
Journal of Biological Physics **27:** 205–215, 2001.

Assessment of the Quality of Energy Functions for Protein Folding by Using a Criterion Derived With the Help of the Noisy Go Model

M. VENDRUSCOLO
Oxford Centre for Molecular Sciences, New Chemistry Laboratory, University of Oxford, Oxford OX1 3QT, UK; e-mail: michelev@bioch.ox.ac.uk

Abstract. We propose a simple criterion based on the Z-score to assess the quality of energy functions for protein folding: one should obtain $Z < -10$ for the equilibrium ensemble at about native conditions. We derive this criterion by studying a Go model with random errors added to the native interactions. The dependence of the Z-score on the thermodynamic parameters, including the noise, can be precisely obtained in this case, as the ground state of the model is known exactly. We apply this criterion to rapidly rule out two otherwise promising pairwise energy approximations. The advantage of adopting the present criterion is that it is not necessary to know the ground state of an energy function to assess its quality. It is sufficient to compute the Z-score from a single equilibrium simulation at around the folding temperature.

Key words: Effective energy functions, Go model, Protein folding, Z-score

1. Introduction

In attempting to perform protein folding by energy minimization an approximation of the extremely complex energy of the system must be used. The basic requirement on an approximate energy function is that, at least for the protein under study, the ground state must coincide with the native state [1]. However, in most practical applications one does not know which is the ground state of the energy function used. Therefore there is the need of a criterion to assess the 'quality' of an energy function, namely whether the native state and the ground state are likely to coincide. An exact criterion has been proposed by Vendruscolo and Domany [1]. By using their approach it is possible to prove whether or not one can choose the interaction parameters for a particular form of the energy in order to assign the lowest energy to known native states of proteins. From this point of view, the question is not whether energy parameters are better derived by using knowledge-based or physico-chemical methods [2], but rather whether suitable parameters exist at all. An alternative, approximate but more easily implementable criterion, has been proposed by Mirny and Shakhnovich [3]. They adopted the Z-score, which measures the difference between the energy of the native state and

the average energy of alternative conformations, measured in units of standard deviations of the energy distribution.

In this paper we give an estimate for the threshold value of the Z-score, namely a value that gives a reasonable confidence that the energy function used can allow to fold proteins to their native states. In order to obtain this result, we introduce the 'noisy Go model', a Go model [4] with a random error added to each native-like pairwise interaction. In this model, the energy E of a particular conformation C is

$$E(C) = -\sum_{j>i} \left[S_{ij} S_{ij}^N \left(1 - \eta\, q_{ij}\right) - \epsilon\, S_{ij} \left(1 - S_{ij}^N\right) \right] \tag{1}$$

where S is the contact map of conformation C, S^N is the contact map of the native state, η is a parameter controlling the strength of the noise, q is a uniform random number in $[0, 1]$ and ϵ is a positive constant which disfavors non-native interactions. In this work, the contact map S is set to 1 if the C_α atoms of a pair of residues are closer than a threshold distance R_c, here set to 8.5 Å, and to 0 otherwise [1]. The main reason to use the Go model with noise is that, at least for small noise, it is possible to know with certainty the ground state of the system. We made two assumptions in the noisy Go model of Eq. (1), (i) the noise acts only upon the native interactions and (ii) the noise is quenched, namely it is fixed before starting a simulation. All the simulations are made by using a Monte Carlo (MC) algorithm in real space, as described elsewhere [1]. The procedure is based on crankshaft moves for individual residues [1], represented by the coordinates of their C_α atoms. In the rest of the paper we set $\epsilon = 0.1$ [5].

The Go model [4] was originally proposed to facilitate folding on a computer. It has been recently the object of renewed attention because it has been argued that it can correctly describe the dynamics of the folding process [5–7] and the geometrical properties of native state conformations [8, 9].

In the present work we suggest that an energy function for protein folding must be characterized by $Z < -10$ when the average energy is computed on a suitable set of alternative conformations, or decoys. By 'suitable' we mean a set of decoys that represent the equilibrium ensemble at the thermodynamic conditions of folding. This result is proved for the Go energy function of Eq. (1) by increasing the strength η of the noise and by computing the Z-score for ensembles of conformations obtained at different temperatures by MC simulations. We suggest that in practice, in order to test the quality of a given energy function, one can generate decoys by carrying out a simulation at T either slightly above T_f and check whether $Z < -10$ or slightly below T_f and check whether $Z < -8$. We observe that one should not be misled by the fact that for some set of poorly chosen decoys one can obtain $Z < -10$. Poor decoys are conformation whose statistical weight at $T \simeq T_f$ is negligible and therefore they do no participate in the folding process.

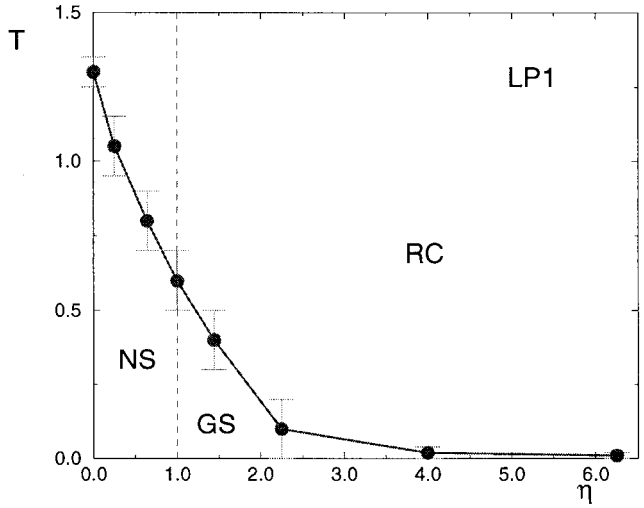

Figure 1. Phase diagram (η, T) of the noisy Go model of Eq. (1) for protein L (LP1), where η is the strength of the noise and T is the temperature. The line is the folding temperature T_f as a function of η. For $T > T_f$ the protein is in a random coil state (RC) and for $T < T_f$ it is in the native state (NS) for $\eta < 1$. For $\eta \geq 1$ the ground state (GS) can be different from the NS. For $\eta > 2$, the transition temperature becomes vanishingly small. A similar phase diagram is found for CI2.

2. Study of the Noisy Go Model

In our MC simulations we considered a fragment of the immunoglobulin light chain-binding domain of protein L from *peptostreptococcus magnus* (LP1, PDB code 2ptl [10]), which has an ubiquitin-like fold. LP1 has no disulfide bonds and no proline residues. We used a truncated form of 60 amino acids, in which the first 18 amino acids were eliminated, since they are disordered in the NMR structure.

For the model described above, we computed the phase diagram (η, T) of LP1, by performing equilibrium MC simulations. The results are shown in Figure 1. For each point in the (η, T) plane we averaged over typically 4 different realizations of the disorder. The errorbars shown are originated from the approximate criterion used to identify the transition temperature T_f, based on the flex point of the $R_g(T)$ curve, where R_g is the radius of gyration of the system [5] and also from the variations due to different realizations of the noise.

For $\eta > 1$, in general, the native state (NS) is not the ground state (GS) of the Go model of Eq. (1) and the 'folding temperature' T_f is the temperature in which the system is stable in its ground state, not in the native state. From simple considerations, we can expect three different regimes, depending on the strength η of the noise. The first regime, called here 'native' (R_N), is realized for $\eta < 1$; in this case all the native interactions are favorable and the GS coincides with the NS. In the second regime, called here 'molten globule' (R_{MG}) [11], for $1 \leq \eta \leq 2$, some

native interactions are repulsive but the protein chain, by expanding its volume, manages to accommodate them by maintaining the overall native topology. For $\eta = 2$ the average perturbed pairwise interaction is $\left[\eta q_{ij}\right] = 1$, where the square brackets indicate an average over different realizations of the noise, and therefore it is equal to the energy gain of forming a native interaction. The energy $[E_N]$ of the native state averaged over the noise is therefore equal to 0. We found that the typical RMSD between the NS and the GS is of about 6 Å for $\eta = 2.0$. In the third regime, called here 'random coil' (R_{RC}), for $\eta > 2$, few native interactions are attractive and they are insufficient to determine a unique structure. In this case, the protein is typically found in a RC state, as in a good solvent, except that at very low temperatures one expects a glassy behavior, due to the competition, not sufficiently weakened by the entropy, between attractive and repulsive interactions. We are not concerned here in studying the properties of this glass phase.

In principle, since as suggested by the ϕ value analysis [12, 13], there are interactions which are more important than others to determine the structure of the native state, averaging over the disorder presents the following problem. For a given η, specific assignments of the disorder q_{ij} with particularly low values on the important interacting pairs can lead to a protein with much better folding properties than another protein for which particularly high values of the noise are assigned to the same important interactions. This fact can be exploited to investigate the role of such special contacts in determining the transition state by tuning their interactions [14]. In the present work, this problem makes it problematic to determine the precise location of the boundary between R_{MG} and R_{RC} which we argued to be at $\eta = 2$. We have not investigated in detail this aspect here.

The main result of this work is presented in Figure 2 which shows the contour levels of Z in the (η, T) plane. In this paper the Z-score is defined as

$$Z = \frac{[E_N] - [\langle E \rangle]}{[\sigma]} \tag{2}$$

where all the quantities are averaged over different realizations of the noise. The angular brackets denote thermal averages and σ is the standard deviation in the distribution of the energy E of the decoys. For $\eta = 0$ this definition coincides with the standard one [3]. In Figure 2 we are particularly interested in the region of the (η, T) plane around the $T = T_f(\eta)$ curve, which reports the dependence of the folding temperature T_f on η (see also Figure 1). The location of the contour levels is determined with a precision in T of about 0.1, due to uncertainty introduced by the different realizations of the noise. For small η, in the R_N regime, one has $Z \simeq -9$ at $T \simeq T_f$. We note that at $T = T_f$, $Z(T)$ has a local maximum, due to the bimodal distribution of the energies. We neglect here in this special case since it is not relevant within this study. Upon increasing η, the Z-score deteriorates, as illustrated by the divergence of the $T_f(\eta)$ curve and the contour level $Z = -9$. For intermediate η, in the R_{MG} regime, one has a dramatic change in the behavior of the Z-score. For $\eta > 1.5$ one must go to $T \gg T_f$ to find the native state

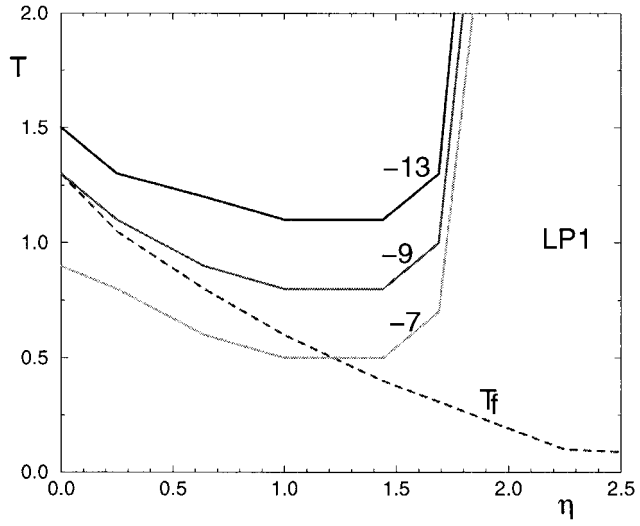

Figure 2. Z-score as a function of the strength η of the noise and of the temperature T. Three contour levels are shown, for $Z = -13, -9$ and -7. The folding temperature T_f as a function of η is also shown for reference (dashed line).

at the bottom of the energy distribution. However, a negative Z-score is in this case misleading, since the NS is not the GS of the system and the value of the Z-score is due to the fact that we have considered an ensemble of decoys typical of equilibrium conditions in which the free energy is very high. For even larger η, in the R_{RC} regime, the Z-score is always positive.

We now explore the effect of the noise on the kinetic properties of the noisy Go model of Eq. (1). Our results should be compared with recent simulations of folding for the Go model without noise [5, 6]. The folding time τ, for $\eta = 0$, is minimal at about $T = 0.7$, when the thermodynamics folding transition is at $T = 1.30 \pm 0.05$ (See Figure 3a). The minimum of τ at $T = 0.7$ arises from the competition between a rapid energy minimization and the necessity to avoid to remain trapped in local minima. At each temperature, τ is averaged over 10 trajectories. When $\eta > 0$ we also averaged τ over 4 different realizations of the noise. Since in the present model there is no energetic preference about the chirality, we consider specular structures as equivalent. This means that a successful folding has a 0.5 probability to reach the specular counterpart of the native state. Upon increasing η, folding becomes slower on average, as shown in Figure 3b, but still possible. The deterioration of the dynamical properties of the model for increasing η parallels the deterioration of the thermodynamic properties discussed above. For $\eta > 1$ there are very few 'fast tracks' [5, 6] available for folding. These fast tracks, which are particular trajectories which allow rapid folding, have been observed in other studies of folding processes using the Go model [5], that is, for $\eta = 0$. We

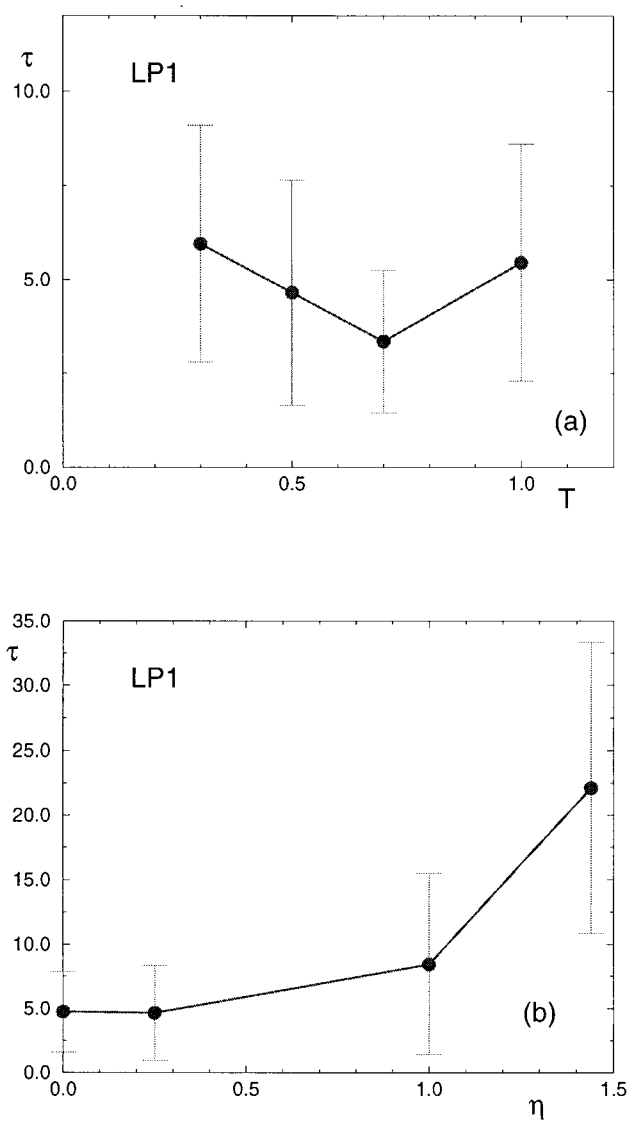

Figure 3. (a) Folding time τ dependence on temperature T for $\eta = 0$. The folding temperature is $T_f = 1.3$ and the folding time is in units of 10^6 MC steps. (b) Folding time τ dependence on the noise strength η. The temperature is $T = 0.5$ in the simulations at $\eta = 0, 0.25$ and 1.0 and $T = 0.25$ at $\eta = 1.44$. There are two sources of variation in the observed folding time, the first is due to different folding trajectories for a given assignment of the noise and the second to different assignments of the noise.

have not investigated the dynamical properties in the R_{MG} regime since in this case the GS is not the NS.

3. Two Applications of the Criterion

Our aim in this section is to assess the quality of two augmented parametrizations of the simple contact pairwise energy approximation for proteins. In the first one we discretize the distance dependence of the pairwise energy of interaction between amino acids. In the second one we separate pairwise contacts according to whether they form local or non-local interactions.

In the first form that we study the idea is to approximate the continuous distance dependence of the inter-residue interaction with a series of discrete steps. To this end, we introduce N_s threshold distances $R_c^{(k)}$ and, correspondingly, N_s contact maps $S^{(k)}$ (with $k = 1, \ldots, N_s$) associated with a certain conformation C of the protein. We define $S_{ij}^{(k)} = 1$ if the distance r_{ij} between residues i and j is such that $R_c^{(k-1)} < r_{ij} < R_c^{(k)}$ (with $R_c^{(0)} = 0$) and $S_{ij}^{(k)} = 0$ otherwise. We define the energy of a conformation C as

$$E_S(C) = \sum_{k=1}^{N_s} \sum_{j>i} S_{ij}^{(k)} W^{(k)}(s_i, s_j) \tag{3}$$

where k runs over the N_s steps of the energy and $W^{(k)}(s_i, s_j)$ is a parameter specifying the energy gained when residues s_i and s_j are in contact within the k-th step. We adopted an all-atom definition of contact with a threshold distance of $R_c = 4.5$Å. We need to specify one 20×20 symmetric matrix for each one of the N_s steps in order to specify the energy in our approximation. The total number of energy parameters is therefore $N_w = 210 N_s$. The standard pairwise contact energy function corresponds to $N_s = 1$.

In the second parametrization that we study we distinguish between short and long-ranged pairwise contacts. In this way we attempt to build in the effect of the local rigidity of the backbone and the tendency of forming secondary structure. This is taken into account by introducing a separate set of 210 pairwise energy parameters for short ranged contacts, where the range is controlled by the separation D of two residues along the chain. In this approximation, the energy is defined as

$$E_D(C) = \sum_{j>i} S_{ij}^{(S)} W^{(S)}(s_i, s_j) + \sum_{j>i} S_{ij}^{(L)} W^{(L)}(s_i, s_j) \tag{4}$$

where $S_{ij}^{(S)} = 1$ if residues i and j are in contact and $|i - j| \leq D$ and 0 otherwise and $S_{ij}^{(L)} = 1$ if i and j are in contact and $|i - j| > D$ and 0 otherwise. In order to specify this form of the energy we introduced two sets of 210 energy parameters, $W^{(S)}$ and $W^{(L)}$, for short and long ranged contacts respectively. For $D = 2$ we

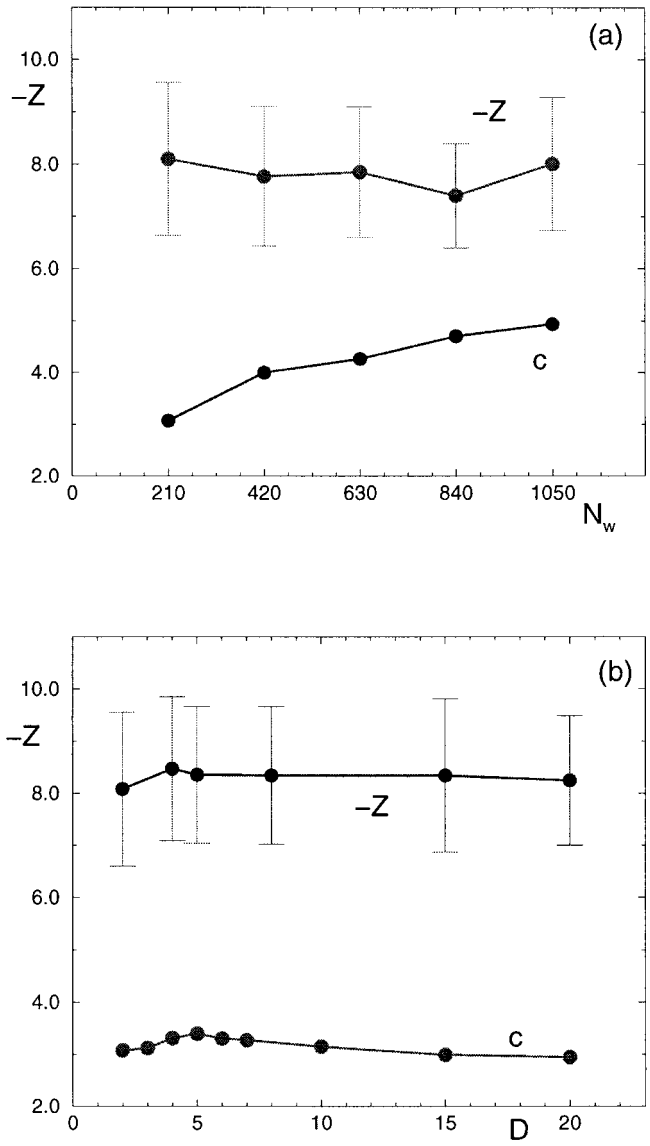

Figure 4. (a) Z-score (full circles with errorbars) as a function of N_w, the number of parameters in the energy function E_S of Eq. (3). These parameters are optimized by maximizing the field c of maximal stability (full circles) for decoys obtained by gapless threading in a database of 50 proteins. The value of $Z \simeq -8$ can be used to discard E_S as a suitable approximation of the energy for protein folding. Errobars in the Z-score represent the standard deviation in the set of analyzed proteins. (b) Assessment of the quality of the energy E_D of Eq. (4). We show the Z-score as a function of D, the separation between local and non-local contacts (see text).

recover the pairwise contact approximation, since contacts between i and $i+1$ are not considered here.

We tested the energy function E_S in the case of decoys obtained by gapless threading in a small database of 50 proteins. In order to optimize energy parameters, we maximized the field of maximal stability c by perceptron learning. We have discussed in detail this approach for energy parameter optimization elsewhere [15]. For each N_s the energy parameters are normalized so that

$$\sum_{k=1}^{N_s} \sum_{j>i} \left[W^{(k)}(s_i, s_j)\right]^2 = 1 \tag{5}$$

We report in Figure 4a the maximal stability field c as a function of the number of parameters N_w in the energy E_S. The tendency which emerges is that the quality of the energy function, as measured by c, is increasing with the number N_s of steps. However, we found $Z \simeq -8$ for any N_s, signalling an energy function of poor quality. It is likely that the Z-score could be somewhat improved by optimizing energy parameters by using the Z-score itself [3]. However, the maximization of the field of maximal stability c guarantees that all the native states are those of minimal energy and the practical equivalence with the Z-score method has been recently discussed [16]. In this study we have verified that sets of parameters with less optimized c did also have a less negative Z-score. Interestingly, upon increasing the number of steps the more difficult proteins are better optimized. For example, for the single step potential ($N_w = 210$), we found 3 proteins with $Z > -4$ while for the 4 steps potential ($N_w = 840$) the worst value found was $Z = -5.2$. The criterion that we suggested prescribes that one should find $Z < -10$ in conditions slightly above the folding ones. We have assumed here that the decoys obtained by gapless threading do correspond to these conditions. By doing so, we imply that a protein can, hypothetically, explore the conformations obtained by threading during its thermal motion at T slightly above T_f but that none of them is actually the one of minimal energy.

We note that for a given N_s one has to specify the set $\{R_s\}$ of threshold distances. The data in Figure 4a correspond to one particular choice of such set, which is different for each N_s. A more quantitative study would consist in finding

$$\tilde{c} = \max_{\{R_s\}} \; c(\{R_s\}) \tag{6}$$

that is the maximal c over all the possible choices for the set $\{R_s\}$ of threshold distances.

For the energy function E_D we optimized the energy parameters following the same procedure described above. Figure 4b shows that there is an improvement of about 10% with respect to the simple pairwise contact energy approximation when the threshold is set at $D = 5$. This result is consistent with the expectation that a better treatment of the energetics is attained when α helices and β hairpins are considered. We compare the improvement obtained in this way with the one for E_S

and two steps, which also involves 420 energy parameters. In that case we obtained an improvement of about 30%. As for E_S, also for E_D we found $Z \simeq -8$ for all the D studied. We therefore suggest that E_D is not a suitable energy function for protein folding studies.

4. Conclusions

By studying the properties of the noisy Go model, we have derived an efficient criterion for assessing the quality of residue-specific functions for protein folding. The motivation is that, at present, performing protein folding by using simple models, which are the only ones amenable to the simulation of the folding process [5, 6, 17–21], is prevented by the lack of a suitable approximation of the energy. The widespread pairwise contact energy function has been shown to be unsuitable, even when the energy parameters are optimized for a single protein [1]. The exploration of new approximations of the energy requires convenient criteria to assess their quality. This is a difficult problem, since in general the ground state of a given energy function is very difficult to find [2]. Here we have presented a method which does not require the knowledge of the ground state in advance, but only of the approximate temperature of the coil-globule collapse.

We proposed that a potential is suitable for protein folding if $Z < -10$ in the equilibrium ensemble for native or nearly-native conditions. This suggestion is based on the finding that for the noisy Go model in the regime of small noise ($\eta < 1$), one has $Z \simeq -10$ for T slightly above T_f. The Go model represent an artificial situation, in which the native state has no competitors. In the general case one can expect to find challenging outliers, whose energy is also at 10 standard deviation or more below the average energy at $T \simeq T_f$. This is why $Z(T_f) = -10$ is our proposed upper threshold for Z. This threshold value for Z is well known from circumstantial evidence: Zhang and Skolnick [22] estimated $Z < -15$, from calorimetric measurements on native proteins; Mirny and Shakhnovich [3] showed that typically $Z > -10$ for various pairwise contact potentials for threading and for lattice simulation tests, which is also in agreement with the result [1] that such potentials are unsuitable for protein folding.

To illustrate the use of the Z-score criterion, we have shown that two augmented forms of the pairwise energy approximation do not meet the criterion introduced in this work. The database that we used in the gapless threading experiment is arguably small. However, even for such an easy case it is not possible to obtain $Z < -10$ for the two energy functions that we tested. Therefore it is unlikely than by making the problem more challenging by including more decoys, for example generated by the Monte Carlo procedure of Ref. [1], the performance of these energy function would improve.

The possibility to assess the quality of a given approximation of the energy by using rapid tests like the one presented here is the main motivation for introducing the criterion discussed in this work. This criterion make it possible a preliminary

screening of several forms of energy functions. Candidates that pass this test can then be analyzed in detail by using other existing more rigorous methods, as for example the perceptron technique [1] or the q-method [23].

Acknowledgements

I am grateful to Harvard University, where part of this work was done, for hospitality and to E. I. Shakhnovich for the suggestion that originated this study. I thank E. Domany, B. Eaton, M. Karplus, E. Kussell, A. Maritan, L. Mirny, H. Orland, E. I. Shakhnovich and J. Shimada for clarifying discussions and EMBO for financial support.

References

1. Vendruscolo, M. and Domany, E.: *J. Chem. Phys.* **109** (1998), 11101–11101.
2. Lazaridis, T. and Karplus, M.: *Proteins* **35** (1999), 133–152.
3. Mirny, L.A. and Shakhnovich, E.I.: *J. Mol. Biol.* **264** (1996), 1164–1179.
4. Go, N.: *Ann. Rev. Biophys. Bioeng.* **12** (1983), 182–210.
5. Zhou, Y. and Karplus, M.: *Nature* **401** (1999), 400–403.
6. Shimada, J., Kussel, E.L. and Shakhnovich, E.I.: *J. Mol. Biol.* **308** (2001), 79–95.
7. Maritan, A., Micheletti, C. & Banavar, J.R.: *Phys. Rev. Lett.* **84** (2000), 3009–3012.
8. Bahar, I., Wallqvist, A., Covell, D.G. and Jernigan, R.L.: *Biochemistry* **37** (1998), 1067–1075.
9. Micheletti, C., Banavar, J.R., Maritan, A. and Seno, F.: *Phys. Rev. Lett.* **82** (1999), 3372–3376.
10. Wikstrom, M., Sjobring, U., Kastern, W., Bjorck, L., Drakenberg, T. and Forsen, S.: *Biochemistry* **32** (1993), 3381–3386.
11. Ptitsyn, O.B.: *Adv. Prot. Chem.* **47** (1995), 83–229.
12. Fersht, A.R.: *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding.* W.H. Freeman & Co., New York, 1999.
13. Vendruscolo, M., Paci, E., Dobson, C.M. and Karplus, M.: *Nature* **409** (2001), 641–645.
14. Dokholyan, N.V., Buldyrev, S.V., Stanley, H.E. and Shakhnovich, E.I.: *J. Mol. Biol.* **296**(5) (2000), 1183–1188.
15. Vendruscolo, M., Najmanovich, R. and Domany, E.: *Proteins* **38** (2000), 134–148.
16. Vendruscolo, M., Mirny, L., Shakhnovich, E.I. and Domany, E.: *Proteins* **41** (2000), 192–201.
17. Hao, M.-H. and Scheraga, H.A.: *Proc. Natl. Acad. USA* **93** (1996), 4984–4989.
18. Skolnick, J. and Kolinski, A.: *Adv. Chem. Phys.* **105** (1999), 203–242.
19. Shea, J.-E., Onuchic, J.N. and Brooks, C.L.: *Proc. Natl. Acad. USA* **96** (1999), 12512–12517.
20. Thirumalai, D. and Klimov, D.K.: *Curr. Opin. Struct. Biol.* **9** (1999), 197–207.
21. Riccle, D.S., Grantcharova, V.P., Santiago, J.V., Alm, E., Ruczinski, I. and Baker, D.: *Nature Struct. Biol.* **6** (1999), 1016–1024.
22. Zhang, L. and Skolnick, J.: *Prot. Sci.* **7** (1998), 1201–1207.
23. Bastolla, U., Vendruscolo, M. and Knapp, E.W.: *Proc. Natl. Acad. USA* **97** (2000), 3977–3981.