Validity of Gō Models: Comparison with a Solvent-Shielded Empirical Energy Decomposition

Emanuele Paci,*† Michele Vendruscolo,‡ and Martin Karplus*§

*ISIS, Laboratoire de Chimie Biophysique, Université Louis Pasteur, 67000 Strasbourg, France; †Biochemisches Institut der Universität Zürich, 8057 Zürich, Switzerland; ‡Cambridge University Chemical Laboratory, Cambridge CB2 1EW, United Kingdom; and §Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138 USA

ABSTRACT Do Gō-type model potentials provide a valid approach for studying protein folding? They have been widely used for this purpose because of their simplicity and the speed of simulations based on their use. The essential assumption in such models is that only contact interactions existing in the native state determine the energy surface of a polypeptide chain, even for non-native configurations sampled along folding trajectories. Here we use an all-atom molecular mechanics energy function to investigate the adequacy of Gō-type potentials. We show that, although the contact approximation is accurate, non-native contributions to the energy can be significant. The assumed relation between residue–residue interaction energies and the number of contacts between them is found to be only approximate. By contrast, individual residue energies correlate very well with the number of contacts. The results demonstrate that models based on the latter should give meaningful results (e.g., as used to interpret ϕ values), whereas those that depend on the former are only qualitative, at best.

INTRODUCTION

Protein folding is one of the essential reactions in living systems. Recently, attention has focused on this reaction not only because of its fundamental role (Fersht, 1999), but also because of the interest in protein folding generated by the availability of many protein sequences from a rapidly increasing number of genomes and the realization that misfolded proteins are involved in disease (Dobson, 1999, 2001). Considerable progress has been made in achieving an understanding of the folding reaction by the use of simplified models (Bryngelson et al., 1995; Chan and Dill, 1998; Dinner et al., 2000). In particular, the problem posed by the "Levinthal Paradox" (namely that a polypeptide chain can find its unique native structure in spite of the very large number of possible denatured conformations) has been solved. It has been shown that a reasonable energy bias toward the native state can reduce the search of conformation space sufficiently for folding to take place on the experimental time scale (Karplus, 1997). One model, referred to as the "Gō model" or "Gō-type model" (Taketomi et al., 1975; Takada, 1999), has been widely used in studies of protein folding (Zhou and Karplus, 1999; Alm and Baker, 1999; Muñoz and Eaton, 1999; Galzitskaya and Finkelstein, 1999; Ozkan et al., 2001; Vendruscolo et al., 2001; Shimada et al., 2001). It is characterized by an energy function that replaces the nonbonded interactions (van der Waals and electrostatic terms) by attractive native-state contact energies; in some cases, non-native repulsions are also present (Zhou and Karplus, 1999; Shimada et al., 2001). Applications of Gō-type models include one-dimensional residue-based phenomenological descriptions of the folding reaction (Alm and Baker, 1999; Muñoz and Eaton, 1999; Galzitskaya and Finkelstein, 1999), lattice model calculations (Ozkan et al. 2001), and three-dimensional C_α or all-atom folding simulations by molecular dynamics (Zhou and Karplus, 1999) and Monte Carlo methods (Shimada et al. 2001); in the latter an excluded volume term is added to prevent collapse of the structure. For the phenomenological descriptions (Alm and Baker, 1999; Muñoz and Eaton, 1999; Galzitskaya and Finkelstein, 1999), the Gō-type model provides an essential simplification, which makes possible the replacement of the three-dimensional structure of the protein by a one-dimensional construct. For simulations in three dimensions (lattice and off-lattice), Gō-type potentials have the important property that the native state is a deep minimum, and that the potential surface corresponding to the non-native configurations is relatively smooth. There results a nearly ideal folding "funnel" leading to the native state, in contrast to the much rougher energy surface obtained with a more realistic molecular mechanics potentials (Duan and Kollman, 1998). As a consequence, trajectories calculated with Gō-type potentials take only on the order of nanoseconds to fold, instead of the experimental timescale which is microseconds or longer. This has made it possible to obtain statistically meaningful results for generic Gō-type models of proteins and polypeptide chains, and for models with native structures corresponding to those of specific proteins (Zhou and Karplus, 1999; Vendruscolo et al., 2001; Shimada et al., 2001).

Given the widespread use of Gō-type models, it is surprising that no direct tests have been made to determine whether they provide an accurate description of the protein energy surface. In this paper, we make such a test by comparing Gō-type model results with those obtained from an extensively validated effective energy function (EEF1)

Submitted March 14, 2002, and accepted for publication May 1, 2002.

Address reprint requests to Martin Karplus, ISIS, Laboratoire de Chimie Biophysique, Université Louis Pasteur, 4 rue Blaise Pascal, 67000 Strasbourg, France. Tel.: +33-90-241560; Fax: +33-90-241562; E-mail: marci@tammy.harvard.edu

0006-3495/02/12/3032/07 $2.00

of the molecular-mechanics type, which combines a standard representation of the nonbonded van der Waals and electrostatic contributions to the energy with an implicit treatment of solvation (Lazaridis and Karplus, 1999). In an earlier paper (Paci et al., 2002b) we showed that the contact approximation with a cut-off radius of 5.5 Å for the EEF1 potential gives an excellent approximation to the total energy for native and non-native configurations of proteins, even when non-native interactions contribute significantly. It was also shown there that the inclusion of solven shielding is essential for the validity of the contact model. Here, we examine the validity of the most widely used form of the Gō-type model, which assumes a simple relation between the residue–residue interaction energies and the number of contacts. Results for the original Gō model (Taketomi et al. 1975), which uses a single parameter to relate a residue–residue contact to its interaction energy, are similar.

EEF1 and Gō-type models correspond to potentials of mean force; i.e., they represent the effective energy of the protein–solvent system for a given configuration of the protein in the presence of a canonically averaged solvent (Karplus and Shakhnovich, 1992). The EEF1 energy function can be decomposed into a sum of pairwise residue–residue interactions, E_{IJ}, where I and J correspond to the residues. For each geometry or for an ensemble of geometries, such as those representing the unfolding state (see Methods), we can write the EEF1 energy in the form

$$E(\mathrm{EEF1}) = \sum_{I} \sum_{J \geq I+N} E_{IJ}, \tag{1}$$

where we excluded N near-neighbor residue interactions (we use $N = 2$ in accord with several implementations of Gō-type models (Muñoz and Eaton, 1999; Vendruscolo et al., 2001; Shimada et al., 2001)); for $N \geq 2$, the bonded terms make no contribution, and E_{IJ} can be written (see Methods)

$$E_{IJ} = E_{IJ}^{\mathrm{cont}} + E_{IJ}^{\mathrm{non\text{-}cont}} = E_{IJ}^{\mathrm{cont,Go}} + E_{IJ}^{\mathrm{cont,non\text{-}Go}} + E_{IJ}^{\mathrm{non\text{-}cont}}. \tag{2}$$

In Eq. 2, cont (non-cont) refer to the fact that the residues are (are not) in contact in a given structure or ensemble of structures (two residues are assumed to be in contact if they have any pair of heavy atoms within 5.5 Å) and the symbols Gō (non-Gō) indicate that the contact between I and J exists (does not exist) in the native state. As is evident from Eqs. 1 and 2, the validity of a Gō-type model requires that, for any configuration of the protein,

$$E_{IJ} \simeq E_{IJ}^{\mathrm{cont}} \simeq E_{IJ}^{\mathrm{cont,Go}}, \tag{3}$$

that is, that only the native contact interactions contribute significantly to the effective energy for both native and non-native configurations or ensembles. The original form of the Gō model (Taketomi et al., 1975) assumes that

$$E_{IJ}^{\mathrm{Go}} = \epsilon \Delta_{IJ}^{\mathrm{Nat}}; \qquad E^{\mathrm{Go}} = \sum_{IJ} \tilde{E}_{IJ}^{\mathrm{Go}}, \tag{4}$$

where $\Delta_{IJ}^{\mathrm{Nat}}$ determines whether residues I and J, which make a contact in the native state, are in contact in the structure under consideration, and ϵ is a constant parameter. In the common implementation of Gō-type models, the assumption is made that the energy for each residue pair in a given structure is proportional to the number N_{IJ} of native heavy-atom contacts in that structure; i.e.,

$$E_{IJ}^{\mathrm{Go}} = \gamma N_{IJ}^{\mathrm{Nat}}; \qquad E^{\mathrm{Go}} = \sum_{IJ} E_{IJ}^{\mathrm{Go}} = \gamma \sum_{IJ} N_{IJ}^{\mathrm{Nat}}, \tag{5}$$

where γ is a proportionality constant determined by a fitting procedure (Muñoz and Eaton, 1999; Shimada et al., 2001).

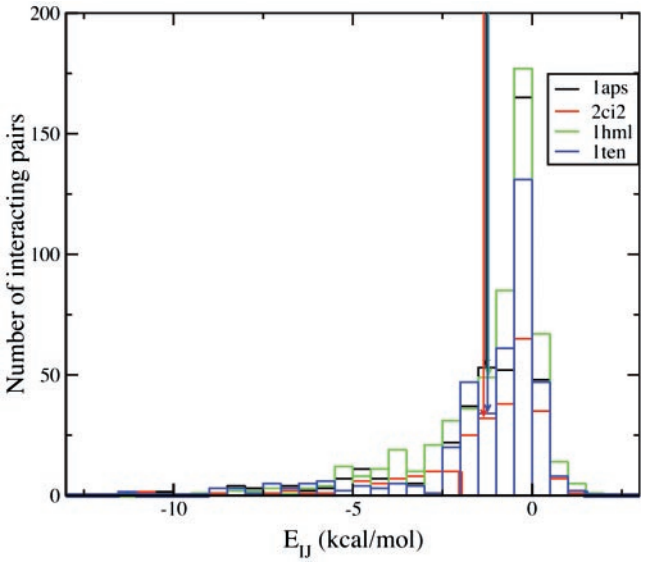

FIGURE 1 Histogram showing the number of residue pairs interacting with a given residue–residue energy. For clarity, results for only four proteins out of the eight studied are shown; their PDB code is 1aps (*black*), 2ci2 (*red*), 1hml (*green*), 1ten (*blue*). The arrows represent the average interaction energy between pairs of residues, corresponding to the parameter ϵ for each protein.

NATIVE-STATE ANALYSIS

Figure 1 shows the distribution of the interaction energy for the interacting residue pairs. It has a peak close to zero, and the average is -1.2 ± 1.8 kcal/mol for the eight proteins. It is evident that the contact energies cover a wide range and that the use of a single coefficient, as in Eq. 4, is a rough approximation.

Figure 2 *a* shows a scatter plot of the relationship between E_{IJ} as calculated for the native structure with EEF1 and the number of contacts between residues I and J. Data for four proteins are included in the figure (see caption); the

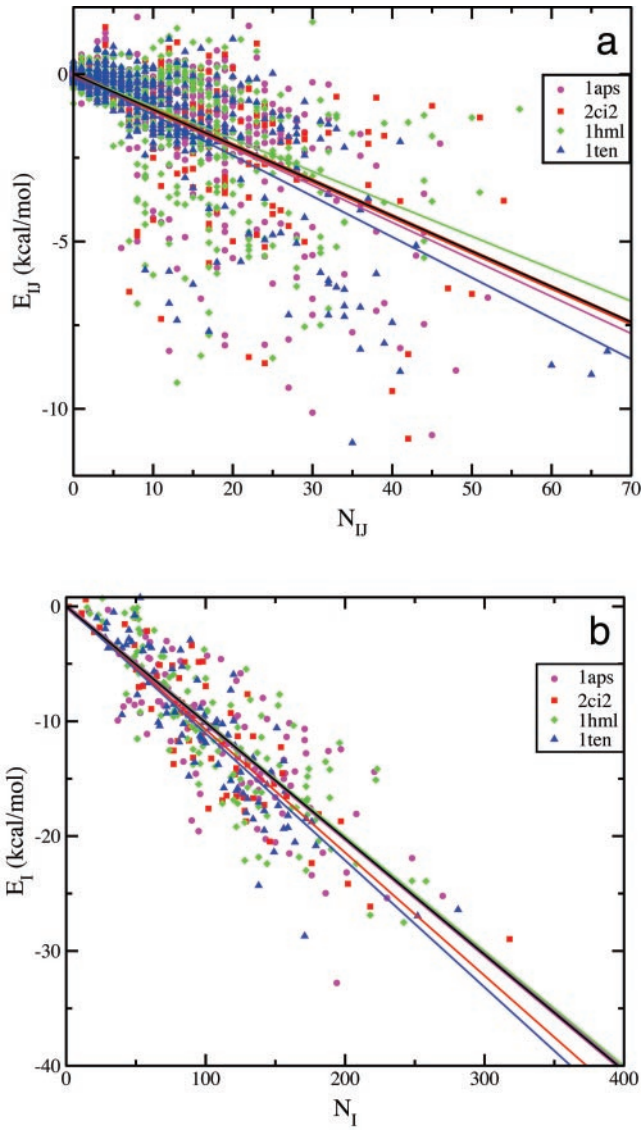

FIGURE 2 Comparison of EEF1 and Gō-type energies. (*a*) Energy between pairs of non-neighboring residues as a function of the number of the heavy atom contacts (with a cutoff of 5.5 Å) between them in the experimental structure. (*b*) Effective residue energy as a function of the numbers of all (heavy) atom contacts made by that residue. The continuous lines in (*a*) and (*b*) are the least-squares fit for the various proteins (*thin colored lines*) and for all the proteins together (*black heavy line*). Results are shown for the set of proteins used in Fig. 1.

lines represent the best (least-squares) fit for individual proteins and for all the proteins simultaneously. Although a qualitative relationship is evident, there is considerable scatter in the distribution.

Table 1 shows the calculated total energies obtained for the native states of eight proteins from the EEF1 energy function in the column headed E(EEF1) (Eq. 1). The next column, E^{cont}(EFF1), shows the result obtained with the contact approximation, $E_{\mathrm{IJ}}^{\mathrm{cont}}$ (Eq. 3), which is equivalent to the exact Gō-type energy for the native state. $E^{\mathrm{cont}}(\mathrm{EFF1}) = E^{\mathrm{Go}}(\mathrm{EFF1})$ is clearly a very good approximation to the true native state energy, E(EFF1). The energies obtained with Eq. 5 are in the next columns. Results obtained by fitting a parameter for each protein (γ^{P}) and with an average parameter for all the proteins ($\bar{\gamma}$) are shown; the values of γ^{P} and $\bar{\gamma}$, are given in Table 2; they correspond to the linear least square fits in Fig. 2a. Deviations from E(EEF1) are as large as 60 kcal/mol with γ^{P}; with $\bar{\gamma}$ significantly larger values occur. We can also introduce a parameter, $\gamma^{\mathrm{P}'}$, which is chosen so that Eq. 5 yields the native state (EEF1) energy for each protein; the values are also listed in Table 2. As can be seen, γ^{P} and $\gamma^{\mathrm{P}'}$ are very similar for each protein (within 0.01 kcal/mol). Nevertheless, because the number of residue–residue contacts is in the range 3000–7000 (as listed in Table 1), such small differences lead to large changes in the total energy. This provides a cautionary note on the use of such formulations.

In applications of Gō-type models to estimate the effect of single-site mutations on the native-state energy (Otzen et al., 1995; Xu et al., 1998; Cota et al., 2000), and for transition states (Vendruscolo et al., 2001), the quantity of interest is not the interaction energy of residue pairs, but rather the interaction energy per residue, $E_{\mathrm{I}}^{\mathrm{Go}}$. This is defined as the sum over all residues in contact with a given residue in the native state (i.e., $N_{\mathrm{I}}^{\mathrm{Nat}} = \sum_{\mathrm{J}} N_{\mathrm{IJ}}^{\mathrm{Nat}}$) so that, from Eq. 5,

$$E_{\mathrm{I}}^{\mathrm{Go}} = \Gamma \sum_{\mathrm{J}} N_{\mathrm{IJ}}^{\mathrm{Nat}}. \tag{6}$$

This energy is to be compared with the energy per residue obtained from EEF1, which is

$$E_{\mathrm{I}} = \sum_{\mathrm{J}} E_{\mathrm{IJ}}. \tag{7}$$

Figure 2 *b* shows a scatter plot of the results for the proteins included in Fig. 2 *a*. In analogy to Fig. 2 *a*, best fit lines for the individual proteins (Γ^{P}) and for all the proteins simultaneously ($\bar{\Gamma}$) are shown. The correspondence is significantly better than that in Fig. 2 *a* with a narrower distribution about the best fit lines. This result is expected from the central limit theorem, which shows that the standard deviation of the average (or sum) is smaller than the standard deviation of individual variables. The total energies and parameters Γ^{P} for the various proteins obtained from the fits to Eq. 6 are given in Tables 1 and 2. The values of Γ^{P} are very similar to γ^{P} and to $\gamma^{\mathrm{P}'}$ and, in most cases, are closer than γ^{P} to $\gamma^{\mathrm{P}'}$, the parameters that fit the total energy. This is in accord with the fact that $E^{\mathrm{Go}}(\Gamma^{\mathrm{P}})$ is a better approximation to E(EEF1) than is $E^{\mathrm{Go}}(\gamma^{\mathrm{P}})$ in most cases (see Table 1). This is an important result because it provides a justification for the use of Eq. 6 in the analysis of protein stability and transition state structures.

TABLE 1 EEF1 total, contact and Gō energies for the experimental structures of eight proteins (see Methods)

	M	N	N'	E(EEF1)	E^{Go}(EEF1)	$E^{Go}(\gamma^P)$	$E^{Go}(\bar{\gamma})$	$E^{Go}(\Gamma^P)$
1aye	78	343	4241	−468.2	−466.2	−489.8	−448.8	−471.5
2ci2	65	261	3312	−348.0	−352.8	−351.8	−350.5	−354.9
1tit	89	381	3754	−323.4	−321.0	−344.9	−397.3	−324.6
SUC1	96	413	5526	−525.8	−528.2	−548.8	−584.8	−528.1
1hml	123	537	6834	−688.4	−684.0	−662.3	−723.3	−683.5
1aey	58	247	3199	−315.8	−319.0	−340.4	−338.6	−311.6
1ten	89	396	4557	−491.9	−493.2	−554.2	−482.3	−503.8
1aps	98	459	5725	−587.9	−587.1	−633.8	−605.9	−574.3

M, Number of residues; N, number of interacting pairs of residues; N', number of interacting pairs of atoms; $E^{Go}(\gamma^P)$ and $E^{Go}(\bar{\gamma})$, energy computed using Eq. 5, where γ is estimated by a best fit to a linear relation between pairwise EEF1 energy E_{IJ} and the number of contacts N_{IJ} for each protein (γ^P) or for all the proteins ($\bar{\gamma}$); $E^{Go}(\Gamma^P)$, calculated using Eq. 6, where Γ^P is computed by a least square best fit to a linear dependence between the single residue EEF1 energy E_I and the total number of contacts N_I.
All energies are in kcal/mol.

NON-NATIVE INTERACTIONS

The determination of contribution of non-native contacts along folding pathways is important for an evaluation of Gō-type models. Because the transition states play an essential role in protein folding, we consider them first and then examine other portions of the energy surface.

Transition states

In Table 3, we show the results for the transition state ensembles (TSE) determined by constraining the calculated ϕ values to be equal to the experimental ones, using molecular dynamics and the EEF1 potential (see Methods). The values in Table 3 are obtained by considering a single representative structure of the TSE; corresponding results are obtained if averages over the TSE are made. The structures in the ensembles have a root mean square deviation (RMSD) in the range 4−6 Å from the native state. The first two columns give the EEF1 energy, E(EEF1), used as a reference, and the contact energy calculated with EEF1, E^{cont}(EEF1). As can be seen, the agreement is as good as it is for the native state (Table 1). The energies calculated using only the native contacts, E^{Go}(EEF1) are, in all cases, less negative than the true energies. The non-native contribution, $E^{non\text{-}Go}$(EEF1), also given in the Table, is in the range of −69 to −122 kcal/mol. This shows that non-native contacts contribute significantly to stabilizing the transition state. Table 3 also lists as $E^{Go}(\gamma^P)$ and $E^{non\text{-}Go}(\gamma^P)$, the corresponding results obtained using Eq. 5 with γ^P, the best fit of the energy parameter to the native state for each protein. There are significant deviations of $E^{Go}(\gamma^P)$ from E^{Go}(EEF1), in addition to the errors in the latter. The deviations are both positive and negative (between −32 and +60 kcal/mol); in comparison to E^{cont}(EEF1), the differences are all positive, as for E^{Go}(EEF1).

Thus, use of the Gō model, whether in the form of E^{Go}(EEF1) or $E^{Go}(\gamma^P)$, results in a deeper well for the native state relative to the transition state than do the actual energy values; e.g., for procarboxypeptidase A2 (1aye), the EEF1 transition state energy is 50.5 kcal/mol above the native state, whereas it is calculated to be 124.4 kcal/mol and 92.9 kcal/mol with E^{Go}(EEF1) and $E^{Go}(\gamma^P)$, respectively. The same is also true relative to the denatured state, in correspondence with the unrealistically deep funnel-like structure of the energy surface obtained with the Gō-type potential, already mentioned in the Introduction.

TABLE 2 Values of γ^P and Γ^P used to compute the energies given in Table 1

	γ^P	$\gamma^{P'}$	Γ^P
1aye	−0.1155	−0.1099	−0.1112
2ci2	−0.1062	−0.1065	−0.1072
1tit	−0.0919	−0.0855	−0.0865
SUC1	−0.0993	−0.0956	−0.0956
1hml	−0.0969	−0.1001	−0.1000
1aey	−0.1064	−0.0997	−0.0974
1ten	−0.1216	−0.1082	−0.1105
1aps	−0.1107	−0.1026	−0.1003
ALL	−0.1058	—	−0.1010

For comparison, the value $\gamma^{P'}$ chosen so that Eq. 5 yields the native state (EEF1) energy for each protein is included (see text).
Units are kcal/mol.

Thermally induced non-native conformations

Table 4 shows results corresponding to those in Table 3 for a set of non-native conformations of CI2 obtained by an unfolding simulation at 450 K, followed by quenching simulation at 300 K (see Methods). The contact approximation is valid for these non-native states, as it is valid for native and transition states. However, the various Gō-type models have errors that increase with the RMSD from the native state; the error in E^{Go}(EEF1) and $E^{Go}(\gamma^P)$ are similar. For the largest RMSD analyzed (11.7 Å) the stabilization arising from non-native contacts, $E^{non\text{-}Go}$(EEF1) and $E^{non\text{-}Go}(\gamma^P)$ is larger than that from the native (Gō-type) contacts. This is because there are more non-native (128) than native (76) contacts, as shown in Table 4.

TABLE 3 Energies for the transition states of eight proteins (see Methods)

	RMSD	E(EEF1)	E^{cont}(EFF1)	E^{Go}(EFF1)	$E^{non\text{-}Go}$(EFF1)	$E^{Go}(\gamma^P)$	$E^{non\text{-}Go}(\gamma^P)$
1aye	4.8	−417.7	−415.7	−343.8	−71.9	−375.3	−47.8
2ci2	5.6	−299.6	−301.2	−224.9	−76.3	−228.2	−48.4
1tit	4.0	−469.0	−471.4	−385.9	−85.5	−325.8	−49.0
SUC1	5.3	−531.0	−534.0	−412.5	−121.5	−395.4	−83.9
1aey	4.0	−321.2	−324.1	−255.0	−69.1	−234.6	−34.4
1ten	4.4	−465.6	−469.3	−390.0	−79.3	−411.3	−67.3
1aps	4.7	−590.1	−592.2	−471.2	−121.0	−423.3	−84.4

For definitions, see Table 2.
$E^{non\text{-}Go}(\gamma^P)$ is computed using the number of non-native all-atom contacts and the values of γ^P given in Table 2. Units are kcal/mol.

In some simulations (Shimada et al., 2001), it has been assumed that non-native contacts are repulsive, which leads to faster folding than the standard Gō models, which include only native contacts. This is not in agreement with the EEF1 decomposition, i.e., the non-native constants are overall attractive, as shown in Tables 3 and 4. The Gō-type potential, used in the folding study of a three-helix bundle protein (Zhou and Karplus, 1999), varied the non-native interactions over a range that included both attractive and repulsive values. A non-native repulsive interaction somewhat weaker than the corresponding native attractive interaction (a ratio of ~0.4) gave a folding time closest to the experimental value.

VALIDITY OF Gō MODELS

The Gō model was proposed in 1975 as an ingenious technical device to make possible computer simulations of protein folding (Taketomi et al., 1975). As such, it has been very successful (Takada, 1999). Now the Gō model is being used not only as a convenient computational tool but also because it is thought by some to capture what they believe to be an essential element (i.e., a smooth deep funnel) of protein energy "landscapes." But does it really do this? Do smooth deep funnels characterize protein energy surfaces?

A comparison between an all-atom molecular-mechanics effective-energy function with solvent shielding and the Gō-type potentials in current use provides a test of these fundamental questions. It is expected that other effective potentials that are pairwise decomposable would give similar results. The practical impossibility of doing this type of analysis with explicit solvent models should be noted. The results obtained show that the most commonly used Gō-type model is only an approximate description of the protein potential energy surfaces for the native state, the transition state, and along folding trajectories. Nevertheless, some global features obtained with Gō-type models are likely to be meaningful. Of particular importance is the demonstration that individual residue energies (rather than residue–residue interaction energies) are rather well described by Gō-type models. This explains the correlations observed for single-site mutations in the native state (Otzen et al., 1995; Xu et al., 1998; Cota et al., 2000) and provides a justification for the determination of the coarse-grained structure of transition-state ensembles based on such models (Vendruscolo et al., 2001; Paci et al., 2002a). Gō-type models are most accurate for transition states that are relatively close in structure to the native state, as they appear to be for many fast-folding small proteins (Li and Daggett, 1994; Vendruscolo et al., 2001). For other portions of the potential energy surface, particularly collapsed misfolded states, the non-native contacts neglected in Gō-type models make important contribution and can lead to significant distortions of the potential energy surface. The present analysis also demonstrates that Gō-type models result in a much deeper well for the native state than do more realistic potentials, due primarily to the neglect of stabilizing non-native contacts. This is an essential element in the simple, fast-folding behavior obtained in molecular dynamics and Monte Carlo calculations based on Gō-type models. Thus, just the elements of the Gō-type models that make them so attractive for folding simulations introduce errors that have to be

TABLE 4 Energies for certain non-native conformations of CI2

RMSD	n_1	n_2	E(EFF1)	E^{cont}(EFF1)	E^{Go}(EFF1)	$E^{non\text{-}Go}$(EFF1)	$E^{Go}(\gamma^P)$	$E^{non\text{-}Go}(\gamma^P)$
0.0	261	0	−348.0	−352.8	−352.8	0.0	−351.7	0.0
3.3	223	29	−339.3	−341.7	−313.6	−28.1	−297.5	−16.9
5.2	167	65	−300.3	−301.0	−240.4	−60.6	−230.0	−50.0
7.8	148	95	−324.9	−329.4	−215.6	−113.8	−201.1	−104.2
11.7	76	128	−266.4	−272.4	−115.1	−157.3	−104.8	−163.0

See text and Tables 1 and 3 for definitions.
n_1 (n_2) is the number of native (non-native) interactions for each configuration.
Energies are in kcal/mol.

considered in evaluating the significance of the quantitative folding result for proteins obtained from such model calculations. Moreover, the neglect of attractive non-native interactions makes it impossible to use Gō-type potentials in the study of misfolding (e.g., the production of fibrils, which appear to be important in certain diseases (Dobson, 1999, 2001)).

METHODS

We use a molecular mechanics potential energy function (EEF1) for the atoms with an implicit solvent term. EEF1 is based on the CHARMM19 polar hydrogen representation (Neria et al., 1996) with a Gaussian model for solvation (Lazaridis and Karplus, 1999). The function, called EEF1, has been used in a variety of applications concerned with the protein folding reaction (Lazaridis and Karplus, 1999), including the high-temperature unfolding of the protein CI2 (Lazaridis and Karplus, 1997), where good agreement was obtained with simulations that used an explicit representation of the solvent (Li and Daggett, 1994).

The effective energy, $E^{\mathrm{EEF1}}(\mathbf{R})$, of a protein with conformation **R** includes the protein internal energy and the solvation free energy. Both can be written as a sum over all residue pairs. Details are given in Lazaridis and Karplus (1999).

Proteins and conformations used for analysis

Eight proteins were used in the analysis. They are acylphosphatase (PDB entry1aps (Pastore et al., 1992)), chymotrypsin inhibitor 2 or CI2 (PDB entry 2ci2) (McPhalen and James, 1987), α-spectrin SRC 3 domain (Blanco et al., 1997) (PDB entry 1aey), the third fibronectin type III repeat from tenascin (Leahy et al., 1992) (PDB entry 1ten), α-LA (Ren et al., 1993) (PDB entry 1hml), procarboxypeptidase A2 (Garcia-Saez et al., 1997) (PDB entry 1aye), an immunoglobulin-like modules from titin I-band (Improta et al., 1996) (PDB entry 1tit), the cell-cycle regulatory protein p13suc1, SUC1 (Endicott et al., 1995). The experimental structure was, in all cases, minimized for 200 steepest descent steps to eliminate bad contacts.

Several types of non-native structures of interest for the understanding of protein folding and unfolding were examined. Transition state ensembles were obtained using an approach based on experimental ϕ values to bias the trajectory (Vendruscolo et al., 2001; Paci et al., 2002a) toward conformations where the fraction of native contacts equals the experimental ϕ value for those residues for which such value has been measured. For CI2, high-temperature unfolded states were obtained by increasing the temperature of the Nosé–Hoover thermostat to 450 K during the simulation of over 1 ns or longer. Collapsed configurations were generated from the high temperature conformations by decreasing the temperature to 300 K over 200-ps trajectories.

E.P. acknowledges financial support from Forschungskredit der Universität Zürich. M.V. is a Royal Society University Research Fellow. The research of M.K. is supported in part by the Centre National de la Recherche Scientifique (ESA 7006), by the Ministère de l'Education Nationale, de la Recherche et de la Technologie (Strasbourg), and by a grant from the National Institutes of Health (Harvard).

REFERENCES

Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11305–11310.

Blanco, F. J., A. R. Ortiz, and L. Serrano. 1997. 1H and 15N NMR assignment and solution structure of the SH3 domain of spectrin: comparison of unrefined and refined structure sets with the crystal structure. *J. Biomol. NMR.* 9:347–357.

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins.* 21:167–195.

Chan, H. S., and K. A. Dill. 1998. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins.* 30:2–33.

Cota, E., S. J. Hamill, S. B. Fowler, and J. Clarke. 2000. Two proteins with the same structure respond very differently to mutation: the role of plasticity in protein stability. *J. Mol. Biol.* 302:713–725.

Dinner, A. R., A. Sali, L. J. Smith, C. M. Dobson, and M. Karplus. 2000. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25:331–339.

Dobson, C. M. 1999. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* 24:329–332.

Dobson, C. M. 2001. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B.* 356:133–145.

Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science.* 282:740–744.

Endicott, J. A., M. E. Noble, E. F. Garman, N. Brown, B. Rasmussen, P. Nurse, and L. N. Johnson. 1995. The crystal structure of p13suc1, a p34cdc2-interacting cell cycle control protein. *EMBO J.* 14:1004–1014.

Fersht, A. R. 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman & Co., New York.

Galzitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11299–11304.

Garcia-Saez, I., D. Reverter, J. Vendrell, F. X. Avilés, and M. Coll. 1997. The three-dimensional structure of human procarboxypeptidase A2. Deciphering the basis of the inhibition, activation and intrinsic activity of the zymogen. *EMBO J.* 16:6906–6913.

Improta, S., A. S. Politou, and A. Pastore. 1996. Immunoglobulin-like modules from titin 1-band: Extensible components of muscle elasticity. *Structure.* 4:323–337.

Karplus, M. 1997. The Levinthal paradox: yesterday and today. *Fold. Des.* 2:S69–S75.

Karplus, M., and E. Shakhnovich. 1992. Protein folding: theoretical studies of thermodynamics and dynamics. *In* Protein Folding. T. E. Creighton, editors. W. H. Freeman & Co., New York. 127–195.

Lazaridis, T., and M. Karplus. 1997. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science.* 278:1928–1931.

Lazaridis, T., and M. Karplus. 1999. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–487.

Leahy, D. J., W. A. Hendrickson, I. Aukhil, and H. P. Erickson. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science.* 258:987–991.

Li, A., and V. Daggett. 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U.S.A.* 91:10430–10434.

McPhalen, C. A., and M. N. James. 1987. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. *Biochemistry.* 26:261–269.

Muñoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. U.S.A.* 96:11311–11316.

Neria, E., S. Fischer, and M. Karplus. 1996. Simulation of activation free energies in molecular dynamics system. *J. Chem. Phys.* 105:1902–1921.

Otzen, D. E., M. Rheinnecker, and A. R. Fersht. 1995. Structural factors contributing to the hydrophobic effect: the partly exposed hydrophobic minicore in chymotrypsin inhibitor 2. *Biochemistry.* 34:13051–13058.

Ozkan, S. B., I. Bahar, and K. A. Dill. 2001. Transition states and the meaning of Φ-values in protein folding kinetics. *Nature Struct. Biol.* 8:765–769.

Paci, E., M. Vendruscolo, and M. Karplus. 2002a. Determination of a transition state at atomic resolution from protein engineering data. In press.

Paci, E., M. Vendruscolo, and M. Karplus. 2002b. Native and non-native interactions along protein folding and unfolding pathways. *Proteins.* 47:379–392.

Pastore, A., V. Saudek, G. Ramponi, and R. J. Williams. 1992. Three-dimensional structure of acylphosphatase. Refinement and structure analysis. *J. Mol. Biol.* 224:427–440.

Ren, J., K. R. Acharya, and D. I. Stuart. 1993. α-lactalbumin possesses a distinct zinc binding site. *J. Biol. Chem.* 268:19292–19298.

Shimada, J., E. L. Kussell, and E. I. Shakhnovich. 2001. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. *J. Mol. Biol.* 308:79–95.

Takada, S. 1999. Gō-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* 96:11698–11700.

Taketomi, H., Y. Ueda, and N. Gō. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* 7:445–459.

Vendruscolo, M., E. Paci, C. M. Dobson, and M. Karplus. 2001. Three key residues form a critical contact network in a protein folding transition state. *Nature.* 409:641–645.

Xu, J., W. A. Baase, E. Baldwin, and B. W. Matthews. 1998. The response of T4 lysozyme to large-to-small substitutions within the core and its relation to the hydrophobic effect. *Prot. Sci.* 7:158–177.

Zhou, Y., and M. Karplus. 1999. Interpreting the folding kinetics of helical proteins. *Nature.* 401:400–403.