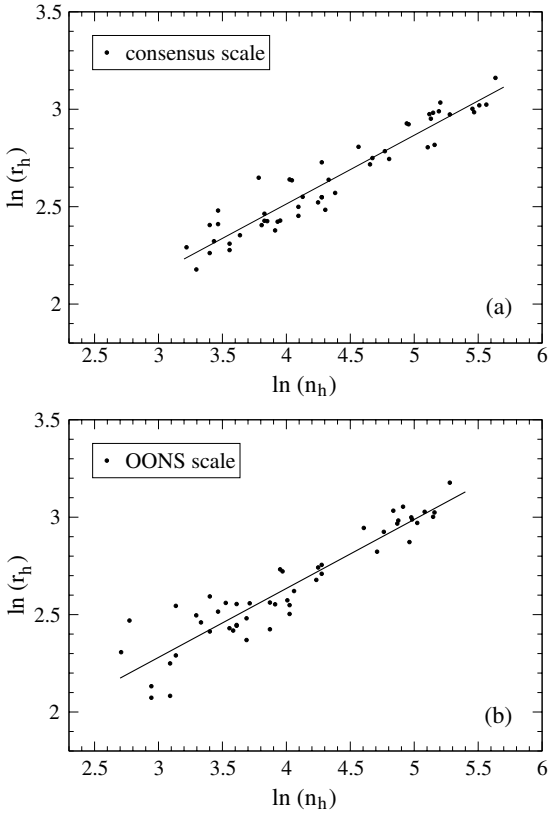

Figure 1. The hydrophobic radius r_h as a function of the number of hydrophobic residues n_h for (a) the consensus scale and (b) the OONS scale.

Table 1. PDB names of proteins and number of residues.

Name	Number	Name	Number	Name	Number	Name	Number	Name	Number
1a26	351	1bgv	449	1gai	472	1neu	115	1utg	70
1arv	123	1cdz	96	1gvp	87	1orc	64	1who	94
1aa2	108	1cq2	153	1ig5	75	1pdo	129	2act	218
1acf	125	1csp	67	1kte	105	1pgb	56	2acy	098
1ail	70	1ctq	166	1lbu	213	1phc	405	2dri	271
1akz	233	1dzo	120	1ldm	329	1phr	154	2sns	141
1at0	242	1e4f	378	1lis	131	1r69	63	2cox	500
1aua	296	1erv	105	1lzl	317	1ris	97	3pbg	468
1aun	208	1feh	574	1mjc	69	1tul	306	4fgf	124
1beo	196	1g8p	321	1msi	66	1uby	348	5pti	58

than the theoretical value $B = 1/3$ for a densely packed hydrophobic core whose volume is proportional to the number of hydrophobic residues. Since the proportionality constant depends only on the density of hydrophobic sites, one expects fewer fluctuations as n_h increases.

The scaling of r_h can also be expressed as a power law in the total number of residues n with the same exponent $\tilde{B} = B \approx 1/3$ and a pre-factor $\tilde{A} \approx c^{1/3}A$ that depends on the relative frequency of hydrophobic residues. According to the OONS scale 35.7% of all residues in

Table 2. Classification of residues according to OONS and consensus scales.

Scale	Hydrophobic	Hydrophilic	Ambivalent
OONS	ala, ile, leu, met, pro, val	arg, asn, asp, cys, gln glu, his, lys, phe, ser, thr, trp, tyr	gly
Consensus	tyr, cys, gly, ala, met, trp leu, val, phe, ile	arg, lys, asp, glu, asn gln, his, ser, thr, pro	

our set of 50 proteins are hydrophobic and 56.3% hydrophilic. The remaining percentage is residues considered neutral. In the consensus scale 50.4% of the residues are hydrophobic and 49.6% hydrophilic. A fit $r_\mathrm{h} = \tilde{A} n^{\tilde{B}}$ leads for the case of the OONS scale to values of $\tilde{A} = 2.02(21)$ and $\tilde{B} = 0.382(21)$ (correlation coefficient 0.93). The corresponding ratio $\tilde{A}/A = 0.60(7)$ is compatible with the predicted value $\tilde{A}/A \approx 0.71$. Similarly, we find for the consensus scale $\tilde{A} = 2.20(19)$ and $\tilde{B} = 0.366(17)$ (correlation coefficient 0.95), leading to a ratio $\tilde{A}/A = 0.74(8)$ that is close to the predicted value $c^{-1/3} \approx 0.79$.

Similar relations hold also for the hydrophilic radius r_p and the radius of gyration r_g itself. Especially, a fit $r_\mathrm{g} = A\, n^B$ leads to the coefficients $A = 2.86(18)$ and $B = 0.335(12) \approx 1/3$ (correlation coefficient 0.97). Note that the coefficients A and B do not depend on the hydrophobicity scale as r_g is calculated over all residues. The larger value of A indicates that overall the protein is slightly less densely packed than its hydrophobic core. The difference is due to the looser packing of hydrophilic residues. This can be seen from the fit $r_\mathrm{p} = An^B$ that leads for the OONS hydrophobicity scale to $A = 3.68(24)$ and $B = 0.331(14)$ (correlation coefficient 0.96), and for the consensus scale to $A = 3.96(20)$ and $B = 0.332(12)$ (correlation coefficient 0.97). The pre-factor A is for both hydrophobicity scales substantially larger than for hydrophobic residues. On the other hand, the exponents B have values that are again consistent with one-third, and differ little from those for other hydrophobicity scales (data not shown). Since the various hydrophobicity scales do not affect the qualitative behaviour of our hydrophobic (hydrophilic) radius, we restrict ourselves in the following to the consensus scale [18]. If not marked otherwise, all quoted values are calculated using this scale.

3. Results and discussion

We start our discussion by first analysing whether the hydrophobic (hydrophilic) radius of gyration describes as a scoring function correctly the characteristics of known protein structures. The scaling of r_h and r_p as a function of number of residues indicates that the hydrophobic residues are densely packed in the core of proteins, and the hydrophilic amino acids on average found at a larger distance from the geometric centre of the protein. We remark that equation (2) can also be written as

$$r_\mathrm{h}^2 = \frac{1}{n_\mathrm{h}^2} \sum_i^{n_\mathrm{h}} (\vec{r}_i - \vec{r}_0)^2 \,, \tag{4}$$

where $\vec{r}_0 = 1/n_\mathrm{h} \sum \vec{r}_i$ (i running over all hydrophobic residues) is the centre of geometry of the hydrophobic residues and differs little from $\vec{r}_0' = 1/n \sum \vec{r}_i$ with the sum going over all residues. While r_h is not the average distance of the hydrophobic residues to the centre of the molecule (which would be given by $\langle r_\mathrm{h} \rangle = 1/n_\mathrm{h} \sum |\vec{r}_i - \vec{r}_0|$), it can be interpreted as a characteristic separation from the protein centre. We can therefore approximate the hydrophobic core of a protein by a sphere of radius r_h around the centre of the molecule. In our ensemble of 50 proteins, 45% of all residues are located within this sphere. This number includes 57%

of the hydrophobic residues (as characterized by the consensus scale) but only 32% of the hydrophilic residues. Hence, only 35% of the residues with $r < r_h$ are hydrophilic according to the consensus scale. On the other hand, while 36% of all residues are outside a sphere with radius r_p around the centre of geometry, this number includes 47% of all hydrophilic residues but only 24% of the hydrophobic ones. In total, 66% of all residues are hydrophilic for $r > r_p$. These numbers demonstrate that r_h characterizes the hydrophobic core while r_p rather describes the onset of the hydrophilic exterior of a protein. We remark that our results seem not to depend on the choice of hydrophobicity scale. We have checked this explicitly for the OONS scale and the 'meta-scale' of [23, 24] where we found similar results (data not shown).

Our previous discussion shows that the hydrophobic radius r_h and the hydrophilic radius r_p describe common characteristics of protein structures. However, it is not clear whether they also allow differentiation between the native structure and other low-energy conformers, i.e. can serve as a scoring function. Answering this question requires a test of these quantities on sets of possible protein structures out of which the correct structure has to be selected. Numerous such decoy sets exist and allow an evaluation of scoring functions. The proteins that we use are listed in table 3 and were taken from the Park–Levitt [9] (http://dd.stanford.edu) and Baker [10] (http://depts.washington.edu/bakerpg) decoy sets. Our choice of proteins allows a direct comparison with the approach of Silverman and collaborators [11] that also attempts to profile the distribution of hydrophobic residues. We use the same proteins and decoys as these authors. As in [11] proteins stabilized strongly by disulfide bridges are excluded since the mechanism of folding and stability is different.

As an example for our analysis we show in figure 2(a) the hydrophobic radius of gyration r_h of protein configurations as a function of their root-mean-square-deviation (rmsd) from the native structure. The rmsd measured in Å is evaluated over the C_α atoms of the two structures. The displayed data are for the protein 3icb and are taken from the Park–Levitt decoy set [9]. The residues that contribute to the calculation of r_h and r_p are selected according to the consensus scale of [18] (see table 2). The horizontal dashed line marks the value of r_h for the native fold. Of the 654 decoys, 99.8% have a hydrophobic radius r_h that is larger than the native structure. The only configuration with a smaller value is very similar to the native structure: the rmsd between the two structures is less than 2 Å. The distribution of decoys exhibits a significant correlation between r_h and the similarity to the ground state: in general, r_h increases with rmsd. This correlation is much weaker for the hydrophilic radius of gyration r_p (as shown in figure 2(b)) where 383 (58.6%) of the decoys have a larger value than the native structure.

Note that no such correlation is observed between the radius of gyration r_g (evaluated over all residues) and the rmsd: r_g is uniformly distributed along the rmsd values in figure 3, with 94 (14.4%) decoys having a smaller radius of gyration than the native structure. Since r_g is a measure for the compactness of protein configurations, it follows that the observed correlation between r_h (and too a lesser degree r_p) and the rmsd values is *not* because decoys with larger rmsd to the native structure may be less compact.

Our above results indicate that the 'hydrophobic radius of gyration' r_h allows one to differentiate between the native structure and other low-energy configurations. But how does r_h depend on the potential energy of the protein? In order to answer this question we approximate for all decoys of 3icb the potential energy by a function often used in protein simulations:

$$E = E_{\mathrm{CHARMM}} + E_{\mathrm{GB}}. \tag{5}$$

Here, one assumes that the intramolecular interactions can be described by the CHARMM force field [27] and the protein–water interactions by a generalized Born ansatz [28]. E as a function of the rmsd to the native structure is drawn in figure 4(a) for the 3icb decoys. Note

Table 3. Number of decoys that have a smaller 'hydrophobic radius of gyration' r_h than the r_h^{PDB} values of the corresponding native structures. The last two columns lists the number of decoys where the hydrophobic score of [11] ('SZ score') leads to a more favourable value than found for the native ones (data taken from [11]).

				$r_h < r_h^{PDB}$					
Decoy Set	PDB name	Size	Total decoys	Consensus	%	OONS	%	SZ score	%
Park–Levitt	1ctf	68	631	28	4.4	2	0.3	4	0.6
	1r69	63	676	0	0	0	0	12	1.8
	2cro	65	675	0	0	0	0	38	5.6
	3icb	75	654	1	0.2	9	1.4	3	0.5
Baker	2ptl	60	1000	27	2.7	65	6.5	381	38.1
	1r69	61	1000	1	0.1	1	0.1	344	34.4
	1c5a	62	991	76	7.6	264	26.4	538	50.2
	1hsn	62	970	913	91.3	920	92.0	726	74.6
	1leb	63	1000	593	59.3	119	11.9	747	74.7
	2ezh	65	1000	453	45.3	398	39.8	43	4.3
	1sro	66	1000	5	0.5	36	3.6	441	44.1
	2fow	66	1000	38	3.7	79	7.9	373	37.3
	1ctf	67	1000	68	6.8	21	2.1	184	18.4
	1mzm	67	1000	301	30.1	106	10.6	156	13.6
	1nkl	70	1000	10	1.0	11	1.1	152	15.2
	1bdo	75	1000	29	29.0	8	0.8		
	1gvp	82	1000	835	83.7	588	58.9		
	1who	88	1000	0	0	4	0.4		
	1ris	92	1000	21	2.1	57	5.7		
	2acy	92	1000	0	0	22	2.2		
	1kte	100	1000	1	0.1	16	1.6		
	1pal	100	1000	20	2.0	62	6.2		
	1aa2	105	1000	0	0	0	0		
	1erv	105	1000	0	0	2	0.2		
	1pdo	121	1000	0	0	3	0.3		
	4fgf	121	1000	0	0	1	0.1		
	1acf	123	1000	0	0	0	0		

that the energies are calculated after minimizing the decoy configurations with respect to the above energy function. Through this minimization one avoids unphysically high energies due to artefacts of the energy function but changes little the rmsd of these configurations to the native structure and their values for the hydrophobic radius r_h. Structures were minimized using at most 500 steps of adopted basis Newton–Raphson (ABNR) minimization or until the tolerance was 1×10^{-5}. Generalized Born implicit solvent calculations as well as decoy minimization were performed using CHARMM22 parameters [29].

We see from figure 4(a) that 145 (22.2%) of the decoys have a lower energy than the (minimized) native structure (whose energy is marked by the straight line). Hence, our energy function is for 3icb less suitable to distinguish between native state and competing decoys. We remark that this poor performance is be due to the solvent approximation. Approximating the solvation energy instead of a generalized Born term by a Poisson–Boltzmann energy, all decoys have higher energy than the native structures. However, calculation of Poisson–Boltzmann electrostatic energies is extremely slow and makes its use in simulations unpractical. For this reason, we continue to compare our results with the more commonly used generalized-Born approximation of solvation energy. Within such an approximation, our score may become

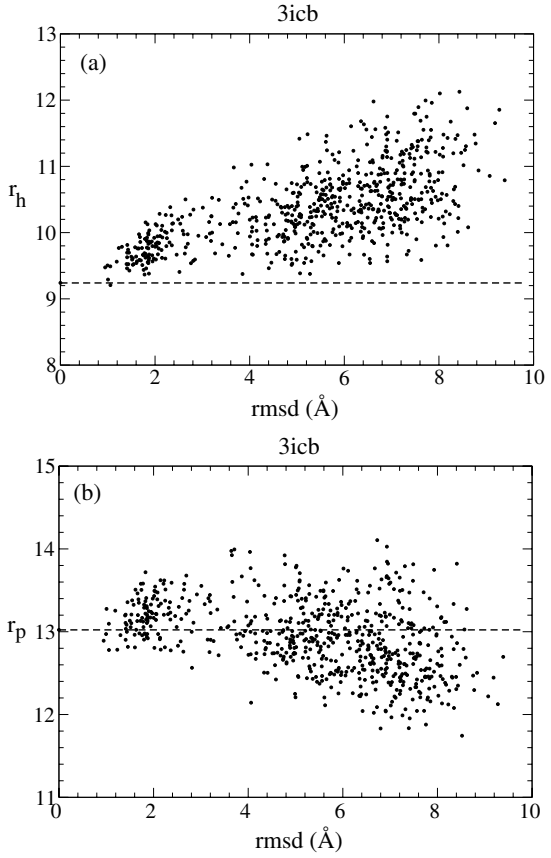

Figure 2. The (a) hydrophobic radius r_h and (b) hydrophilic radius r_p as a function of the rmsd to the native structure. Shown are data for the 3icb decoys of the Park–Levitt set.

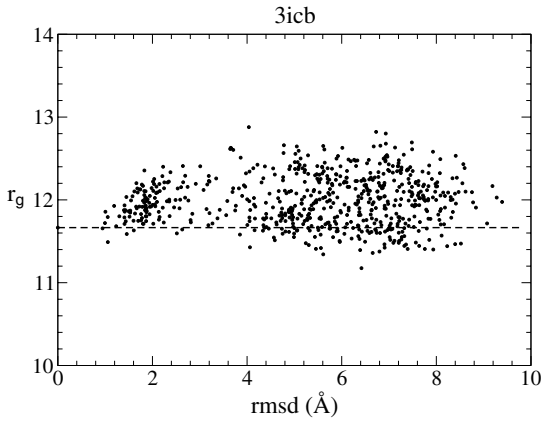

Figure 3. Radius of gyration r_g (taken over all heavy atoms) as a function of the rmsd to the native structure. Shown are data for the 3icb decoys of the Park–Levitt set.

a valuable tool to distinguish between low-energy structures as E and r_h are only weakly correlated. This can be seen in figure 4(b). While E increases with r_h, configurations with similar values of r_h can vary substantially in energy. On the other hand, configurations with similar energies may differ considerably in their hydrophobic radius. Hence, r_h is an independent scoring function and not only an approximation of the internal energy of the molecule.

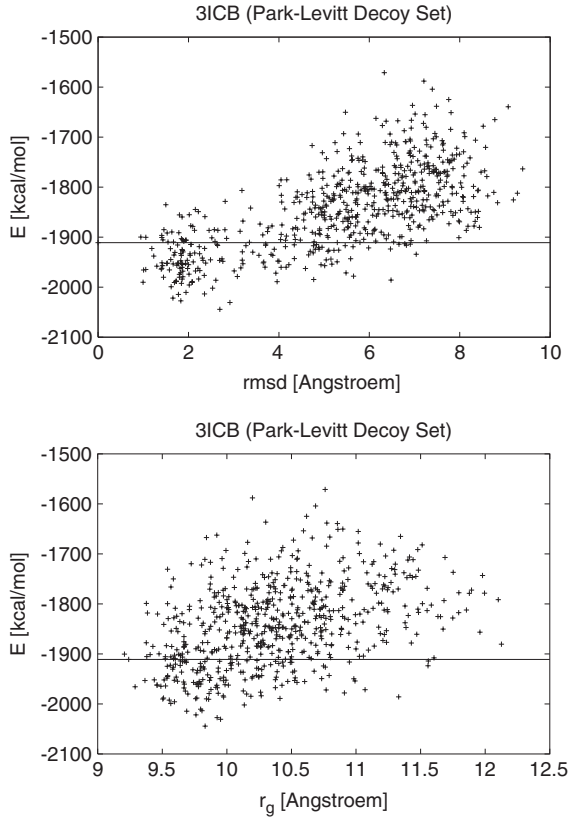

Figure 4. Potential energy versus (a) rmsd and (b) hydrophobic radius of gyration r_h. Shown are data for the 3icb decoys of the Park–Levitt set. The horizontal line marks the energy of the PDB structure.

We summarize in table 3 results for our 'hydrophobic radius of gyration' as a scoring function. Shown are values for the consensus scale and the OONS scale. Our results are compared with those of the hydrophobic score of Silverman and collaborators [11] (shown in the last column) that also rely on the consensus scale. For the Park–Levitt decoy set our results are comparable to or better than the Silverman score for three of the four proteins, but perform slightly worse for 1ctf. However, the performance of our score improves for 1ctf when the OONS scale is utilized. For the Baker set, we observe that the proteins where the Silverman score performed extremely badly (1hsn and 1leb) are also the ones where r_h leads to the largest percentage of false positives. Both 1hsn and 1leb are DNA binding proteins, and we observe a similar poor performance of our score for a third DNA binding protein (2ezh) in our set. Hence, we conclude that r_h is not a suitable scoring function for DNA binding proteins—and probably not for lipid binding proteins either as a similar poor performance is observed for 1mzm, a lipid binding protein. No such conclusion can be drawn for the Silverman score as 2ezh has the lowest failure rate for the Baker set while 1hsn and 1leb have the highest failure rates. For the proteins that are not DNA binding r_h performs on average better than the Silverman score, albeit in general worse than for the decoys of the Park–Levitt decoy sets. No substantial differences are observed between the consensus scale and the OONS scale. This demonstrates again that the qualitative behaviour of r_h is independent from the details of the underlying hydrophobicity scale.

However, a correlation is observed between the failure rate of the hydrophobic radius r_h and the size of proteins. Table 3 also lists proteins of the Baker set that were not used in [11]. We have ordered these molecules according to their size (measure in the number of amino acids that constitute the protein chain). One of these additional 12 proteins, 1gvp, is a DNA binding protein, and r_h again performs poorly as a scoring function: the failure rate is 83.7%. However, focusing on the non-DNA-binding proteins in the Baker set, one observes that the failure rate of the hydrophobic radius decreases with increasing size. For proteins with more than 100 residues no false positives are observed! This observation is not surprising as it is well known that formation of a hydrophobic core is more pronounced in proteins of more than 100 residues. Hence, both the 'hydrophobic radius' r_h and the hydrophobic ratio of the Silverman group work best for larger sized single-domain proteins with more than 100 residues [11].

Our above results demonstrate the ability of the hydrophobic radius to score low-energy structures. Similar to the hydrophobic score of [11] it classifies protein configurations according to formation of a hydrophobic core. This allows both scores with comparable probability to differentiate between native-like structures and competing low-energy structures. However, both scores are imperfect and may lead to false positives, especially for proteins smaller than 100 residues. Hence, scores based on hydrophobic profiling should be used only as one of many criteria in prediction of native states of proteins. Within this group of scores the hydrophobic radius r_h has the advantage that its calculation is simple and fast. As the quantity is only weakly correlated with the energy it may also be suitable as a coordinate for designing generalized ensembles [30] that allow an improved sampling of low-energy configurations.

For instance, in *energy landscape paving* (ELP) [31] one performs low-temperature Monte Carlo simulations with a modified energy expression that steers the search away from regions already explored:

$$w(\tilde{E}) = \mathrm{e}^{-\tilde{E}/k_B T} \qquad \text{with } \tilde{E} = E + f(H(q,t)). \tag{6}$$

Here, T is a (low) temperature, $\tilde{E}$ serves as a replacement of the energy E and $f(H(q,t))$ is a function of the histogram $H(q,t)$ in a pre-chosen 'order parameter' q. It follows that within ELP the weight of a local minimum state decreases with the time the system stays in that minimum until the local minimum is no longer favoured. The system will then explore higher energies until it falls into a new local minimum. Of critical importance for the working of this method is the choice of the 'order parameter' q that differentiates between the various local minima. We expect that the hydrophobic radius r_h is a suitable choice as ELP works optimally if q is only weakly correlated with the energy.

We have tested this conjecture by simulating the 36-residue protein HP-36 [32]. As one of the few small proteins that have a well defined secondary and tertiary structure and can fold autonomously [32], it is sufficiently complex and with 596 atoms of a size that numerical simulations become a challenge [33]. In order to compare our data with previous results [31, 34], we have again performed here all-atom simulations of this molecule that rely on of the ECEPP/3 force field [35] (as implemented in the program package SMMP [36]) and a solvent-accessible surface term [26] to approximate the protein–water interactions. We chose $f(H(q,t) = H(r_h))$ and $T = 100$ K.

In [34] it was observed that the native structure as deposited in the PDB (1vii) and shown in figure 5(a) is *not* the dominant structure at $T = 300$ K in simulations with ECEPP force field [35] and a solvent accessible surface term [26] to approximate the protein–water interaction. Instead, 90% of configurations resemble at this temperature the structure shown in figure 5(b) that is also (the minimized) lowest-energy configuration found in our simulations. Its energy is $E = -343$ kcal mol^{-1}. Both types of configurations have at room temperature similar average energies [34]. Hence, without *a priori* knowledge of the experimental structure

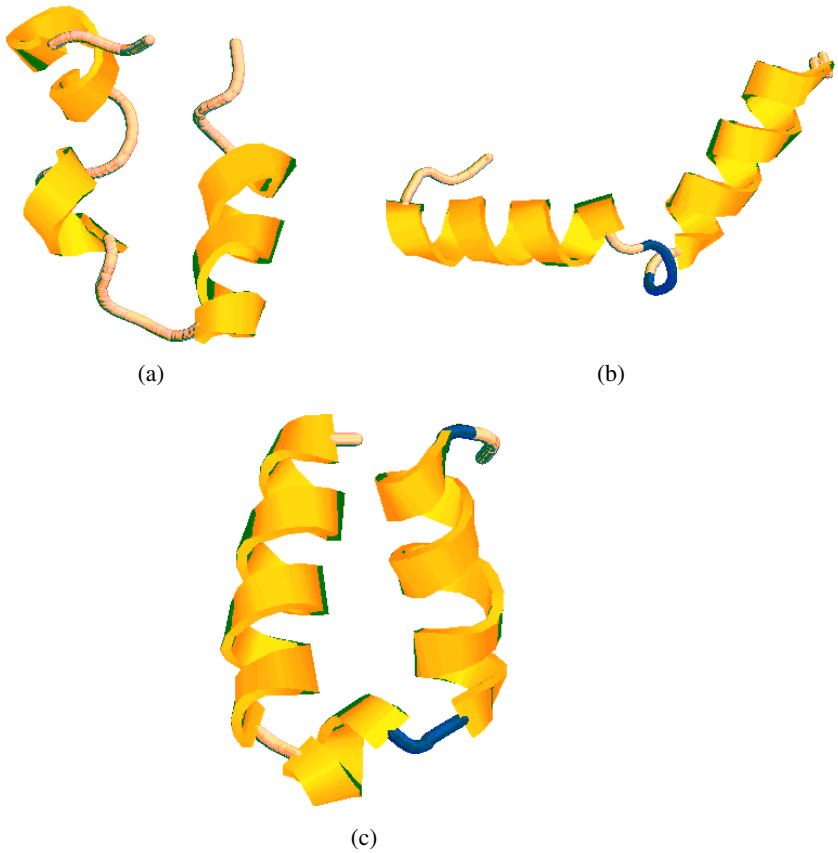

Figure 5. (a) Structure of HP-36 as deposited in the PDB (1vii); (b) lowest-energy configuration of HP-36 as found in our ELP simulation; and (c) configuration of HP-36 that has the smallest value of the hydrophobic radius r_h. The figures have been prepared with RASMOL [38].

(This figure is in colour only in the electronic version)

it is not possible to identify the native structure, as in our simulations (and in contradiction to the experimental results of [32]) it is not the global free energy minimum. However, both structures differ in their hydrophobic radius: using the consensus scale we find $r_h = 8.7$ Å for the native structure and $r_h = 12.4$ for the competing structure that dominates in the ECEPP simulations. On the other hand, the configuration shown in figure 5(c), that has a hydrophobic radius $r_h = 9.8$ (the smallest value found in our simulation), has a higher similarity with the native structure despite the fact that its energy is at $E = -328$ kcal mol^{-1} higher than that of figure 5(b). Its rmsd from the native structure is 5.4 Å and comparable to that of structures found in earlier work by different methods [33, 31]. We remark that modifications of the solvent term lead to structures closer to the native structure [37]. The above-presented results need to be taken with a grain of salt as the observed failure rates for both the Park–Levitt and the Baker decoy sets indicate that the native configuration is not always the one with smallest hydrophobic radius. However, our preliminary results from ELP simulations of HP-36 show that the choice of r_h as an ensemble coordinate in ELP and other generalized-ensemble methods leads indeed to an improved sampling of low-energy configurations.

4. Conclusion

We have proposed a simple measure, the hydrophobic radius of gyration r_h and derived quantities, as a new tool for discriminating the native structure out of an ensemble of decoys. It can therefore complement an energy minimization in search of the native fold and may proof a useful tool in evaluating candidate structures in structure prediction methods. The performance of the new score is comparable to the hydrophobic moment profiling approach by Silverman and collaborators [11], but its evaluation is much easier. Preliminary results from ELP simulations of the 36-residue protein HP-36 suggest that our score may also be useful for guided molecular dynamics simulations (or other biased search schemes) and as an ensemble coordinate in generalized-ensemble simulations [30].

Acknowledgments

Support by a research grant from the National Institutes of Health (GM62838) is gratefully acknowledged. Part of this article was written while N A Alves was visiting Michigan Technological University. He thanks the MTU Physics Department for kind hospitality.

References

[1] Bryngelson J D and Wolynes P G 1987 *Proc. Natl Acad. Sci. USA* **84** 7524
[2] Boczko E M and Brooks C L III 1995 *Science* **269** 393
[3] Onuchic J N, Luhey-Schulten Z and Wolynes P G 1997 *Annu. Rev. Phys. Chem.* **48** 545
[4] Hansmann U H E, Okamoto Y and Onuchic J N 1999 *Proteins: Struct. Funct. Genet.* **34** 472
[5] Hansmann U H E and Onuchic J N 2001 *J. Chem. Phys.* **115** 1601
[6] Lin C-Y, Hu C-K and Hansmann U H E 2003 *Proteins* **52** 436
[7] Halgren T A 1995 *Curr. Opin. Struct. Biol.* **5** 205
[8] Lazaridis T and Karplus M 2000 *Curr. Opin. Struct. Biol.* **10** 139
[9] Park B and Levitt M 1996 *J. Mol. Biol.* **258** 367
[10] Simons K T, Ruczinski I, Kooperberg C, Fox B A, Bystroff C and Baker D 1999 *Proteins* **34** 82
[11] Zhou R, Silverman B D, Royyuru A K and Athma P 2003 *Proteins: Struct. Funct. Genet.* **52** 561
[12] Seok C, Rosen J B, Chodera J D and Dill K A 2003 *J. Comput. Chem.* **24** 89
[13] Berglund A, Head R D, Welsh E A and Marshall G R 2004 *Proteins: Struct. Funct. Bioinformat.* **54** 289
[14] Kauzmann W 1959 *Adv. Protein Chem.* **14** 1
[15] Dill K A 1990 *Biochemistry* **29** 7133
[16] Southall N T, Dill K A and Haymet A D J 2002 *J. Phys. Chem.* B **106** 521
[17] Zhou H and Zhou Y 2002 *Proteins: Struct. Funct. Genet.* **49** 483
[18] Eisenberg D, Weiss R M, Terwilliger T C and Wilcox W 1982 *Faraday Symp. Chem. Soc.* **17** 109
[19] Silverman B D 2001 *Proc. Natl Acad. Sci. USA* **98** 4996
[20] Silverman B D 2003 *Protein Sci.* **12** 586
[21] Silverman B D 2003 *Proteins: Struct. Funct. Genet.* **53** 880
[22] Karplus P A 1997 *Protein Sci.* **6** 1302
[23] Biswas K M, DeVido D R and Dorsey J G 2003 *J. Chromatogr.* A **1000** 637
[24] Trinquier G and Sanejouand Y-H 1998 *Protein Eng.* **11** 153
[25] Eisenberg D and McLachlan A D 1986 *Nature* **319** 199
[26] Ooi T, Oobatake M, Némethy G and Scheraga H A 1987 *Proc. Natl Acad. Sci. USA* **84** 3086
[27] Brooks B R, Bruccoleri R E, Olafson B D, States D J, Swaminathan S and Karplus M 1983 *J. Comput. Chem.* **4** 187
[28] Qiu D, Shenkin P S, Hollinger F P and Still W C 1997 *J. Phys. Chem.* **101** 3005
[29] Dominy B and Brooks C L III 1999 *J. Phys. Chem.* **103** 3765–73
[30] Hansmann U H E 2004 Protein folding in silico—the quest for better algorithms *New Optimization Algorithms in Physics* ed A Hartmann and H Rieger (Weinheim: VCH–Wiley)
[31] Hansmann U H E and Wille L T 2002 *Phys. Rev. Lett.* **88** 068105
[32] McKnight C J, Doehring D S, Matsudaria P T and Kim P S 1996 *J. Mol. Biol.* **260** 126

[33] Duan Y and Kollman P A 1998 *Science* **282** 740
[34] Lin C-Y, Hu C-K and Hansmann U H E 2003 *Proteins: Struct. Funct. Genet.* **52** 436
[35] Némethy G, Gibson K D, Palmer K A, Yoon C N, Paterlini G, Zagari A, Rumsey S and Scheraga H A 1992 *J. Phys. Chem.* **96** 6472
[36] Eisenmenger F, Hansmann U H E, Hayryan Sh and Hu C-K 2001 *Comput. Phys. Commun.* **138** 192
[37] Hansmann U H E 2004 *Phys. Rev.* E **70** 012902
[38] Sayle R and Milner-White E J 1995 *Trends Biochem. Sci.* **20** 9
Bernstein H J 2000 *Trends Biochem. Sci.* **25** 9