Native and non-native interactions along protein folding and unfolding pathways

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MAY 2002

Impact Factor: 2.63 · DOI: 10.1002/prot.10089 · Source: PubMed

3 AUTHORS, INCLUDING:

Emanuele Paci
University of Leeds
124 PUBLICATIONS **6,436** CITATIONS
SEE PROFILE

Michele Vendruscolo
University of Cambridge
378 PUBLICATIONS **12,622** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Genetics 47:379–392 (2002)

Native and Non-Native Interactions Along Protein Folding and Unfolding Pathways

Emanuele Paci,[1,2] Michele Vendruscolo,[1] and Martin Karplus[2,3]*
[1]*Oxford Centre for Molecular Sciences, Central Chemistry Laboratory, University of Oxford, Oxford United Kingdom*
[2]*Laboratoire de Chimie Biophysique, Institut Le Bel, Université Louis Pasteur, Strasbourg, France*
[3]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts*

ABSTRACT **Gō-type models, which include only native contact interactions in the energy function, are being used increasingly to describe the protein folding reaction. To investigate the validity of such models, we determine the role of native and non-native interactions along folding and unfolding pathways. For this purpose, we use a molecular mechanics energy function with an implicit solvation model (an effective energy function or potential of mean force) that can be expressed in a pairwise decomposable form. We find that for the native state and a wide range of other configurations, the contact energy is an accurate description, in part due to the cancellation of non-zero contributions from more distant residues. However, significant errors in the energy are introduced for non-native structures if the energy is calculated from the native contacts alone. Non-native contacts tend to make a significant contribution, particularly for molten globules and collapsed states along the unfolding pathways. The implication of these results for the use of Gō-type models in studies of protein folding are discussed. Proteins 2002;47:379–392.**

Key words: Gō models; contact energy; effective energy function; residue-residue interactions

INTRODUCTION

An understanding of protein folding and misfolding is of great importance for the use of genomic data, for combating diseases involving protein aggregates, and because of the fundamental role of the folding reaction in structural biology.[1] The direct approach to folding a protein by computer simulations with an all-atom representation of the potential function and explicit solvent is not yet possible,[2] although the folding of peptides that form α-helices, β-hairpins, and double hairpin constructs has been performed successfully with explicit[3] and implicit[4–6] representations of the solvent. Attempts to overcome the problem raised by the time scale of the folding process (μs or longer) by use of distributed computing on a large number of processors are of interest in this regard.[7] However, even when this type of brute force approach to proteins folding is successful, the complexity of the resulting trajectories will continue to require a simplified model for their interpretation.

One type of model that has been widely used to obtain information about the general mechanism of folding treats proteins in terms of a C_α ("bead") representation of each amino acid residue of the polypeptide chain and simplifies the interactions to contact potentials on a lattice[8] or to square-well[9] and Lennard-Jones potentials[10,11] in off-lattice molecular dynamics simulations. Such models have been particularly useful because the calculations are fast enough to make possible the determination of the large numbers of folding trajectories necessary to obtain statistically meaningful results. Simulations with such models,[12] as well as analytical approaches,[13–16] have provided an answer to the general problem raised by the Levinthal paradox (namely, that a polypeptide chain can find its unique native state in spite of the very large number of possible denatured conformations) by demonstrating that a reasonable energy bias toward the native state can reduce the required search of conformation space sufficiently so that folding can take place on the experimental time scale.[17]

Recently a number of studies have been made that apply simplified models to specific proteins (see, for example, Shea et al.[11]). The most widely used have been what are referred to as "Gō-type" models based on the pioneering work of Gō and coworkers[18] (see also Gō[19] and Takada[20]) using lattice models. The calculations have concerned primarily physical off-lattice C_α representations[9,21–23] or one-dimensional (helix-coil type) representations of the polypeptide chain.[24,25] The essential assumption in a Gō-type model is that the only interactions that contribute to the energy function, other than a repulsive (excluded-volume-type) term between C_α atoms in the off-lattice models, are the attractive interactions between "neighbors" in the native structure. Although the specific form of the potential used is variable, as is the definition of "neighbor," the essential point is there are no attractions between atoms that are not neighbors in the native state. This type of potential clearly yields the native state as the one of lowest energy, and the potential can be parametrized so that the native state free energy is so low that the thermodynamic criterion for the stability of a protein is

*Correspondence to: Martin Karplus, Université Louis Pasteur, Laboratoire de Chimie Biophysique, ISIS, 4 rue Blaise Pascal, 67000 Strasbourg, France. E-mail: marci@brel.u-strasbg.fr

Received 14 June 2001; Accepted 6 December 2001

satisfied.[26] Moreover, it has been demonstrated that such models can introduce the energy bias necessary to solve the Levinthal paradox and that folding can be achieved in the off-lattice models by Monte Carlo or molecular dynamics approaches on a time scale that is easily accessible with present-day computers. A Gō-type model with an all-atom representation has been used in multiple folding simulations of crambin.[27]

Since the known structure of the native state is used to determine the potential, applications to specific proteins can be made. Examples include the use of Gō-type models in off-lattice representations to search for transition states, either by direct methods[22] or with biases based on experimental (ϕ value) data.[23] The helix-coil type models have been employed to correlate the experimental relative rates of folding of a series of two-state proteins[24,25] and to justify the relation between folding rate and contact order.[28,29]

The exact significance of some of these applications of Gō-type potentials is not yet clear. However, their increasing use makes it essential to determine whether such potentials provide a realistic representation of the energy surface of a protein. In this report, we use an all-atom potential for the protein[30] with an implicit representation of the solvent[31] to examine this question. The potential function we use, referred to as EEF1, has been shown to yield a sequence of events for high temperature unfolding[32] comparable to that obtained by explicit solvent simulations[33] and to discriminate native structures from misfolded decoys.[34] Importantly, both for its speed of execution and for the present analysis, EEF1 permits a pairwise decomposition of the nonbonded interaction energy of the atoms; the bonded energy terms can also be so decomposed on a residue basis. EEF1, like the Gō-type potentials, is a potential of mean force so that it provides the effective energy of the protein in solution; i.e., the energy of the solvent-protein system for a given configuration of the protein, averaged over the solvent degrees of freedom. Simulations with an explicit representation of the solvent cannot be analyzed in such a simple way because the solvation free energy is very difficult to calculate.

We first determine the contribution of native and nonnative interactions to the stability of the native state. For this case, a contact (Gō-type) potential with any reasonable neighbor criterion is found to be an excellent approximation because pairs not in contact make negligible contributions to the total energy. We then consider nonnative states, including those corresponding to the transition state ensemble and those accessed along the unfolding pathway at elevated temperature[32] or in the presence of external forces,[35–37] as well as for collapsed states from thermal unfolding pathways and molten globule states. We find that non-native contacts can make a sizable contribution to the effective stabilizing energy, so that Gō-type models are less satisfactory. In a separate paper, we examine the Gō-type approximation that the interaction energy between residues is simply related to the number of native contacts in native and non-native states, as for instance in Muñoz and Eaton,[24] and the determination of residue-residue contact interactions from statistical data for native proteins,[38–40] a widely used approach in protein structure prediction or threading.

We outline first the methods used in the effective energy decomposition and then present the results. A concluding section discusses the relevance of the results to an evaluation of the simplified Gō-type models for studies of protein folding.

METHODS

To analyze contact energies and Gō-type models for describing the native and non-native states of proteins, we use a molecular mechanics potential energy function for the atoms with an implicit solvent term. The function, called EEF1, is based on the CHARMM19 polar hydrogen representation[30] with a Gaussian model for solvation.[34] We describe first the decomposition of this function to obtain a pairwise representation of the energy that can be used for the present analysis. Then, we outline the various type of approximate models based on EEF1.

Pairwise Decomposition of EEF1

To show that the EEF1 effective energy, $E_{\text{eff}}(\mathbf{R})$, of a protein with conformation $\mathbf{R}$ can be written as the sum over all pairs of residues, we write[31]

$$E_{\text{eff}}(\mathbf{R}) = E_{\text{prot}}(\mathbf{R}) + \Delta G^{\text{solv}}(\mathbf{R}) \tag{1}$$

where $E_{\text{prot}}(\mathbf{R})$ is the solute (protein) energy and $\Delta G^{\text{solv}}(\mathbf{R})$ is the solvation free energy consisting of the protein-solvent and the solvent-solvent terms. The degrees of freedom of the solvent have been eliminated in that $E_{\text{eff}}(\mathbf{R})$ corresponds to a potential of mean force; i.e., it is the effective energy of the solute with configurations $\mathbf{R}$, in the presence of the canonically averaged free energy of the solvent at 300 K.[26] In using $E_{\text{eff}}(\mathbf{R})$, the temperature dependence of $\Delta G^{\text{solv}}(\mathbf{R})$ is usually neglected, although it can be introduced by the use of a temperature-dependent parameter set.[31,41]

The contribution $E_{\text{prot}}(\mathbf{R})$ is a sum of terms, which can be written in the form

$$\begin{aligned} E_{\text{prot}}(\mathbf{R}) &= \sum_i \sum_{j \neq i} H_{ij}(r_{ij}) + H^{\text{bonded}} \\ &= \sum_i \sum_{j \neq i} [E_{ij}^{\text{elec}}(r_{ij}) + E_{ij}^{\text{vdW}}(r_{ij})] + H^{\text{bonded}} \end{aligned} \tag{2}$$

where H^{bonded} includes the bond, angle and dihedral angle terms in the energy function for atoms separated, respectively, by one, two, or three covalent bonds.[42] The $H_{ij}(r_{ij})$ in Eqn 2 are the electrostatic ($E_{ij}^{\text{elec}}(r_{ij})$) and van der Waals ($E_{ij}^{\text{vdW}}(r_{ij})$) interaction between protein atoms i and j; in EEF1, the electrostatic terms is chosen to correspond to the Coulomb interaction with a distance-dependent dielectric factor ($\epsilon_{ij} = r_{ij}$).

The solvation free energy, $\Delta G^{\text{solv}}(\mathbf{R})$, in EEF1, is written as a sum of atomic terms

$$\begin{aligned} \Delta G^{\text{solv}} &= \sum_i \Delta G_i^{\text{ref}} + \sum_i \sum_{j \neq i} G_{ij}(r_{ij}) \\ &= \sum_i \left(\Delta G_i^{\text{ref}} - \sum_j \int_{V_j} d\mathbf{r} f_i(\mathbf{r}) \right) \simeq \sum_i \left(\Delta G_i^{\text{ref}} - \sum_{j \neq i} f_i(r_{ij}) V_j \right) \end{aligned} \tag{3}$$

where ΔG_i^{ref} (the reference solvation free energy) is the solvation free energy of the atom/group i in a small molecule in which the group itself is fully exposed and the term $G_{ij}(r_{ij})$ is the intramolecular solvent shielding that takes a pairwise form with $\int_{V_j} d\mathbf{r} f_i(\mathbf{r})$ corresponding to the reduction in solvation of atom i due to the presence of the surrounding group j; the integral over the solvation free energy $f_i(r)$ is approximated by a summation over a Gaussian function dependent on the atom i times the atomic volumes V_j. Both ΔG_i^{ref} and $\sum_{j \neq i} f_i(r_{ij})V_j$ include electrostatic and hydrophobic solvation contributions. In what follows, we subtract the term $\Delta G^{\text{ref}} = \sum_i \Delta G_i^{\text{ref}}$ since it depends only on the composition and not on the conformation $\mathbf{R}$ of the molecule and so is a constant for a given protein; that is, we use $E'_{\text{eff}}(\mathbf{R}) = E_{\text{eff}}(\mathbf{R}) - \Delta G^{\text{ref}}$ in the analysis. We drop the prime superscript in what follows for simplicity.

The total energy can be expressed as a sum over all pairs of *residues,*

$$E_{\text{eff}}(\mathbf{R}) = \sum_{J \geq I} E_{IJ}(\mathbf{r}_I, \mathbf{r}_J) \tag{4}$$

where capital indexes run over the residues, $\mathbf{r}_I$ is the set of coordinates of the atoms of residue I, and

$$E_{IJ}(\mathbf{r}_I, \mathbf{r}_J) = \sum_{i \in I, j \in J} [H_{ij}(r_{ij}) + \Delta G_{ij}(r_{ij})] + H_{IJ}^{\text{bonded}} \tag{5}$$

The sum over i and j runs over all the pairs of atoms of residues I and J, where a residue is defined as the mainchain atoms N, C_α, C, O, plus the sidechain; the sums over bonds, angles, and dihedral angles in H_{IJ} run over all the bonds, angles, and dihedrals that involve atoms of residues I and J. From the form of the bonded energy terms, those that occur in residues I and J include no other residues; i.e., on a residue basis the bonded terms are pairwise additive.

It is convenient for the subsequent analysis to define the partial effective energy of a protein, which includes only interactions that arise from residues that are more than a certain distance along the sequence; i.e.,

$$E^N(\mathbf{R}) = \sum_I \sum_{J \geq I+N} E_{IJ}(\mathbf{r}_I, \mathbf{r}_J) \tag{6}$$

where $E^N(\mathbf{R})$ leaves out the contribution from residues that are less than N residues apart. Existing formulations of Gō-type models have used $N = 2$ to 4 (in references 23, 24 and 27, $N = 2$ was used) on the assumption that the contributions for smaller N are not strongly dependent on the conformation or to emphasize non-local contributions.

Gō-Type Analysis and Definition of Contacts

In this section, we describe the various models that we are going to examine. It should be noted that the atom-atom cut-off used in EEF1 is 9 Å with a smooth switching function[43] between 7 and 9 Å.

Definition of Contact Energy

The contact approximation to the effective energy for a given conformation $\mathbf{R}$ can be written

$$E^N_{\text{cont}}(\mathbf{R}) = \sum_{J \geq I+N} E_{IJ}\Delta_{IJ} \tag{7}$$

where the factor Δ_{IJ}, is the contact matrix. It is 1 when residues I and J are in contact and 0 otherwise; the definition of "contact" depends on the model. It should be noted that Eqn 7 is applicable to any structure, native or not.

Contact matrices Δ_{IJ} can be chosen based on *all heavy atoms* or on C_α *atoms,* in correspondence with the Gō-model that have been described in the literature. In the all-heavy-atom definition, two residues are considered to be in contact if they have any pair of non-hydrogen atoms within a certain cut-off distance. For the C_α atom definition, two residues are considered in contact if their C_α atoms are within the contact distance. The value of the cut-off distance, r_c, is a parameter of the model. Only the all-atom analysis is described in the main text; some C_α results are given in the Appendix.

Definition of Gō and non-Gō Energy Contributions

The contact energy $E^N_{\text{cont}}(\mathbf{R})$ defined in Eqn 7 can be written as

$$E^N_{\text{cont}}(\mathbf{R}) = \sum_{J \geq I+N} E_{IJ}\Delta_{IJ}[\Delta^{NS}_{IJ} + (1 - \Delta^{NS}_{IJ})] = \sum_{J \geq I+N} E_{IJ}\Delta_{IJ}\Delta^{NS}_{IJ} + \sum_{J \geq I+N} E_{IJ}\Delta_{IJ}(1 - \Delta^{NS}_{IJ}) \tag{8}$$

where Δ^{NS}_{IJ} is the contact matrix for the native state, usually that given by an X-ray or NMR structure. The first summation of the right-hand-most equation is a generalization of the Gō energy (see below) and the second term is the correction to the Gō energy; the latter corresponds to the pairwise contact energy that arises from non-native contacts in configuration $\mathbf{R}$. It is referred to as the *non-native* contact energy, or sometimes as the *non-Gō* energy; we use this definition in what follows. Non-contact and non-native pair energies can contribute in non-native states, such as in molten globules, transition states, and, more generally, conformers that occur during the folding and unfolding reaction. Justification of Gō-type models for studying non-native conformations requires that *both* the non-contact and non-native contact energy contributions are negligible; that is

$$E^N_{\text{eff}}(\mathbf{R}) \simeq E^N_{\text{cont}}(\mathbf{R}) \simeq E^N_{\text{Go}}(\mathbf{R}) \tag{9}$$

The central purpose of this paper is to examine the validity of the approximate equalities in Eqn (9). If E^N_{cont} differs significantly from E^N_{eff}, no contact model (Gō or otherwise) is generally valid; if $E^N_{\text{eff}}(\mathbf{R}) \simeq E^N_{\text{cont}}(\mathbf{R})$, it is still maybe true that there are significant contribution from non-native contacts to the contact energy. Thus, we have to determine also whether $E^N_{\text{cont}}(\mathbf{R}) \simeq E^N_{\text{Go}}(\mathbf{R})$.

The original Gō model[18] corresponds to a particular choice of E_{IJ} in Eqn (7), namely

$$E_{IJ} = -\epsilon\Delta^{NS}_{IJ} \tag{10}$$

where ϵ is a well-depth parameter for residues in contact in the native state independent of the residue type. Given

Eqn 10, the Gō energy of a protein can be written as a sum over residue pairs

$$E_{\mathrm{Go}}(\mathbf{R}) = -\epsilon \sum_{I} \sum_{J \geq I} \Delta_{IJ}^{NS} \Delta_{IJ}(\mathbf{R}) \qquad (11)$$

This model and its extension (see, for example, Muñoz and Eaton[24]) will be examined in a separate paper (EP, MV, MK, in preparation).

If the native state is represented by a single conformation, such as the X-ray structure, its non-Gō energy is identically zero by definition. However, especially for non-native states (e.g., the transition state) it is generally appropriate to consider an ensemble of structures. For comparison with such ensembles, the most appropriate choice of "native contacts" is that obtained from a native state simulation. To extend the equations of this section to ensembles of structures (e.g., the native state or the transition state ensemble), the contact matrices and the contact energies for such ensembles have to be defined. To obtain the "native" contacts, the average distance $\langle r_{IJ} \rangle_{NSE}$ between each pair of residues (I, J) was computed for the native state ensemble and $\Delta_{IJ}^{NS} = \langle \Delta_{IJ} \rangle_{NSE}$ was set equal to 1 or 0 if $\langle r_{IJ} \rangle_{NSE}$ was smaller or larger than r_c. Correspondingly, the pairwise energy for an ensemble, i.e., a "macrostate," or a "state" for short was taken to be equal to the average energy $\langle E_{IJ} \rangle$ of the conformations representing the state. The contact energy for the state is given by $E_{\mathrm{cont}}^{N} = \Sigma_{J \geq I+N} \langle E_{IJ} \Delta_{IJ} \rangle$ and the Gō-type energy by $E_{\mathrm{Go}}^{N} = \Sigma_{J \geq I+N} \langle E_{IJ} \Delta_{IJ} \Delta_{IJ}^{NS} \rangle$.

Proteins and Conformations Used for Analysis

Four proteins were used in the analysis. They are acylphosphatase[44] (AcP, entry 1APS in the PDB), chymotrypsin inhibitor 2[45] (CI2, entry 2CI2 in the PDB), α-lactalbumin[46] (α-LA, entry 1HML in the PDB), and the third fibronectin type III repeat from tenascin[47] (TNfn3, entry 1TEN in the PDB). Much of the analysis refers to AcP, with the other proteins being discussed as appropriate to complement or extend the AcP results. The experimental structure was, in all cases, minimized for 200 steepest descent steps to eliminate bad contacts. The native state ensembles were obtained by simulating the proteins for 2 ns at a constant 300 K temperature (in the canonical ensemble using a Nosé-Hoover thermostat[48]), after heating and equilibrating over 1 ns.

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.[23] and Paci et al., unpublished data). Pathways for the mechanical unfolding (such as the unfolding induced by atomic force microscopy) were determined with the biased molecular dynamics approach described by Paci et al.[35]; samples at a fixed "length" were obtained by performing 200-ps simulations with a constraint on the distance between N- and C-termini. High-temperature unfolded states were obtained by increasing the temperature of the Nosé-Hoover thermostat during the simulation over 1 ns or longer. Collapsed configurations were generated from the high-temperature conformations by gradually decreasing in 200-ps trajectories the temperature to 300 K.

RESULTS

The goal of the present work is to determine if and when Gō-type models provide a meaningful description of native and non-native states that are populated in protein folding and unfolding. Specially, we examine whether for different configurations **R** or averages over configurations, either or both equalities in Eqn 9 are satisfied to a good approximation.

Definition of Residue Pair Contacts: Analysis of the Native Structure

As a first step in the analysis, an appropriate definition of a contact for a residue pair has to be determined to define the cutoff-value r_c. The number of atom pairs as a function of the distance has been examined previously[49] and it was found to have a broad maximum in the range 4.5 to 5.5 Å.

Table I shows the EEF1 energy (Eqn. 6) and the contact energy (Eqn. 7) of the native AcP structure (1APS) for a range of r_c and $N = 2$ to 4. The choice $r_c = 5.5$ Å gives agreement between 0.8 and 2 kcal/mol (out of a total of about 600 kcal/mol) for the two types of energy; the exact difference depends on the choice of N. The largest deviation (2 kcal/mol) occurs for $N = 4$ (i.e., only residues 4 or more apart are included). This suggests that while all the atoms of residues that are 2 or 3 positions apart in the sequence are within $r_c = 5.5$ Å, this is not the case for residues that are at a distance of 4 or greater. The best agreement between the EEF1 and the contact energy is obtained for $r_c = 5.5$ Å and $N = 2$. The energies for larger values of r_c differ by only a small amount, while values smaller than 5 Å introduce significant errors. Thus, the choice $r_c = 5.5$ Å with $N = 2$ is justified both by the nearest neighbor distribution[49] and the agreement between E_{cont}^{2} and E^2. We use the values $r_c = 5.5$ Å and $N = 2$ for most of the subsequent analysis.

Table II shows calculated values of E^N for $N = 0$, 1, and 2. We note that there is a large difference between E^0 obtained from the slightly minimized NMR structure (1APS) and from the native state ensemble (NSE). This difference arises because a minimized structure has a very small contribution from the harmonic bond length and angle terms, while an ensemble of structures at room temperature has a significantly more positive average potential energy (relative to that of the minimum) resulting from equipartition of energy at a finite temperature. These bonded terms are not present in E^2 so that the results for the minimized structure and for the native state ensemble are very similar. Correspondingly, the two values for E^1, which excludes most of the bond length and angle terms, are also very similar. However, E^1 contains large nearest-neighbor terms, which vary only slightly as a function of the structure and, therefore, are of little interest. Consequently, in what follows, we focus on E^2.

TABLE I. E^N_{cont} as a Function of r_c Compared With E^N for Different Values of N†

			E^N_{cont}					E^N				
N	r_c	N_{pairs}	Total	vdW	Elec	[csc]	Solv	Total	vdW	Elec	[csc]	Solv
2	4.0	291	−542.8	−463.9	−472.7	[−577.4]	393.7	−587.9	−578.3	−481.3	[−605.6]	471.7
2	4.5	344	−569.1	−510.6	−477.7	[−583.9]	419.2	−587.9	−578.3	−481.3	[−605.6]	471.7
2	5.0	407	−583.2	−540.2	−482.6	[−593.7]	439.5	−587.9	−578.3	−481.3	[−605.6]	471.7
2	**5.5**	**459**	**−587.1**	**−553.7**	**−482.3**	**[−593.1]**	**449.0**	**−587.9**	**−578.3**	**−481.3**	**[−605.6]**	**471.7**
2	6.0	528	−589.2	−564.8	−482.0	[−600.2]	457.6	−587.9	−578.3	−481.3	[−605.6]	471.7
2	6.5	602	−590.1	−571.2	−481.5	[−603.5]	462.6	−587.9	−578.3	−481.3	[−605.6]	471.7
2	7.0	707	−590.2	−575.4	−481.3	[−605.8]	466.4	−587.9	−578.3	−481.3	[−605.6]	471.7
2	8.0	985	−588.6	−578.2	−481.4	[−605.6]	470.9	−587.9	−578.3	−481.3	[−605.6]	471.7
3	5.5	363	−515.7	−440.3	−422.4	[−544.4]	347.1	−516.6	−464.9	−421.5	[−556.9]	369.8
4	5.5	326	−441.0	−385.4	−356.6	[−447.2]	300.9	−443.0	−407.3	−356.8	[−461.3]	321.0

†One of the NMR structures of AcP (1APS, model 5) is used. N_{pairs} is the number of interacting pairs for a given N and r_c. The reference energy ΔG^{ref} is omitted. The electrostatic contribution with charged side chains [csc] is also reported.

TABLE II. Effective Energies and Partial Contributions Computed for the Native Structure of AcP†

	RMSD	E_{eff}	E^0	E^1	E^2
1APS	0	−2926.3	−1066.9	−1372.8	−587.9
NSE	3.0	−2379.0	−519.6	−1378.0	−621.1

†All quantities in kcal/mol. Quantities for the NSE are computed for the 2 ns trajectory performed in native conditions. $r_c^{atom} = 5.5$ Å is used throughout. The reference free energy for AcP is $\Delta G_{ref} = -1{,}859.4$ kcal/mol; it is included in EEF1; if it is omitted, E^0 is obtained.

Origin of Agreement Between Contact and EEF1 Energy of the Native Structure

The result that excellent agreement between the EEF1 and the contact energy with $r_c = 5.5$ Å is obtained for the native structure (Table I) requires additional analysis since only 459 pairs out of 1,225 within EEF1 cut-off (9 Å) are included in E^2_{cont}. This corresponds to the fact that 99.8% of the energy is obtained from the contact pairs, which are only 37% of the total number of pairs.

Figure 1(a) shows a histogram of the number of pairs that interact with a given EEF1 energy; we use $N = 2$ and those with zero energy ($r_{ij} > 9$ Å) are not considered. The histograms show the results for all the 1,225 residue pairs and the 1,225 − 459 = 766 pairs of residues that are not in contact (i.e., residues with all pairs of heavy atoms more than 5.5 Å apart). Most of the pairs belonging to the latter set interact with an energy less than 0.1 kcal/mol in absolute value. The largest interaction that is disregarded if only pairs in contact are considered is equal to −0.4 kcal/mol while the total energy omitted by including only those in contacts is 0.8 kcal/mol.

The contributions to E^2_{cont} are generally stabilizing, though a few are not. By contrast, as is clear from the insets in Figure 1, the additional non-contact contributions to EEF1 are essentially symmetrically distributed about zero; i.e., some are stabilizing (with a total energy of 10.5 kcal/mol) and some are destabilizing (with a total energy of 9.7 kcal/mol) leading to the net contribution of only 0.8 kcal/mol.

Table I also shows the decomposition of the total energy E^N and E^N_{cont} into the components that contribute to the empirical energy function (see Eqn. 2). If we focus on $r_c^{atom} = 5.5$ Å and $N = 2$ (shown in bold), we see that for the van der Waals interaction, the E^2_{cont} contribution deviates significantly from E^2; i.e., the value is nearly 25 kcal/mol less negative than the latter. However, when the free energy of solvation is added to the van der Waals contribution, the totals are nearly identical (they differ by only 1.9 kcal/mol). Corresponding behavior is observed for other rows of Table I. The electrostatic contribution with neutralized charges (and a distance-dependent electrostatic constant) are essentially equal in EEF1 and E^2_{cont}. If full charges (csc) are used, this is no longer true; i.e., the contact contribution is less negative by 13 kcal/mol, indicating that attractive terms (that would be shielded by the solvent) are present in the latter.

To determine the origin of the results just described, we plot in Figure 2 the pairwise energy contribution to EEF1 as a function of the inter-residue distance (see Fig. 2 legend). The convergence of the total EEF1 energy at about 5.5 Å is evident; small contributions, randomly distributed around zero, occur at larger distances. For the electrostatic contribution, the same behavior is observed; i.e., the values for distances greater than 5.5 Å are small and randomly distributed around zero. However, although the van der Waals energy is about the same magnitude as the electrostatic energy at distances greater than 5.5 Å, all contributions are negative because the attractive component (proportional to r^{-6}) dominates, and the terms that are omitted by taking any cut-off lower than that of the EEF1 potential make a significant contribution, in accord with Table I. The solvation free energy term has a similar monotonic behavior but in this case the contributions are positive and they essentially cancel the negative van der Waals contribution; i.e., as shown in Figure 2, the sum of E_{vdW} and E_{solv} goes to zero at a distance of 5.5 Å (or somewhat less) with the contribution for larger distances being small and random so as to cancel. With neutralized charges and an r-dependent dielectric, the solvation term accounts primarily for the hydrophobic contribution and is, therefore, positive, as seen in Figure 2. Because it is based on a Gaussian exclusion model, it is relatively short range, but the fact that there such a near perfect cancellation between the long-range contribution (≥5.5 Å) between the van der Waals energy and the solvation free energy is an interesting and far from obvious result.

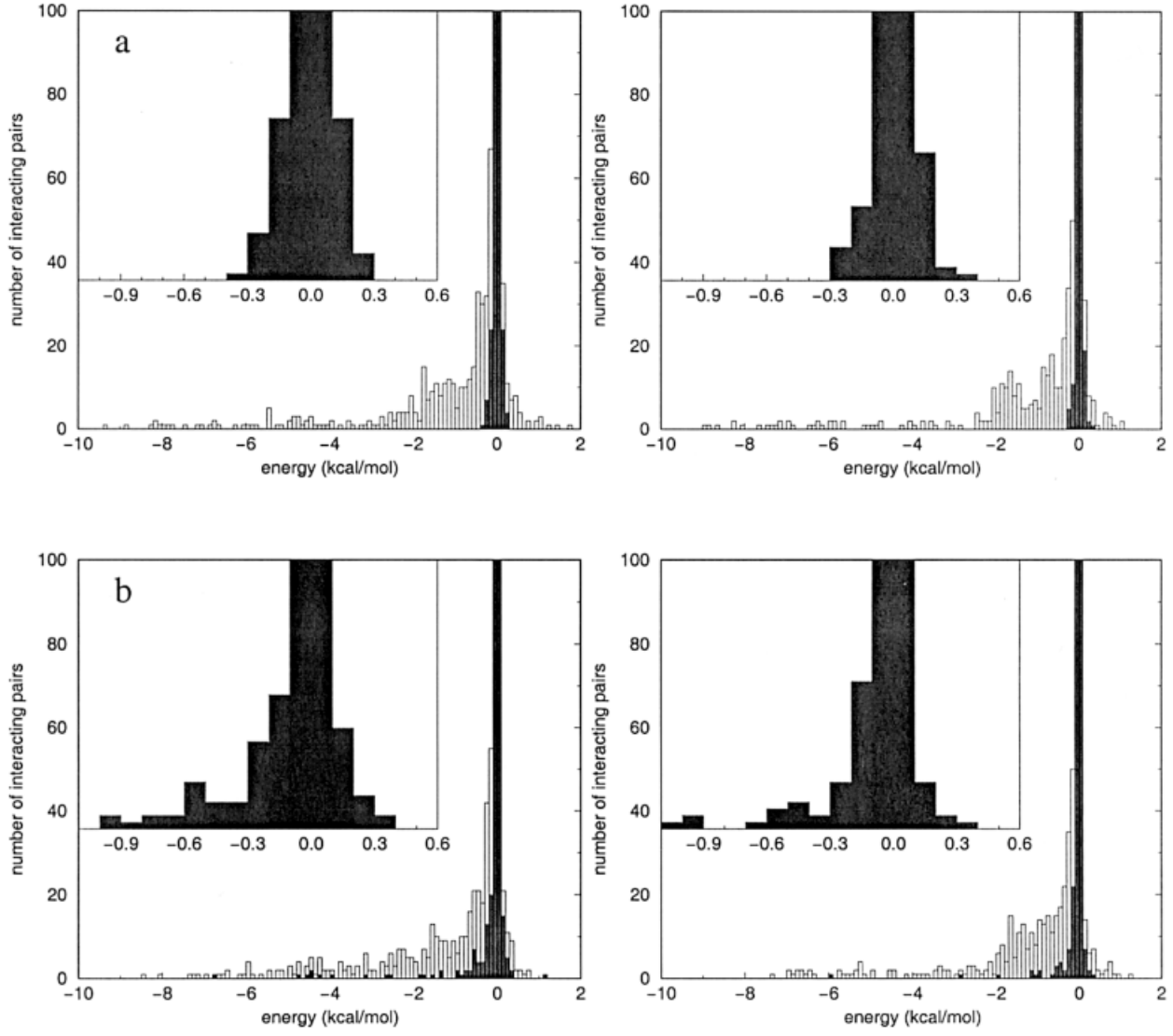

Fig. 1. Distributions of the average pairwise effective energy for the experimental structure (**a**) and the native state ensemble (**b**). **Left:** Result for AcP. **Right:** Result for TNfn3. **Insets:** Histograms of the pairwise energy distribution of the non-contact contribution to the energy.

Summarizing this section, we have shown that a contact model is an excellent approximation to the total effective energy of the native state, as described by a slightly minimized experimental structure. This is a necessary, but not sufficient, condition for the validity of a Gō-type model. It should be noted also, as evident from Figure 2, that if one considers $E_{I,J}$ for individual residue pairs, the cancellation leading to a zero value for contributions from $r > r_c$ is somewhat variable; analysis of the significance of this result for effective residue-residue potentials will be given separately.

Native State Ensembles

Since we use simulations to obtain ensembles representative of non-native states, averages over a set of structures (e.g., the transition state ensemble or TSE) are of primary interest. Consequently, it is necessary first to analyze the native state when it is represented by an ensemble of structures (NSE) generated by molecular dynamics simulations with the EEF1 potential.

As described in Methods, the native state ensemble was generated for several proteins by a 2-ns simulation and a set of 2,000 structures (one each ps) were used for analysis. In Table III, for the NSE of the proteins considered, E^2 and E^2_{cont} show excellent agreement. Between 99.1 and 99.8% of the native effective energy is "contact energy"; these values are very similar to those obtained using the experimental native structure. The calculated energies for the NSE can be slightly more negative (AcP, CI2) or less negative (TNfn3) than for the minimized experimental structure. The non-contact energy, which varies between 0.7 to 5.7 kcal/mol, according to the protein considered, can be attributed to the broadness of the NSE and it increases with the RMSD from the experimental structure (i.e., the starting structure used for the dynamics).

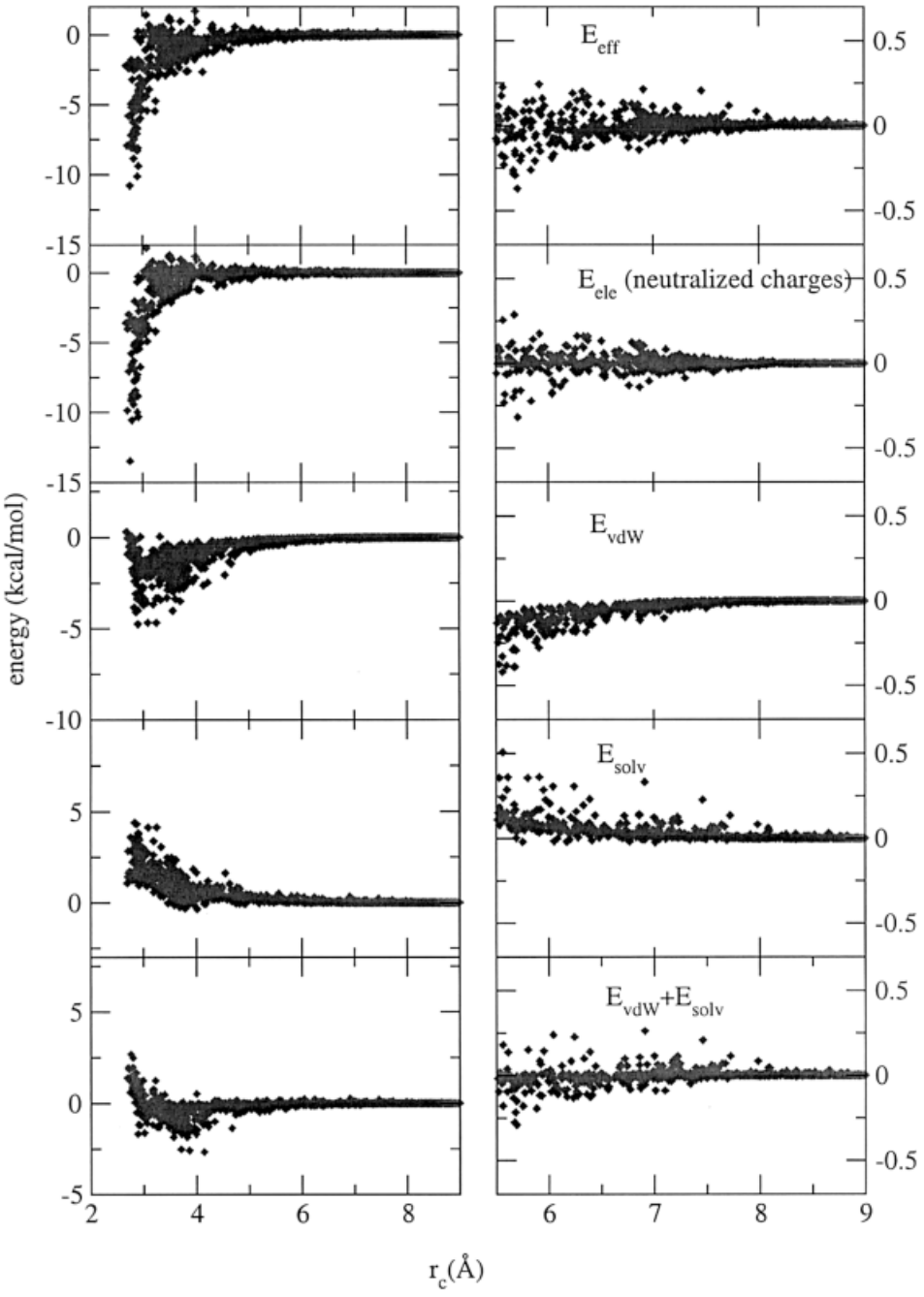

Fig. 2. Pairwise energy contributions between residues as a function of their average distance (i.e., the minimum distance between any two heavy atoms of the two residues) for the experimental structure of AcP. **Left:** Energy contributions between 2 and 9 Å; the latter is the cut-off in EEF1. **Right:** Same (expanded), between 6.5 and 9 Å. The various panels show (top to bottom): the total effective energy (without the constant reference contribution), the electrostatic energy, the van der Waals energy, the solvation energy, and the sum of van der Waals and solvation energies.

To further validate the NSE results, we analyze the pairwise energies E_{IJ} obtained from the NSE of AcP. Histograms of E_{IJ} for pairs in contact and for non-contacting pairs are shown in Figure 1b. The results are very similar to those based on the experimental native structure, shown in Figure 1a. The distribution of the energies of non-contacting pairs is symmetric and centered on $E_{IJ} = 0$. Only very few pairs beyond the 5.5 Å interact with an energy larger than 0.2 kcal/mol in absolute value. The distribution from the NSE is somewhat broader than for the experimental structure and there are a few relatively large non-contact stabilizing interactions (the largest is -0.4 kcal/mol). Thus, as for the experimental structure, the non-contact pairwise energies are small and their effect corresponds to a weakly perturbing random "background," relative to the contact interac-

TABLE III. Various Energies Computed for "States" for Several Proteins (See Text)[†]

	RMSD (C_α)	E^2	E^2_{cont}	E^2_{Go}	$E^2_{non\text{-}Go}$
States of AcP					
NSE	3.0	−621.3	−622.5	−614.3	−8.2
TSE	5.2	−590.1	−592.2	−486.1	−106.1
States of TNFn3					
NSE	2.1	−476.0	−477.4	−472.2	−5.2
450 K	6.7	−336.8	−340.8	−258.4	−82.5
TSE	5.1	−481.0	−484.3	−432.5	−51.8
r_{NC} = 36.4 Å	13	−447.4	−448.8	−407.7	−41.1
r_{NC} = 45.1 Å	14	−450.6	−454.2	−410.8	−43.4
r_{NC} = 139 Å	37	−311.5	−315.2	−292.8	−22.5
r_{NC} = 326 Å	93	−19.0	−19.5	−19.5	0.0
States of CI2					
NSE	1.6	−353.6	−355.8	−352.2	−3.6
TSE	5.8	−299.6	−301.2	−247.1	−54.1
States of α-LA					
NSE	1.8	−703.5	−701.3	−682.9	−18.4
Group 1	4.8	−691.2	−691.4	−607.6	−83.8
Group 2	8.3	−667.8	−668.4	−539.9	−128.5
Group 3	14	−629.9	−632.9	−443.0	−189.9

[†]All quantities in kcal/mol. Values for the NSE are computed from the 2-ns trajectory performed under native conditions. For E^2_{cont}, a cutoff of 5.5 Å between heavy atoms is used and pairs $I, J \geq I + 2$ are considered. The distance r_{NC} is the distance imposed between the N- and C-termini of the protein along an unfolding pathway along which the protein is mechanically stretched (see text). The definition of the groups for α-LA is discussed in the text; see also Paci et al.[37] The RMSD value is that from the minimized experimental structure.

tions. The corresponding results obtained for TNfn3 are shown in Figure 3(c) and (d) and confirm the analysis for AcP.

The results in the present section justify the use of ensembles generated by molecular dynamics simulations for assessing the validity of Gō-like models for non-native states.

Non-Gō Interactions in Non-Native States

In this section, we determine first whether contact interactions provide a good approximation to the total energy of non-native state ensembles (i.e., whether $E^2_{eff} \simeq E^2_{cont}$), and given that they do, whether the magnitude of the contribution of interactions of the non-Gō type, $E^2_{non\text{-}Go}$, to the contact energy are small or large (i.e., whether $E^2_{cont} \simeq E^2_{Go}$ or not). Table III shows the results obtained for the quantities E^2, E^2_{cont}, E^2_{Go}, and $E^2_{non\text{-}Go}$ (as defined by Eqn 7 and 8 for different states of the proteins generated as described in Methods). For these states, the quantities E_{IJ} and Δ_{IJ} in the definitions correspond to averages over ensembles (see Methods).

Transition State Ensembles

We consider the transition state ensembles generated for AcP, TNfn3, and CI2 (see Methods). These are all two-state proteins for which the TSE play an essential role in the folding. The results for individual structures from the ensembles (not shown) and for averages over the ensembles are very similar. For AcP, about 20% of the total (or contact) energy is contributed by non-Gō terms ($E^2_{non\text{-}Go}$), while for TNfn3 and CI2 the contribution are about 12 and 20%, respectively. Thus, the Gō-type model, which is based on contacts present in the native state ensemble, gives the major part of the energy in the TSE for these proteins, but the non-Gō contribution is not negligible. It should be noted also that the overall non-Gō contribution is stabilizing, rather than destabilizing as has been assumed in some models; e.g., for certain values of the bias parameter studied by Zhou and Karplus[9] for a C_α model of a three-helix-bundle protein and in the Monte Carlo simulation of crambin by Shimada et al.[27]

Figure 3 compares E^2_{Go} and $E^2_{non\text{-}Go}$ for the ensembles of structures corresponding to the native and transition state of AcP, TNfn3, and CI2. It can be seen that in the NSE, the E^2_{Go} contributions are strongly stabilizing, while the $E^2_{non\text{-}Go}$ terms are small and distributed nearly symmetrically about zero (see also Fig. 1). For the transition state ensemble, E^2_{Go} has a behavior similar to that in the NSE, but the result for $E^2_{non\text{-}Go}$ is different; for the latter there are small but significant contributions that are stabilizing.

Thermal Unfolding

For an evaluation of the Gō-type model along the entire reaction path, we show in Figure 4(a) the results obtained as a function of time for a 2-ns high temperature (450 K) trajectory of CI2, started from the native state. The protein unfolds in less than 1 ns. While close to the native state (RMSD ≤ 2.5 Å) the non-Gō energy is very small, for more unfolded conformations the Gō and non-Gō contributions to the contact energy are of comparable magnitude, e.g., at 1.5 ns, where the RMSD from the native structure is in the range 10 to 15 Å, E^2_{Go} and $E^2_{non\text{-}Go}$ are both equal to about −100 kcal/mol. As noted earlier for other structures, E^2_{cont} is essentially identical to the total energy in all cases. This means that the total energy is determined by the interactions of residues that are in contact, whether or not the contacts are native or non-native. Also, as found for the transition state, both E^2_{Go} and $E^2_{non\text{-}Go}$ contributions to the effective energy are stabilizing. This behavior is demonstrated very clearly in the unfolding trajectory of CI2 (see Fig. 5). For t = 1 ps and up to 501 ps, the non-Gō contributions are small and relatively symmetrically distributed about zero. However, as the structure becomes more non-native (t > 751 ps), significantly stabilizing non-Gō energies become important. At very long times (t = 1751 ps), the Gō and non-Gō terms become more symmetric again about zero and the contribution of the latter to the stabilization of the protein becomes smaller.

Twenty conformations were chosen from the unfolding trajectory at equally spaced (100 ps) intervals and quenched smoothly by decreasing the temperature to 300 K during 200-ps simulations. Figure 4(b) shows the components of the energy for the last configuration of each simulation; the time corresponds to that along the 450-K simulation from which the simulations were initiated. From the R_g values, it is clear that the quenched structures are significantly more compact than the high-temperature structures [compare R_g in Fig. 4(b)], but the RMSD from the native structure does not change significantly; i.e., the

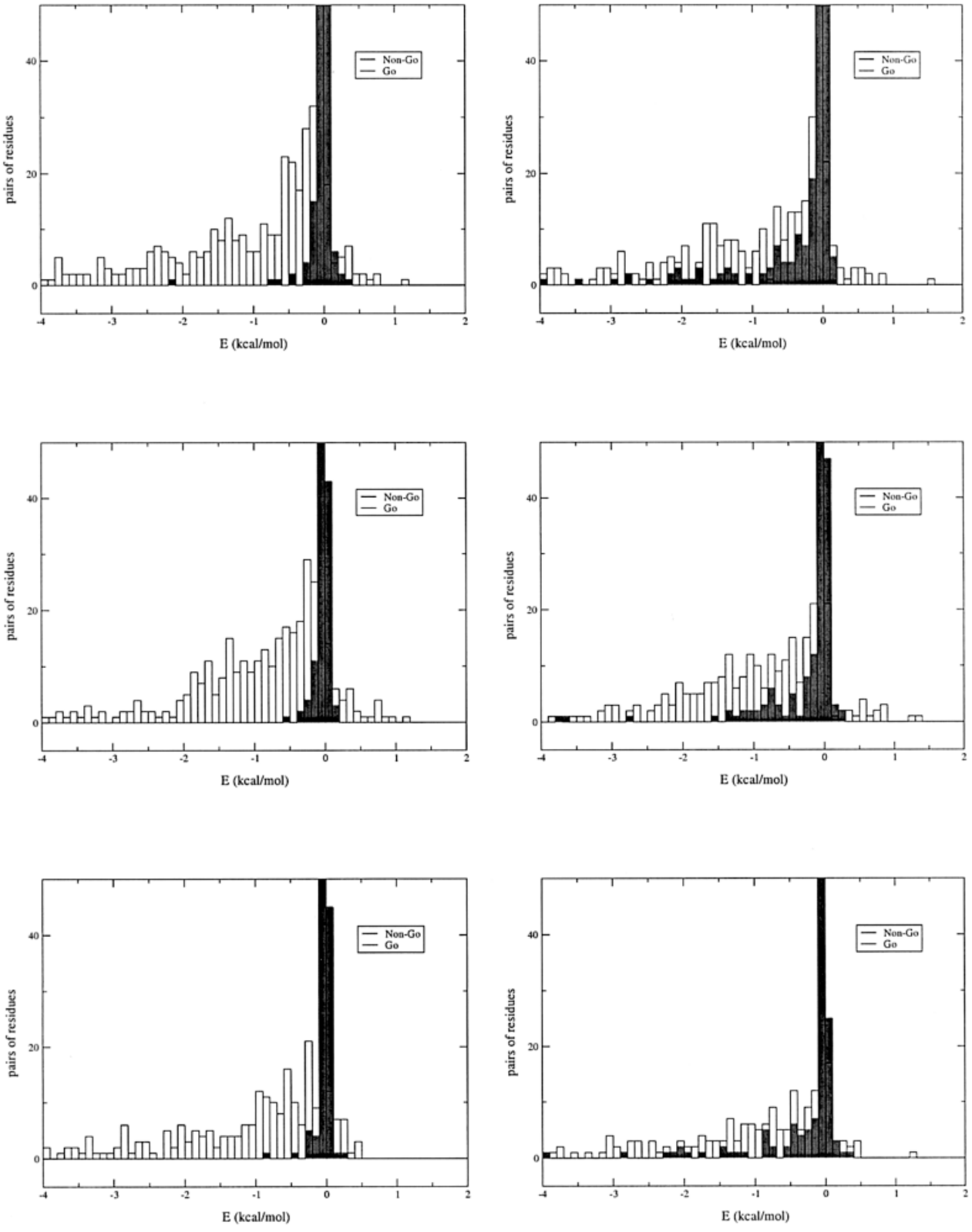

Fig. 3. Distribution of Gō and non-Gō energies for ensemble of structures representing the native (**left**) and the transition state (**right**). From the top: AcP, TNfn3, CI2. The Gō-terms are shown in open bars and the non-Gō terms in solid bars.

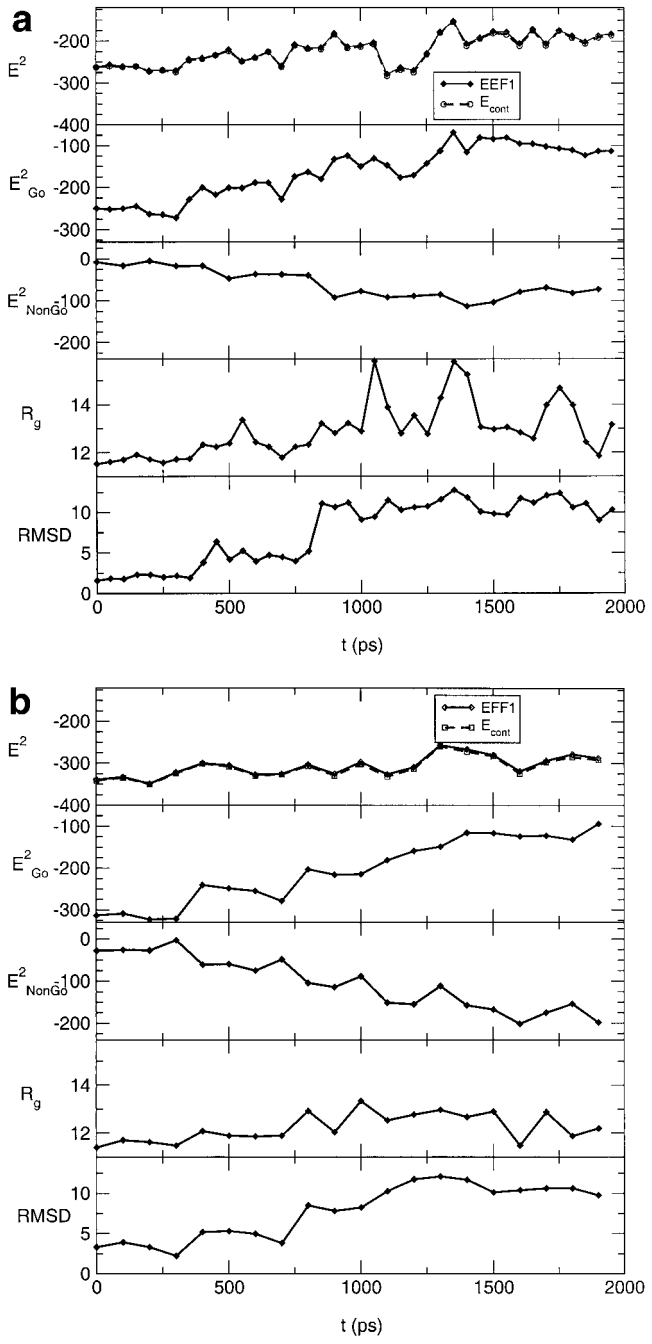

Fig. 4. The energy components shown in Table III (EEF1, contact, Gō, and non-Gō) are plotted as a function of time (**a**) for an unfolding trajectory of CI2 at T = 450 K started from the equilibrated native structure and (**b**) for more compact temperature-quenched structures along that trajectory (see text). The C_α-RMSD and R_g along the trajectory are also shown.

structures are not more native-like and correspond to a collapsed disordered globule. Comparison of the same time points in Figure 4(a) and (b) shows that E^2 is considerably more negative in the collapsed structures. Since E^2_{Go} is very similar for the two sets of structures, the decrease in effective (and contact) energy is due essentially to $E^2_{\text{non-Go}}$. For the least native conformations, $E^2_{\text{non-Go}}$ is the dominant contribution to E^2.

The results for CI2 are confirmed by those for TNfn3 (not shown). A 1.4-ns simulation was performed at 450 K, which samples relatively compact non-native states characterized by a RMSD between 3 and 10 Å, and R_g between 13 and 15 Å (compared to the native value 13.0 ± 0.1 Å). The non-Gō fraction of the contact energy reaches 30%; e.g., for the last conformation of the trajectory E^2 equals −305 kcal/mol, E_{Go} equals −217 kcal/mol, and $E_{\text{non-Go}}$ equals −91 kcal/mol.

Unfolding Trajectories Produced by External Forces

For TNfn3, we generated an unfolding pathway by pulling the two ends of the chain apart, in analogy with AFM experiments.[35,36] The five conformations selected for analysis correspond to metastable states under a constant force; they are "bottlenecks" in the unfolding pathways along the pulling direction. Starting from these conformations, 200-ps simulations were performed with a holonomic constraint to keep constant the distance between the main chain nitrogen of the N-terminal residue to the carbon of the C-terminal residue. Table III shows the results obtained for these non-native "states." The total energy E^2 decreases in magnitude over the trajectory and E^2_{Go} is a good approximation. Unlike the thermal unfolding trajectory, the contribution of $E^2_{\text{non-Go}}$ remains small. It never exceeds 40 kcal/mol, on the order of 10% of the total energy.

Unfolding of α-LA

A corresponding analysis of the various energy contributions for a series of structures of α-LA is also reported in Table III. These structures[37] are representative of the native state (NSE), the molten globule (group 1), and structures populated during equilibrium unfolding in urea (groups 2 and 3). In each case, average quantities are calculated over the most representative structures (between 60 and 140) for each group.

Also in this case, the contact energy, E^2_{cont}, is an excellent approximation to the total EEF1 energy, E^2. As the structures become less native-like, the contribution from $E^2_{\text{non-Go}}$ increases in importance. Although E^2_{Go} remains larger in magnitude than $E^2_{\text{non-Go}}$, for strongly denatured (10 M urea) but still relatively compact structures ($R_g \simeq$ 20 Å vs. 14.5 Å of the native state) $E^2_{\text{non-Go}}$ is about 30% of the total effective energy.

CONCLUDING DISCUSSION

Gō-type models have come into common use for studies of protein folding, in part because it has not been possible to fold most proteins with more realistic potentials of the molecular mechanics type. As realized many years ago by Gō and his collaborators,[18,19] use of a potential function that has only attractive native contacts and neglects all other contacts (or makes them repulsive) leads to rapid folding to the (known) native structure. Gō and coworkers[18] used lattice models, but off-lattice C_α and all-atom Gō-type models (usually augmented by other potential energy terms, describing excluded volume or dihedral angle preferences) have been shown to be useful for obtaining insights into the protein folding reaction by means of both Monte Carlo and molecular dynamics

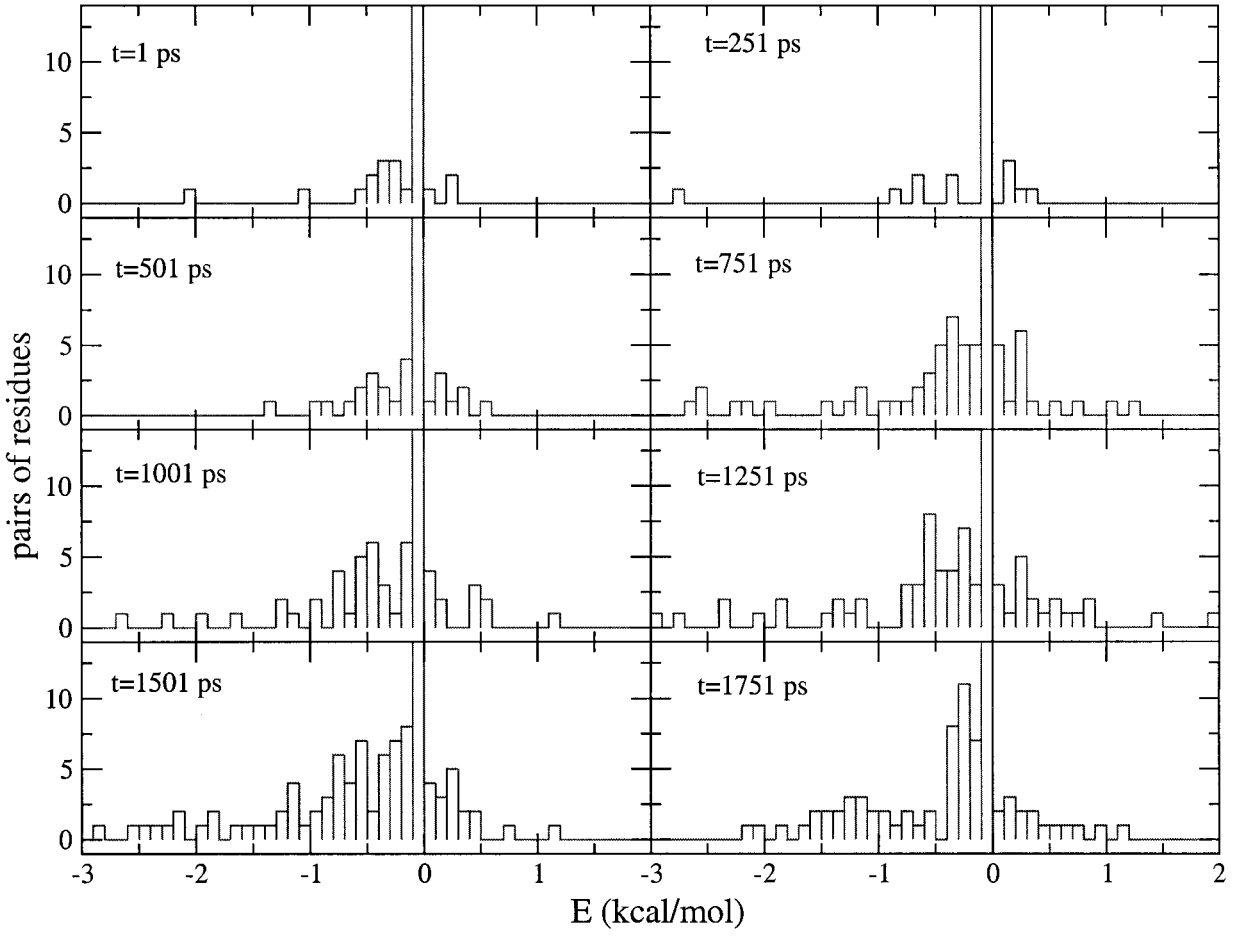

Fig. 5. Distribution of non-Gō energies for some of the snapshots along the high-temperature unfolding pathway of CI2 shown in Figure 4(a).

simulations.[9,11,22,27,50,51] Also, more simplified one-dimensional models with Gō-type interactions have been described that provide estimates of relative folding rates of a set of small, experimentally studied proteins that fold in a two-state manner.[24,25,52] The low-resolution structures of the transition state of a series of proteins have been determined by use of experimental ϕ values with a C_α Gō-type model,[23] which complements the all-atom results used here.

It is important, therefore, to determine whether or not a Gō-type model is an adequate description of the interactions between residues of the polypeptide chain for configurations ranging from the native to the denatured state. The obvious approach is to use molecular mechanics potentials for examining this question. Although they are only approximate, there is considerable evidence that they provide a good description of the potential energy surface when the effect of the solvent is included in the calculations. Gō-type potentials are essentially potentials of mean force (PMF); i.e., they represent the effective energy for each configuration of the polypeptide chain in the presence of the canonically averaged solvent. To obtain such an effective potential or PMF by simulations with explicit solvent is an extremely time-consuming task. The closest approach to such a calculation are the free energy surfaces along folding pathways determined by Brooks and collaborators.[53–55] However, even if one has such surfaces, their decomposition to test Gō-type models is still complex because it requires a separation of the contributions of Gō and non-Gō type terms to the effective energy. Consequently, we have approached the problem by use of a pairwise decomposable all-atom molecular mechanics energy function with an implicit solvent model. The effective (PMF) energy function, referred to as EEF1,[31] has been used in this paper to study a range of native and non-native structures; the latter include transition states, unfolding pathways at high temperature and in the presence of external forces, as well as collapsed unfolded and molten globule states. It was found in all cases that the contact contribution is an excellent approximation to the total effective energy, even though only about one third of the pairwise interactions included in EEF1 are taken into account; a cut-off of 5.5 Å based on the maximum in the radial distribution function was used. A detailed analysis of the various energy contributions demonstrated that the contact energy is so accurate because of a nearly exact cancellation between the van der Waals interactions and the implicit solvation terms beyond 5.5 Å. Also, the long-range electrostatic interactions average to zero, in part because the side-chains of charged residues are neutralized to account for solvent shielding. For the native state, represented either by an experimental structure or an ensemble of structures gener-

ated by molecular dynamics with the same energy function, a Gō-type model (i.e., inclusion only of native contacts or average native contacts from ensembles) is consequently an excellent approximation. For non-native states, while the contact approximation is still valid (i.e., the contributions to the effective energy from interactions arising from distances larger than 5.5 Å is essentially negligible), a considerable fraction of the effective energy can arise from non-native interactions; e.g., for the thermal unfolding of CI2, the non-Gō energy can be equal to the Gō-energy. Moreover, the non-Gō term always make an overall stabilizing contribution to the energy. When the transition state is rather native-like, which is the case for some proteins, particularly for fast-folding proteins that have recently been studied experimentally and theoretically,[56] the Gō approximation is still approximately valid. For the three examples considered in this paper (CI2, TNfn3, and AcP), the non-Gō contribution to the transition state ensemble is between 12 and 20% of the total effective energy.

In summary, the results obtained here clearly demonstrate that the contact term (for $r_c = 5.5$ Å) is an excellent approximation to the total effective energy calculated with the EEF1 model for a wide range of native and non-native structures. The effect of the small error due to non-native contacts in protein fold predictions remains to be investigated; a more detailed analysis of contact models[39] will be given separately. The analysis of non-native states shows that stabilizing contributions to the effective energy from non-native contacts can be as large as those from native contacts. This result suggests that caution is required in interpreting the results from models of the Gō-type that include only native interactions. Particularly for systems where there is early collapse to a disordered compact globule and the search through the compact globule state is the slow step, non-native interactions have to be considered. By contrast, for proteins that collapse directly to a native-like transition state, the non-Gō terms are less important. More generally, it should be noted that errors due to the neglect of non-native terms, even if small, may be significant because the difference in the effective energy between the transition state and the native state is often not large.

Other studies, which have commented on the role of non-native contacts, include the lattice model analysis of Li et al.[57] and the all-atom unfolding simulations of Kazmirski and Daggett.[58] The former found in lattice model simulations that non-native interactions in the transition state (nucleus) can be important in the kinetics but, not surprisingly, have no effect on the stability of the native state.

The present results strongly suggest that the stabilizing non-native (non-Gō) terms are part of a true description of the effective energy surface of a polypeptide chain sampled during the folding reaction, as well as of other non-native states, such as molten globules. Such terms, of course, make non-native states more stable than they are in Gō models and thus are one of the factors that result in a "rough" effective energy surface, which makes difficult molecular dynamics or Monte Carlo simulations of protein folding with more realistic potentials. The "roughness" appears to be an important element, even for relatively fast-folding proteins, because they require milliseconds or longer to reach the native state in most cases. If the potential energy function corresponded to a steep ideal "funnel," folding could be significantly faster, faster even than the "free" diffusion limit of microseconds.[59,60]

A referee asked us to comment on "Why do Gō potentials work so well?" implying that the deviations we find should introduce significant errors. The simple answer to his/her question is that we really do not know how well they work because the available experimental data are too limited for a definitive test. Two types of uses have been made of Gō potentials. The first is to obtain generic models of protein folding and the second is to do calculations for specific proteins. There are many generic model calculations with Gō potentials and they have given useful insights about possible folding mechanisms. However, no actual tests of the results exist. In more specific models (e.g., Zhou and Karplus[9] and Shimada et al.[27]), the results do appear to correspond to experimental data in some cases. In one such model applied to three-helix bundle proteins,[9] the contribution of non-native contact interactions was explored and the most realistic behavior was obtained when they are significantly weaker than the native contacts. In another study (E. Shakhnovich, personal communication), where a number of parameters were adjusted to obtain the correct stability of secondary structural elements, certain aspects of the experimental folding behavior were reproduced. Thus, it appears that, with some type of empirical adjustment, detailed aspects of folding can be mimicked by Gō models. An example is the use of a Gō-type model to determine the transition state with experimental data based on ϕ values.[23] Further exploration of Gō-type models, particularly when augmented by experimental data, thus seems to be appropriate, in spite of the deviations found in the present analysis.

We believe that the conclusions described are of general significance, even though they are obtained by studying a small number of proteins and are based on an approximate molecular mechanics potential with a simple solvation correction.

ACKNOWLEDGEMENTS

EP was supported by a Marie Curie Fellowship of the European Union. MV was an EMBO long term fellow.

REFERENCES

1. Dinner AR, Sali A, Smith LJ, Dobson CM, Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment. Trends Biochem Sci 2000;25:331–339.
2. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
3. Daura X, Jaun B, Seeback D, van Gunsteren WF, Mark AE. Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 1998;280:925–932.
4. Schaefer M, Bartels C, Karplus M. The solution conformation and thermodynamics of structured peptides: Molecular dynamics simulation with and implicit solvation model. J Mol Biol 1998;284:835–847.
5. Wang H, Varady J, Ng L, Sung SS. Molecular dynamics simulations of β-hairpin folding. Proteins 1999;37:325–333.
6. Ferrara P, Caflisch A. Folding simulations of a three-stranded antiparallel β-sheet peptide. Proc Natl Acad Sci USA 2000;97:10780–10785.
7. Shirts M, Pande V. COMPUTING: screen savers of the world unite! Science 2000;290:1903–1904.

8. Karplus M, Šali A. Theoretical studies of protein folding and unfolding. Curr Opin Struct Biol 1995;5:58–73.
9. Zhou Y, Karplus M. Interpreting the folding kinetics of helical proteins. Nature 1999;401:400–403.
10. Guo Z, Thirumalai D. The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. Fold Des 1997;2:377–391.
11. Shea JE, Onuchic JN, Brooks CL. Exploring the origins of topological frustration: design of a minimally frustrated model of fragment B of protein A. Proc Natl Acad Sci USA 1999;96:12512–12517.
12. Dobson CM, Šali A, Karplus M. Protein folding: a perspective from theory and experiment. Angew Chem Int Ed 1998;37:868–893.
13. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy-model (with applications to protein folding). J Phys Chem 1989;93:6902–6915.
14. Durup J. On "Levinthal paradox" and the theory of protein folding. Theochem J Mol Struct 1998;424:157–169.
15. Bicout DJ, Szabo A. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. Prot Sci 2000;9:452–465.
16. Finkelstein AV, Badretdinov AY. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Fold Des 1997;2:115–121.
17. Karplus M. The Levinthal paradox: yesterday and today. Fold Des 1997;2:S69–S75.
18. Taketomi H, Ueda Y, Gō N. 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 Prot Res 1975;7:445–459.
19. Gō N. Theoretical studies of protein folding. Ann Rev Biophys Bioeng 1983;12:183–210.
20. Takada S. Gō-ing for the prediction of protein folding mechanisms. Proc Natl Acad Sci USA 1999;96:11698–11700.
21. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Discrete molecular dynamics studies of the folding of a protein-like model. Fold Des 1998;3:577–587.
22. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details transition state ensemble and "En-route" intermediates for protein investigation for small globular proteins. J Mol Biol 2000;298:937–953.
23. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature 2001;409:641–645.
24. Muñoz V, Eaton WA. A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc Natl Acad Sci USA 1999;96:11311–11316.
25. Alm E, Baker D. Matching theory and experiment in protein folding. Curr Opin Struct Biol 1999;9:189–196.
26. Karplus M, Shakhnovich E. Protein folding: Theoretical studies of thermodynamics and dynamics. In: Creighton TE, editor. Protein folding. New York: W. H. Freeman; 1992. p 127–195.
27. Shimada J, Kussell EL, Shakhnovich EI. The folding thermodynamics and kinetics of crambin using an all-atom Monte Carlo simulation. J Mol Biol 2001;308:79–95.
28. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rate of single domain proteins. J Mol Biol 1998;277:985–994.
29. Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology, stability, sequence, and length: defining the determinants of folding kinetics. Biochemistry 2000;39:11177–11183.
30. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular dynamics system. J Chem Phys 1996;105: 1902–1921.
31. Lazaridis T, Karplus M. Effective energy function for protein dynamics and thermodynamics. Proteins 1999;35:133–152.
32. Lazaridis T, Karplus M. "New view" of protein folding reconciled with the old through multiple unfolding simulations. Science 1997;278:1928–1931.
33. Li A, Daggett V. Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. J Mol Biol 1996;257:412–429.
34. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 1999;288:477–487.
35. Paci E, Karplus M. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. J Mol Biol 1999;288:441–459.
36. Paci E, Karplus M. Unfolding proteins by external forces and high temperatures: the importance of topology and energetics. Proc Natl Acad Sci USA 2000;97:6521–6526.
37. Paci E, Smith LJ, Dobson CM, Karplus M. Exploration of partially unfolded states of human α-lactalbumin by molecular dynamics simulation. J Mol Biol 2001;306:329–347.
38. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal-structures. Quasi-chemical approximation. Macromolecules 1985;18:534–552.
39. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
40. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci USA 2000;97:2550–2555.
41. Lazaridis T, Archontis G, Karplus M. Enthalpic contribution to protein stability: insights from atom-based calculation statistical mechanics. Adv Prot Chem 1995;47:231–306.
42. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 1983;4: 187–217.
43. Steinbach PJ, Brooks BR. New spherical-cutoff methods for long-range forces in macromolecular simulation. J Comp Chem 1994;15:667–683.
44. Pastore A, Saudek V, Ramponi G, Williams RJ. Three-dimensional structure of acylphosphatase. Refinement and structure analysis. J Mol Biol 1992;224:427–440.
45. McPhalen CA, James MN. Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds. Biochemistry 1987;26:261–269.
46. Ren J, Acharya KR, Stuart DI. α-lactalbumin possesses a distinct zinc binding site. J Biol Chem 1993;268:19292–19298.
47. Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science 1992;258:987–991.
48. Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A 1985;31:1695–1697.
49. Zhou Y, Vitkup D, Karplus M. Native proteins are surface-molten solids: application of the Lindemann criterion for the solid versus liquid state. J Mol Biol 1999;285:1371–1375.
50. Shoemaker BA, Wang J, Wolynes PG. Structural correlations in protein folding funnels. Proc Natl Acad Sci USA 1997;94:777–782.
51. Portman JJ, Takada S, Wolynes PG. Variational theory for site resolved protein folding free energy surfaces. Phys Rev Lett 1998;81:5237–5240.
52. Galzitskaya OV, Finkelstein AV. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. Proc Natl Acad Sci USA 1999;96:11299–11304.
53. Boczko EM, Brooks CL. First-principles calculation of the folding free energy of a three-helix bundle protein. Science 1995;269:393–396.
54. Guo Z, Brooks CL, Boczko EM. Exploring the folding free energy surface of a three-helix bundle protein. Proc Natl Acad Sci USA 1997;94:10161–10166.
55. Sheinerman FB, Brooks CL. Molecular picture of folding of a small α/β protein. Proc Natl Acad Sci USA 1998;95:1562–1567.
56. Jackson SE. How do small single-domain proteins fold? Fold Des 1998;3:R81–R91.
57. Li L, Mirny LA, Shakhnovich EI. Kinetics, thermodynamics and evolution of non-native interactions in a protein folding nucleus. Nature Struct Biol 2000;7:336–342.
58. Kazmirski SL, Daggett V. Simulation of the structural and dynamical properties of denatured proteins: the "molten coil" state of bovine pancreatic tripsin inhibitor. J Mol Biol 1998;277:487–506.
59. Hagen SJ, Hofrichter J, Szabo A, Eaton WA. Diffusion-limited contact formation in unfolded cytochrome *c:* Estimating the maximum rate of protein folding. Proc Natl Acad Sci USA 1996;93: 11615–11617.
60. Gruebele M. The fast protein folding problem. Annu Rev Phys Chem 1999;50:485–516.
61. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.

APPENDIX

C_α-Based Contact Definition and Gō-Models

For C_α-based contacts, the value of 8.5 Å has often been used in analyses based on Gō-type models.[61] To determine an appropriate cut-off, the total number of residues in contact using the heavy atom and C_α definitions is plotted

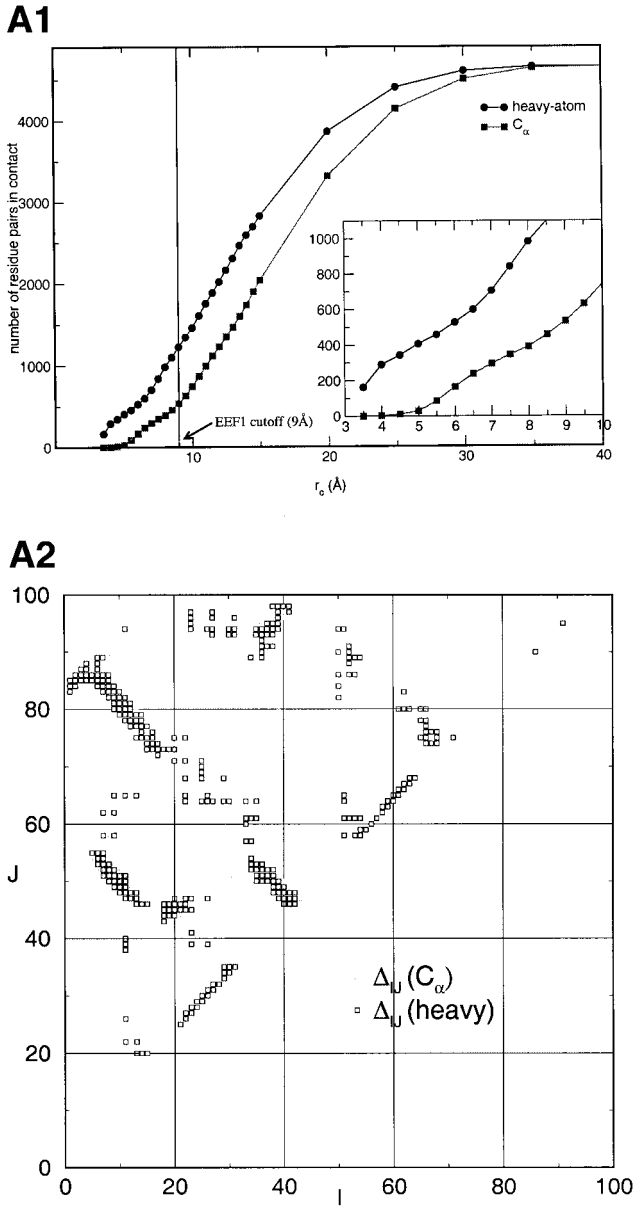

Fig. A-1. Comparison of heavy-atom and C_α-atom approximations for defining "contacts." The total number of pairs within the EEF1 cut-off is 1,225, so that about 37% are included with a 5.5 Å heavy atom cut-off.

Fig. A-2. Contact matrix: $\Delta_{IJ} = 1$ are represented as squares. Contacts between heavy atoms are shown as black squares; those between C_α atoms as light squares. Pairs less than 4 residues apart in the sequence are not shown.

in Figure A-1 as a function of the cut-off distance, r_c, for the experimental structure of AcP. Essentially the same number of contacts is found when the cut-off based on the C_α definition is increased by 3 Å relative to the heavy-atom definition. A consistent choice is $r_c^{\text{atom}} = 5.5$ Å for heavy-atoms and $r_c^{C_\alpha} = 8.5$ Å for C_α-atoms. With such a choice and $N = 2$ (i.e., contacts in the same residue and nearest neighbor residues are not included), there are 459 and 458 pairs in contact, respectively. The contact maps obtained with the two cut-offs are shown in Figure A-2. They are very similar although there are some differences. The C_α-atom approximation favors the existence of secondary structural contacts (where a gray square is not overlapped by a black box in Fig. A-2) while several native tertiary contacts are lost (indicated by an empty black box in Fig. A-2). The latter is most prominent for large hydrophobic residues (e.g., Tyr11, Phe22, Tyr25, Trp64, Phe80) but also for some residues that are completely buried (such as Gln50, Val51).

For $r_c^{C_\alpha} = 8.5$ Å, which corresponds to the same numbers of residue pairs as included in $r_c^{\text{atom}} = 5.5$ Å, the agreement between E^2_{cont} and E^2 is not as good (Table A-1). Independent of N for $N \leq 4$, the value of E^N_{cont} is significantly less than the corresponding value of E^N (by about 33 kcal/mol). Even for $r_c^{C_\alpha} = 11$ Å, the two values still differ by 4 kcal/mol; E^N_{cont} converges to E^N only when essentially all pairs within the EEF1 threshold are included. Therefore, C_α-based definitions of contacts are not satisfactory, unless a large r_c is used. Beyond $r_c^{C_\alpha} = 9$ Å, however, non-neighboring residue pairs are included. One reason for the poorer performance of the C_α relative to the heavy-atom definition of contact is that the EEF1 energy function is based on an all-atom description of the protein. A force field directly based on a C_α definition of the interactions is likely to be better approximated by E^N_{cont}.

TABLE A-1. Contact (E^N_{cont}) and Total (E^N) Energy Compared for One of the NMR Structures of AcP (1APS) as a Function of $r_c^{C_\alpha}$ and N†

			$E^{\text{N}}_{\text{cont}}$				E^N			
N	r_c	N_{total}	Total	vdW	Elec [csc]	Solv	Total	vdW	Elec [csc]	Solv
2	6.0	165	−312.0	−265.0	−298.5	251.4	−587.9	−578.3	−481.3 [−605.6]	471.7
2	7.0	295	−493.6	−418.9	−445.4	370.7	−587.9	−578.3	−481.3 [−605.6]	471.7
2	8.0	391	−545.1	−488.4	−469.7	412.9	−587.9	−578.3	−481.3 [−605.6]	471.7
2	**8.5**	**458**	**−554.5**	**−517.7**	**−466.9**	**430.1**	**−587.9**	**−578.3**	**−481.3 [−605.6]**	**471.7**
2	9.0	534	−567.1	−535.6	−471.7	440.2	−587.9	−578.3	−481.3 [−605.6]	471.7
2	10.0	742	−582.7	−556.2	−481.5	454.9	−587.9	−578.3	−481.3 [−605.6]	471.7
2	11.0	995	−584.2	−568.6	−481.6	466.1	−587.9	−578.3	−481.3 [−605.6]	471.7
3	8.5	362	−483.1	−404.3	−407.0	328.2	−516.6	−464.9	−421.5 [−556.9]	369.8
4	8.5	320	−408.6	−348.7	−341.5	281.6	−443.0	−407.3	−356.8 [−461.3]	321.0

†The reference solvation term ΔG^{ref} is omitted. The number of interacting pairs is reported.