Protein Science (1997), 6:676–688. Cambridge University Press. Printed in the USA.

Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?

JEFFREY SKOLNICK,[1] LUKASZ JAROSZEWSKI,[2] ANDRZEJ KOLINSKI,[1,2] AND ADAM GODZIK[1]

[1]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037
[2]Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

(RECEIVED September 18, 1996; ACCEPTED December 5, 1996)

Abstract

Many existing derivations of knowledge-based statistical pair potentials invoke the quasichemical approximation to estimate the expected side-chain contact frequency if there were no amino acid pair-specific interactions. At first glance, the quasichemical approximation that treats the residues in a protein as being disconnected and expresses the side-chain contact probability as being proportional to the product of the mole fractions of the pair of residues would appear to be rather severe. To investigate the validity of this approximation, we introduce two new reference states in which no specific pair interactions between amino acids are allowed, but in which the connectivity of the protein chain is retained. The first estimates the expected number of side-chain contacts by treating the protein as a Gaussian random coil polymer. The second, more realistic reference state includes the effects of chain connectivity, secondary structure, and chain compactness by estimating the expected side-chain contact probability by placing the sequence of interest in each member of a library of structures of comparable compactness to the native conformation. The side-chain contact maps are not allowed to readjust to the sequence of interest, i.e., the side chains cannot repack. This situation would hold rigorously if all amino acids were the same size. Both reference states effectively permit the factorization of the side-chain contact probability into sequence-dependent and structure-dependent terms. Then, because the sequence distribution of amino acids in proteins is random, the quasichemical approximation to each of these reference states is shown to be excellent. Thus, the range of validity of the quasichemical approximation is determined by the magnitude of the side-chain repacking term, which is, at present, unknown. Finally, the performance of these two sets of pair interaction potentials as well as side-chain contact fraction-based interaction scales is assessed by inverse folding tests both without and with allowing for gaps.

Keywords: empirical parameter sets; inverse protein folding; protein structural database; protein threading; quasichemical approximation

Recently, it has become increasingly recognized that the key to the solution of the protein folding problem lies in the development of potentials that can distinguish the native conformation from the myriad possible alternative structures (Jernigan & Bahar, 1996). Any successful folding algorithm must also be able to find this global energy minimum conformation (Hao & Scheraga, 1994). One possible way of addressing this multiple minimum problem is the use of reduced or simplified protein models (Kolinski & Skolnick, 1996; Skolnick & Kolinski, 1996). These represent each amino acid by a small number of united atoms, e.g., the α-carbons plus the side-chain center of mass positions. Such reduced models are designed to produce low to moderate resolution folds, after which recovery of full atomic detail becomes possible (Kolinski et al., 1993). Although a number of examples of de novo folding have been reported in the literature (Hansmann & Okamoto, 1993; Sun, 1993; Hao & Scheraga, 1995; Kolinski & Skolnick, 1996; Olszewski et al., 1996; Skolnick & Kolinski, 1996), it is clear that better, more specific potentials are necessary (Kolinski et al., 1996). Thus, the formulation of empirical energy functions consistent with a given reduced protein model has become an area of active investigation. In this paper, we present a new derivation of a knowledge-based, amino acid pair-specific potential that accounts for chain connectivity, compactness, and the presence of secondary structural elements. A key question is under what circumstances, if any, is the quasichemical approximation, which ignores these features of proteins and treats the chain as being chopped into individual residues, correct? Then, the performance of the set of derived potentials is assessed by the ability of the sequence to find the

Reprint requests to: Jeffrey Skolnick, Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Rd., La Jolla, California 92037; e-mail: skolnick@scripps.edu.

native topology in a variety of inverse folding tests of increasing rigor. Such tests constitute a minimal test of the potential; the ultimate arbiter of the validity of any potential is the ability to fold many proteins from the random state.

Over the years, a variety of amino acid pair-specific potentials have been formulated (Miyazawa & Jernigan, 1985, 1996). By way of illustration, we focus on potentials that are contact based, i.e., where an interaction between two residues is allowed if their distance is less than a given threshold, taken to be 4.5 Å between any pair of side-chain heavy atoms. In this instance, the average number of contacts per residue ranges from about 1.5 for Gly to 6.8 for Tyr. However, the general approach can be used to derive potentials of mean force of any functional form, in particular longer distance cutoffs can be used (Jernigan & Behar, 1996). In the classical approach formulated originally by Tanaka and Scheraga (1976) and subsequently followed by many other investigators, one uses a library of native-like structures to extract the relative observed frequency of side-chain contacts between a given pair of amino acids γ and μ, $p_{\gamma\mu}$. This frequency is then compared to that expected in some reference state where there are no specific side-chain interactions $p^{\circ}_{\gamma\mu}$. The potential of mean force between amino acids of type γ and μ is then given by

$$\epsilon_{\gamma\mu} = -k_B T \ln(p_{\gamma\mu}/p^{\circ}_{\gamma\mu}). \tag{1}$$

Thus, one implicitly assumes a Boltzmann distribution of energies holds for the distribution of contacts obtained in a library of native-like folds. Recently, a number of theoretical and empirical arguments have suggested that this assumption is justified (Bryant & Lawrence, 1991; Finkelstein et al., 1995).

The origin of the differences between extant pair interaction scales resides predominantly in the different choices of the reference state (Godzik, 1996). In a recent paper (Godzik et al., 1995), we presented a preliminary comparison of various scales and their associated reference states. Jernigan and Behar (1996) have also given an excellent review of this problem. In practice, different investigators have chosen various reference states. These include the unfolded state, a compact state of inert residues, a compact state where hydrophobic residues prefer to be buried, and an ideal mixture where the excess energy of mixing is zero (see Equations 3a and 3b) (Godzik et al., 1995).

To date, all statistical potentials based on Equation 1 have in common the quasichemical approximation, i.e., all neglect the connectivity of the chain. Basically, the chain is viewed as a collection of disconnected units (that may or may not be of different size) that undergo random mixing. Thus, the expected frequency of $\gamma\mu$ contacts in a quasichemical reference state is

$$p^{\circ}_{\gamma\mu} = f_\gamma f_\mu, \tag{2a}$$

where f_γ is a residue-dependent, compositional-based property such as the mole fraction, x_γ, or the side-chain contact fraction, ϕ_γ, of residue type γ. Here,

$$x_\gamma = \frac{n_\gamma}{\sum n_\eta} \tag{2b}$$

and

$$\phi_\gamma = \frac{n_\gamma z_\gamma}{\sum n_\eta z_\eta}. \tag{2c}$$

Here, n_γ is the number of residues and z_γ is the average number of contacts of residue type γ.

A number of conceptual problems are associated with the quasichemical approximation. First of all, it ignores chain connectivity. As pointed out by Jernigan and Bahar (1996), this neglect of chain connectivity and its attendant influence on the correlation between interactions might be a very severe approximation. It also ignores the presence of secondary structural elements in the native state. Ideally, one would like a reference state having both features, but where the interaction between pairs of residues is nonspecific. One path toward the creation of an approximate reference state that possesses such properties was suggested by Maiorov and Crippen (1992). They developed a potential by fitting a large number of parameters to a library of structures and demanded that the native sequence in its native structure has the lowest energy. Interestingly, we have shown (Godzik et al., 1995) that the resulting pair interaction scale is highly correlated to one of the pair scales derived by Miyazawa and Jernigan (1985, 1996), a knowledge-based, statistical scale based on the quasichemical approximation.

In this paper, we propose two new reference states for the calculation of the empirical pair interaction energy. The first reference state simply includes the restraint of chain connectivity and is based on the statistics of Gaussian random coil chains (Flory, 1953; Mattice & Suter, 1994). Gaussian random coil chains describe the conformational behavior of polymer chains that lack excluded volume interactions (Flory, 1953; Mattice & Suter, 1994). The advantage of this approach is that the contact probability can be calculated analytically, and thus, this reference state serves to illustrate the basic features of the more general treatment. Next, in the spirit of Maiorov and Crippen (1992), we propose a more realistic "native" reference state, whereby the reference contact probability is obtained by inserting or "threading" each sequence through a library of structures that are essentially as compact as the native state, but where all interaction specificity is ignored. This reference state accounts for the constraints of chain connectivity, protein chain compactness, and the presence of regular secondary structure.

In all cases, it is very illustrative to dissect a given interaction scale into its ideal and excess components (Godzik et al., 1995). The ideal component of the pair interaction energy is defined by

$$\epsilon_{\gamma\mu,ideal} = (\epsilon_{\gamma\gamma} + \epsilon_{\mu\mu})/2. \tag{3a}$$

This term may also include properties that depend on a single residue (e.g., burial preferences). In an infinite system, it does not provide any pair interaction specificity, but it can contribute in a finite system due to surface effects. In general, such specificity is provided by the excess pair interaction component defined by

$$\epsilon_{\gamma\mu,excess} = \epsilon_{\gamma\mu} - \epsilon_{\gamma\mu,ideal}, \tag{3b}$$

In previous work, we demonstrated that the excess component is essentially independent of the reference state used to derive the statistical potential (Godzik et al., 1995).

The outline of the remainder of this paper is as follows. We begin with a derivation of the compact Gaussian chain (Flory, 1953; Mattice & Suter, 1994) reference state and compare the resulting scale with that obtained using the quasichemical approximation. Next, we present the native reference state, which in-

cludes the effects of chain connectivity, protein compactness, and a native-like distribution of secondary structure on the expected contact frequency in a system lacking preferential side-chain interactions. We also explore the relationship of this scale to the corresponding scale based on the quasichemical approximation. Then, we examine a contact fraction-based reference state and clarify the difference between scales derived using mole fraction and contact fraction-based reference states. This is followed by a comparison of the various derived scales. To validate the scales in inverse folding tests, we examine their relative performance. We assess the relative ability of the potentials to do threading when gaps in the sequence are not permitted. This is termed "gapless threading" in what follows. In gapless threading calculations, the object is to match the sequence to its native structure. This test is a necessary but far from sufficient means of assessing the utility of a given interaction scale. For a set of test sequences, we then examine the ability of the potentials to identify the correct topology in a structural library, at least one of whose members has the same fold and where gaps are permitted. All proteins in this structural library have random homology to the test sequences. We conclude with a discussion of the significance of the results and the directions of future research.

Results

In what follows, we present two new derivations of the contact potential between pairs of amino acids. Furthermore, because a variety of scales based on different reference states are introduced throughout this paper, to avoid confusion, we summarize the properties of the reference states in Table 1A. The various scales that are based on the different reference states are summarized in Table 1B.

Gaussian chain reference state

By way of a very simple illustration, we consider a protein whose contact frequency obeys Gaussian chain statistics (Flory, 1953). Let the ℓth protein contain $N(\ell)$ residues and $C(\ell)$ total side-chain contacts. Let a_i be the amino acid at the ith position in the sequence. In a Gaussian random chain, the probability that the side chains at positions i and j in the sequence would be in contact and are occupied by amino acids of type γ and μ is

$$p_g(\gamma,\mu,i,j) = \nu(i-j)^{-1.5}\delta_{\gamma,ai}\delta_{\mu,aj}, \tag{4a}$$

ν is a constant that will cancel out in subsequent analysis and

Table 1. *Properties of the reference states and various scales that are based on the different reference states*

A. Summary of reference states used to estimate side-chain contact probability in noninteracting systems

Reference state	Description of reference state
Quasichemical-mole fraction reference state	Disconnected collection of units of identical size that interact randomly. Contact probability is the product of the mole fractions. See Equations 1, 2a, and 2b.
Quasichemical-contact fraction reference state	Disconnected collection of units that interact randomly and where each residue type has a different number of contacts. Contact probability is the product of the contact fractions. See Equations 1, 2a and 2c.
Gaussian reference state	Contact probability is calculated from that exhibited by a Gaussian-random coil polymer chain. See Equations 7, 9c, and 9d.
Native reference state	Library of structures excised from real proteins that have comparable compactness as the native fold. The presence of protein-like secondary structure and packing patterns are included, but there is no pair specificity. The side-chain contact maps that dictate the allowed interactions are static and excised from the experimental structure. See Equations 12a–12d.

B. Summary of various scales and the reference states on which they are based

Scale	Description of reference state
Gaussian	Gaussian chain reference state. Accounts for chain connectivity in the calculation of the expected contact frequency in a randomly interacting system.
Native	Native reference state. Accounts for chain connectivity, secondary structure, and protein compactness in the calculation of the expected contact frequency in a randomly interacting system.
Native-filtered	Based on the native scale, but considers only strongly interacting pairs of residues.
Contact fraction-averaged	Quasichemical-contact fraction-based reference state that includes both buried and exposed residues and calculates the contact fraction from the actual protein structure. See Equations 16a and 16b.
GKS scale	Quasichemical-contact fraction-based reference state that includes only buried residues and calculates the contact fraction from the actual protein structure (Godzik et al., 1992).
Native-contact fraction	Quasichemical-based reference state threads each sequence through a library of structures of comparable compactness to the native state and calculates the expected number of contacts using the contact fraction of every residue type in each of these compact structures. See Equation 21a.

$$\delta_{\gamma,a_i} = \begin{cases} 1 & \text{if } \gamma = a_i \\ 0 & \text{otherwise} \end{cases}. \tag{4b}$$

δ_{γ,a_i} is the probability that the amino acid at position i is occupied by γ.

Observe that p_g can be written as the product of a structural part that depends on the location in the structure and a part that depends on the amino acid identities at positions i and j. Parenthetically, we note that the total number of contacts in a Gaussian chain is $4(\sqrt{N(\ell)} - 1)^2$, which, for reasonable values of $N(\ell)$ is $4N(\ell)$. The proportionality of the number of contacts with protein size is consistent with observations on protein structures.

Thus, the total probability that residues of type γ and μ will be in contact in the ℓth structure is

$$P_g(\gamma,\mu,\ell) = \sum_{i=1}^{N(\ell)} \sum_{j=1}^{N(\ell)} p_g(\gamma,\mu,i,j)\delta_{\gamma,a_i}\delta_{\mu,a_j}, \tag{5a}$$

The relative probability that amino acids γ and μ are in contact is

$$P_{g,rel}(\gamma,\mu,\ell) = \frac{P_g(\gamma,\mu,\ell)}{\sum_{\epsilon=1}^{20}\sum_{\eta=1}^{20} p_g(\epsilon,\eta,\ell)}. \tag{5b}$$

Let us define the Gaussian chain contact probability by

$$\rho(i,j) = \nu(|i-j|)^{-1.5}. \tag{6a}$$

Substituting Equation 5a into Equation 5b gives

$$P_{g,rel}(\gamma,\mu,\ell) = \frac{\sum_{i=1}^{N(\ell)}\sum_{j=1}^{N(\ell)} \rho(i,j)\delta_{\gamma,a_i}\delta_{\mu,a_j}}{\sum_i \sum_j \rho(i,j)}. \tag{6b}$$

The denominator of Equation 6b follows from Equation 5b, because the summation is over all possible pairs of amino acid types.

The expected number of contacts for a chain whose relative contact probability is determined by Gaussian chain statistics is

$$N_{\text{exp},gauss}(\gamma,\mu,\ell) = C(\ell)P_{g,rel}(\gamma,\mu,\ell). \tag{7}$$

Now, the total number of observed contacts in the ℓth protein, $C(\ell)$, can be calculated from

$$C(\ell) = \sum_\gamma \sum_\mu N_{obs}(\gamma,\mu,\ell). \tag{8}$$

$N_{obs}(\gamma,\mu,\ell)$ is the actual number of observed side-chain contacts between amino acids in the ℓth protein of type γ and μ.

Assuming that the contacts between amino acid pairs in the database of S_{tot} protein structures obey a Boltzmann distribution, the pair interaction energy between amino acids of type γ and μ is given by

$$\epsilon_{gauss}(\gamma,\mu) = -k_B T \ln\left(\frac{\sum_{\ell=1}^{S_{tot}} N_{obs}(\gamma,\mu,\ell)}{\sum_{\ell=1}^{S_{tot}} N_{\text{exp},gauss}(\gamma,\mu,\ell)}\right). \tag{9a}$$

Here, k_B is Boltzmann's constant, and T is the absolute temperature. The scale that uses the Gaussian chain approximation to the reference state to estimate the relative contact probability is called "Gaussian" in what follows. The parameters for this scale are compiled in Table 2.

A slightly more convenient formulation that holds when the total number of expected and observed contacts are different is obtained by expressing Equation 9a in terms of the ratio of the observed and expected contact probabilities for residues of type γ and μ in all S_{tot} protein structures, that is,

$$P_{obs}(\gamma,\mu) = \frac{\sum_{\ell=1}^{S_{tot}} N_{obs}(\gamma,\mu,\ell)}{\sum_{\ell=1}^{S_{tot}} C(\ell)}, \tag{9b}$$

and

$$P_{\text{exp},gauss}(\gamma,\mu) = \frac{\sum_{\ell=1}^{S_{tot}} N_{\text{exp},gauss}(\gamma,\mu,\ell)}{\sum_{\ell=1}^{S_{tot}} C(\ell)}. \tag{9c}$$

Thus, Equation 9a can be rewritten as

$$\epsilon_{gauss}(\gamma,\mu) = -k_B T \ln\left(\frac{P_{obs}(\gamma,\mu)}{P_{\text{exp},gauss}(\gamma,\mu)}\right). \tag{9d}$$

Relationship of the Gaussian chain and quasichemical-mole fraction-based, reference states

We first observe that $\rho(i,j)$ defined in Equation 6a is a function of $|i-j|$. If we neglect essentially inconsequential end effects, the sum over amino acids of type γ is simply $x_\gamma(\ell)$ times the sum over all i, where $x_\gamma(\ell)$ is the mole fraction of amino acid type γ in the ℓth structure.

Now, we invoke the fact that the probability of finding an amino acid at any position j is random and is just proportional to the mole fraction of μ. These two approximations give the quasichemical-mole fraction-based approximation for $P_{g,rel}$:

$$P^{quasi}_{g,rel}(A_\gamma,A_\mu,\ell) = x_\gamma(\ell)x_\mu(\ell). \tag{10a}$$

In fact, we note that Equation 10a holds for any functional form $\rho(i,j) = \rho(|i-j|)$. Thus, with any contact function of this type and given a random amino acid sequence distribution, $P_{g,rel}$ should be well approximated by the quasichemical approximation. Finally, we note that, even if $\rho(i,j) \neq \rho(|i-j|)$, we can repeat the derivation by approximating Equation 6a as

$$\sum_{i=1}^{N(\ell)}\sum_{j=1}^{N(\ell)} \rho(i,j)\delta_{\gamma,a_i}\delta_{\mu,a_j} \approx \sum_{i=1}^{N(\ell)}\sum_{j=1}^{N(\ell)} \rho(i,j)\langle\delta_{\gamma,a_i}\delta_{\mu,a_j}\rangle, \tag{10b}$$

with

$$\langle\delta_{\gamma,a_i}\delta_{\mu,a_j}\rangle = x_\gamma(\ell)x_\mu(\ell). \tag{10c}$$

where the brackets denote the average over all positions i and j in the chain. Thus, we are factorizing the contact probability into a geometric part and an amino acid composition-dependent part.

Table 2. *Pair interaction parameters derived using the Gaussian chain reference state*

	Gly	Ala	Ser	Cys	Val	Thr	Ile	Pro	Met	Asp	Asn	Leu	Lys	Glu	Gln	Arg	His	Phe	Tyr	Trp
Gly	1.8	1.4	1.0	1.1	0.8	0.6	0.6	1.0	0.6	0.6	0.6	0.6	0.7	1.0	0.7	0.3	0.8	0.4	0.2	0.2
Ala	1.4	1.1	0.9	0.5	−0.1	0.6	−0.3	0.8	−0.2	0.9	0.8	−0.1	1.0	1.1	0.6	0.5	0.4	−0.2	−0.2	−0.5
Ser	1.0	0.9	0.6	0.4	0.5	0.4	0.4	0.7	0.3	0.3	0.5	0.4	0.6	0.2	0.3	0.2	0.0	0.2	0.2	0.1
Cys	1.1	0.5	0.4	−1.7	−0.5	0.2	−0.6	0.3	−0.3	0.5	0.3	−0.4	1.0	0.8	0.2	0.4	−0.1	−0.8	−0.5	−0.9
Val	0.8	−0.1	0.5	−0.5	−1.0	−0.1	−1.1	0.2	−0.7	0.9	0.5	−1.1	0.5	0.5	0.2	0.0	0.1	−1.1	−0.7	−1.0
Thr	0.6	0.6	0.4	0.2	−0.1	0.2	−0.2	0.4	−0.1	0.1	0.1	−0.1	0.5	0.1	0.2	−0.1	0.0	−0.2	−0.2	−0.2
Ile	0.6	−0.3	0.4	−0.6	−1.1	−0.2	−1.3	0.1	−0.9	0.5	0.5	−1.3	0.3	0.3	0.0	−0.2	0.0	−1.2	−0.9	−1.3
Pro	1.0	0.8	0.7	0.3	0.2	0.4	0.1	0.7	−0.1	1.0	0.5	0.1	0.7	0.6	0.2	0.1	0.1	0.0	−0.4	−0.7
Met	0.6	−0.2	0.3	−0.3	−0.7	−0.1	−0.9	−0.1	−1.1	0.5	0.2	−1.0	0.3	0.1	−0.1	0.2	−0.3	−1.3	−1.0	−1.4
Asp	0.6	0.9	0.3	0.5	0.9	0.1	0.5	1.0	0.5	0.4	0.0	0.6	−0.2	0.4	0.2	−0.6	−0.1	0.4	−0.1	−0.1
Asn	0.6	0.8	0.5	0.3	0.5	0.1	0.5	0.5	0.2	0.0	−0.1	0.4	0.2	0.1	−0.1	−0.1	0.0	0.0	−0.2	−0.1
Leu	0.6	−0.1	0.4	−0.4	−1.1	−0.1	−1.3	0.1	−1.0	0.6	0.4	−1.2	0.3	0.4	0.1	−0.1	−0.1	−1.3	−0.9	−1.3
Lys	0.7	1.0	0.6	1.0	0.5	0.5	0.3	0.7	0.3	−0.2	0.2	0.3	1.5	−0.4	0.2	0.7	0.5	0.3	−0.4	−0.1
Glu	1.0	1.1	0.2	0.8	0.5	0.1	0.3	0.6	0.1	0.4	0.1	0.4	−0.4	0.9	0.3	−0.6	0.0	0.3	−0.2	−0.2
Gln	0.7	0.6	0.3	0.2	0.2	0.2	0.0	0.2	−0.1	0.2	−0.1	0.1	0.2	0.3	−0.1	0.0	0.0	−0.1	−0.3	−0.4
Arg	0.3	0.5	0.2	0.4	0.0	−0.1	−0.2	0.1	0.2	−0.6	−0.1	−0.1	0.7	−0.6	0.0	−0.3	−0.2	−0.3	−0.6	−0.6
His	0.8	0.4	0.0	−0.1	0.1	0.0	0.0	0.1	−0.3	−0.1	0.0	−0.1	0.5	0.0	0.0	−0.2	−0.4	−0.4	−0.7	−0.8
Phe	0.4	−0.2	0.2	−0.8	−1.1	−0.2	−1.2	0.0	−1.3	0.4	0.0	−1.3	0.3	0.3	−0.1	−0.3	−0.4	−1.5	−1.0	−1.5
Tyr	0.2	−0.2	0.2	−0.5	−0.7	−0.2	−0.9	−0.4	−1.0	−0.1	−0.2	−0.9	−0.4	−0.2	−0.3	−0.6	−0.7	−1.0	−0.8	−1.2
Trp	0.2	−0.5	0.1	−0.9	−1.0	−0.2	−1.3	−0.7	−1.4	−0.1	−0.1	−1.3	−0.1	−0.2	−0.4	−0.6	−0.8	−1.5	−1.2	−1.4

Using Equation 10c, the quasichemical approximation to Equation 7 is

$$N^{quasi}_{exp,gauss}(A_\gamma, A_\mu, \ell) = C(\ell) x_\gamma(\ell) x_\mu(\ell). \tag{11a}$$

The quasichemical-mole fraction-based approximation to the scale obtained using the Gaussian reference state is:

$$\epsilon^{quasi}_{gauss}(\gamma,\mu) = -k_B T \ln\left(\frac{\sum_{\ell=1}^{S_{tot}} N_{obs}(\gamma,\mu,\ell)}{\sum_{\ell=1}^{S_{tot}} N^{quasi}_{exp,gauss}(\gamma,\mu,\ell)} \right). \tag{11b}$$

The correlation coefficient between the two scales defined in Equations 9a and 11b is 0.99. We can also calculate the difference between the energy obtained using the Gaussian reference state and that obtained from the quasichemical-mole fraction approximation. The average value of this difference is $0.043 k_B T$. The correction term is small because the distribution of amino acids along the protein sequence is essentially random, and one can factorize the contact probability in a system without preferential side-chain interactions into a structural part and an amino acid composition-dependent part. Thus, we conclude that the quasichemical approximation to the Gaussian reference state is, on average, excellent.

Native reference state

Next, we present a new derivation of the contact potential between pairs of amino acids designed to account for chain connectivity, the fact that native conformations of proteins are compact, and that they possess substantial amounts of regular secondary structure. That is, we wish to obtain the potential of mean force between residues, suitably (or at least approximately) corrected for the presence of protein-like environments, but where no specific interactions between pairs of residues occur. In this native reference state, we estimate the expected contact probability for different pairs of amino acids where there are no specific amino acid pair interactions by threading each protein sequence through a library of compact fragments excised from native protein structures. To enforce native-like compactness, the radius of gyration of all such excised protein fragments must be within 10% of the native conformation of the sequence of interest.

For the kth such compact subfragment in the mth protein in a library of S_{tot} total structures, the probability that residues γ and μ are in contact is

$$p^{\circ}(\gamma,\mu,N(\ell);k,m) = \frac{\sum_{i=1}^{N(\ell)} \sum_{j=1}^{N(\ell)} C_{i+k,j+k}(m)\delta_{\gamma,a_i}\delta_{\mu,a_j}}{\sum_{i=1}^{N(\ell)} \sum_{j=1}^{N(\ell)} C_{i+k,j+k}(m)}, \tag{12a}$$

where $N(\ell) \le N(m)$; $C_{ij}(m) = 1$ (0) if residues at positions i and j in the mth protein are (not) in contact. In Equation 12a, the denominator is just the total number of contacts in the compact subfragment starting at residue k and finishing at residue $N(\ell) + k - 1$ in the mth protein and is given by

$$C(k,m,\ell) = \sum_{i=1}^{N(\ell)} \sum_{j=1}^{N(\ell)} C_{i+k,j+k}(m). \tag{12b}$$

Note that the contact map is assumed to remain the same when the sequence from the ℓth protein is threaded into the structure of the mth protein. This is a crucial approximation whose effect on the resulting potential requires additional investigation. However, it would be true if all amino acids were of the same size so that a sequence can be threaded into one of these protein-like structures without causing the side chains to repack.

The total probability that residues of type γ and μ will be in contact in the native reference state is

$$p^{\circ}(\gamma,\mu) = \frac{\sum_{\ell=1}^{S_{tot}} \sum_{m=1}^{S_{tot}} \sum_{k=0;\text{compact}}^{N(m)-N(\ell)} p^{\circ}(\gamma,\mu,N(\ell);k,m)C(k,m,\ell)}{\sum_{\ell=1}^{S_{tot}} \sum_{m=1}^{S_{tot}} \sum_{k=0;\text{compact}}^{N(m)-N(\ell)} C(k,m,\ell)}. \tag{12c}$$

In Equation 12c, the numerator is simply the number of contacts between amino acids of type γ and μ in appropriately compact substructures, and the denominator is simply the total number of contacts in the entire library of such compact subfragments.

We next require the relative contact probability that is actually observed in the library of S_{tot} structures, a quantity given by Equation 9b. If we once again assume that the observed contact frequency in the structural database obeys a Boltzmann-like distribution, then the potential of mean force of a γ and μ contact relative to the case where they lack pair-specific interactions is

$$\epsilon_{native}(\gamma,\mu) = -k_B T \ln(P_{obs}(\gamma,\mu)/P^{\circ}(\gamma,\mu)). \tag{12d}$$

The resulting scale based on the native reference state is presented in Table 3A.

Relationship of the native and quasichemical-mole fraction-based reference states

The quasichemical-mole fraction-based approximation to Equation 12a is

$$p^{\circ}_{quasi}(\gamma,\mu,N(\ell);k,m) = x_{\gamma}(\ell)x_{\mu}(\ell), \tag{13a}$$

and to Equation 12c is

$$P^{\circ}_{quasi}(\gamma,\mu) = \frac{\sum_{\ell=1}^{S_{tot}} x_{\gamma}(\ell)x_{\mu}(\ell) \sum_{m=1}^{S_{tot}} \sum_{k=0;\text{compact}}^{N(m)-N(\ell)} C(k,m,\ell)}{\sum_{\ell=1}^{S_{tot}} \sum_{m=1}^{S_{tot}} \sum_{k=0;\text{compact}}^{N(m)-N(\ell)} C(k,m,\ell)}. \tag{13b}$$

Thus, the scale based on the quasichemical-mole fraction-based approximation to the native-like reference state, native-mole fraction is

$$\epsilon_{native-mole\,fraction}(\gamma,\mu) = -k_B T \ln\left(\frac{P_{obs}(\gamma,\mu)}{P^{\circ}_{quasi}(\gamma,\mu)}\right). \tag{13c}$$

The correlation coefficient between the scale based on the native reference state, Equation 12d, and the quasichemical approximation to the native reference state is 0.96. The mean value of the difference between the pair energy calculated on the basis of the native reference state and the analogous quantity calculated in the quasichemical approximation, i.e., the native-mole fraction reference state, is 0.027. This demonstrates that the quasichemical approximation is again an excellent one. It does not arise from neglect of chain connectivity or the presence of protein-like secondary structure (Equation 12a explicitly includes both features). Rather, this result emerges because the amino acid sequence distribution is essentially random and repacking effects accompanying the threading of a given sequence into another structure are ignored.

Filtered version of the scale based on the native reference state

We have also considered a filtered version of the pair interaction scale based on the native reference state. The "native-filtered" scale only considers strongly interacting residues and sets the remainder of the interactions to zero. If $|\epsilon_{native}(\gamma,\mu)| < 0.8$, then set $\epsilon_{native,filtered}(\gamma,\mu) = 0$. Otherwise, if $\epsilon_{native}(\gamma,\mu) > 0.8$, then it is replaced by the average value of ϵ_{native} over all such residue pairs, $\epsilon_{p,av} = 1.1$. Similarly, if $\epsilon_{native}(\gamma,\mu) < -0.8$, then it is replaced by the average value over all such residue pairs, $\epsilon_{n,av} = -1.2$.

Quasichemical-contact fraction-based reference state

We next consider a reference state where the frequency of expected side-chain contacts is calculated from the contact fraction of residues defined in Equation 2c; this is a quasichemical-contact fraction-based reference state averaged over a library of structures; hence, it is called "contact fraction-averaged." Let the total number of side-chain contacts made by amino acids γ and μ in the ℓth structure be $N_{obs}(\gamma,\mu,\ell)$. For the ℓth structure, the fraction of contacts made by amino acid type γ is

$$\phi_{\gamma}(\ell) = \frac{\sum_{\mu} N_{obs}(\gamma,\mu,\ell)}{C(\ell)}. \tag{14}$$

The numerator is the total number of actual contacts made by residues of type γ with all other residues in the ℓth structure. The denominator is the total number of contacts in the ℓth structure and is given by Equation 8. The expected number of contacts between amino acids γ and μ is

$$N_{\exp,\phi} = C(\ell)\phi_{\gamma}(\ell)\phi_{\mu}(\ell). \tag{15}$$

The pair potential in the quasichemical-contact fraction-averaged approximation is simply

$$\epsilon_{\phi}(\gamma,\mu) = -k_B T \ln\left(\frac{P_{obs}(\gamma,\mu)}{P_{\phi}(\gamma,\mu)}\right), \tag{16a}$$

where the observed contact probability is given by Equation 9b, and the expected contact probability is

$$P_{\phi}(\gamma,\mu) = \frac{\sum_{\ell=1}^{S_{tot}} C(\ell)\phi_{\gamma}(\ell)\phi_{\mu}(\ell)}{\sum_{\ell=1}^{S_{tot}} C(\ell)}. \tag{16b}$$

The resulting pair scale differs from the contact fraction-based, GKS scale derived originally by Godzik et al. (1992, 1995) in that the GKS scale considers only buried residues.

Comparison of the scales based on different reference states

As conjectured by a number of investigators (Park & Levitt, 1996), the values of the energy parameters presented in Tables 2 and 3, as well as other scales (Godzik et al., 1997), suggest that one does not need a full set of 210 parameters (involving the 20 amino acids) to describe the interactions between the various amino acids. Rather,

Table 3. *Pair interaction parameters, ideal energy contribution to the pair interaction scale, and excess energy contribution to the pair interaction scale, all derived using the native-like reference state*

	Gly	Ala	Ser	Cys	Val	Thr	Ile	Pro	Met	Asp	Asn	Leu	Lys	Glu	Gln	Arg	His	Phe	Tyr	Trp
A. Pair interaction parameters																				
Gly	1.5	1.2	0.8	1.3	0.5	0.5	0.4	0.8	0.5	0.8	0.7	0.4	0.9	1.1	0.8	0.2	0.7	0.3	0.1	0.0
Ala	1.2	0.8	0.9	0.6	−0.4	0.5	−0.6	0.6	−0.3	1.2	0.9	−0.3	1.3	1.2	0.6	0.6	0.4	−0.2	−0.4	−0.6
Ser	0.8	0.9	0.6	0.6	0.4	0.4	0.3	0.6	0.4	0.7	0.7	0.4	0.9	0.6	0.6	0.2	0.0	0.1	0.1	−0.1
Cys	1.3	0.6	0.6	−1.3	−0.6	0.5	−0.5	0.4	−0.3	0.8	0.6	−0.4	1.3	0.9	0.3	0.7	0.1	−0.6	−0.1	−0.7
Val	0.5	−0.4	0.4	−0.6	−1.2	−0.2	−1.2	−0.1	−0.8	1.0	0.5	−1.2	0.7	0.5	0.2	0.0	0.0	−1.1	−0.8	−1.1
Thr	0.5	0.5	0.4	0.5	−0.2	0.1	−0.3	0.2	0.0	0.5	0.3	0.0	0.8	0.4	0.4	0.0	0.1	−0.2	−0.2	−0.2
Ile	0.4	−0.6	0.3	−0.5	−1.2	−0.3	−1.4	0.0	−1.0	0.5	0.5	−1.3	0.5	0.4	0.1	−0.2	−0.1	−1.3	−1.0	−1.3
Pro	0.8	0.6	0.6	0.4	−0.1	0.2	0.0	0.4	−0.2	0.9	0.5	0.0	0.9	0.7	0.2	0.1	0.0	−0.1	−0.5	−0.8
Met	0.5	−0.3	0.4	−0.3	−0.8	0.0	−1.0	−0.2	−1.1	0.6	0.3	−1.0	0.6	0.4	0.1	0.2	−0.4	−1.3	−1.1	−1.5
Asp	0.8	1.2	0.7	0.8	1.0	0.5	0.5	0.9	0.6	0.9	0.4	0.7	0.2	0.9	0.6	−0.5	−0.1	0.5	−0.2	0.0
Asn	0.7	0.9	0.7	0.6	0.5	0.3	0.5	0.5	0.3	0.4	0.1	0.4	0.7	0.5	0.2	0.0	0.2	0.1	−0.2	0.0
Leu	0.4	−0.3	0.4	−0.4	−1.2	0.0	−1.3	0.0	−1.0	0.7	0.4	−1.2	0.5	0.6	0.2	−0.1	−0.1	−1.3	−0.9	−1.4
Lys	0.9	1.3	0.9	1.3	0.7	0.8	0.5	0.9	0.6	0.2	0.7	0.5	2.1	0.0	0.5	1.1	1.0	0.4	−0.2	−0.1
Glu	1.1	1.2	0.6	0.9	0.5	0.4	0.4	0.7	0.4	0.9	0.5	0.6	0.0	1.2	0.8	−0.4	0.1	0.4	−0.2	−0.1
Gln	0.8	0.6	0.6	0.3	0.2	0.4	0.1	0.2	0.1	0.6	0.2	0.2	0.5	0.8	0.3	0.2	0.1	−0.1	−0.3	−0.4
Arg	0.2	0.6	0.2	0.7	0.0	0.0	−0.2	0.1	0.2	−0.5	0.0	−0.1	1.1	−0.4	0.2	−0.1	−0.1	−0.4	−0.7	−0.6
His	0.7	0.4	0.0	0.1	0.0	0.1	−0.1	0.0	−0.4	−0.1	0.2	−0.1	1.0	0.1	0.1	−0.1	−0.8	−0.4	−0.8	−0.9
Phe	0.3	−0.2	0.1	−0.6	−1.1	−0.2	−1.3	−0.1	−1.3	0.5	0.1	−1.3	0.4	0.4	−0.1	−0.4	−0.4	−1.5	−1.0	−1.5
Tyr	0.1	−0.4	0.1	−0.1	−0.8	−0.2	−1.0	−0.5	−1.1	−0.2	−0.2	−0.9	−0.2	−0.2	−0.3	−0.7	−0.8	−1.0	−0.8	−1.2
Trp	0.0	−0.6	−0.1	−0.7	−1.1	−0.2	−1.3	−0.8	−1.5	0.0	0.0	−1.4	−0.1	−0.1	−0.4	−0.6	−0.9	−1.5	−1.2	−1.2
B. Ideal energy contribution to the pair interaction scale																				
Gly	1.5	1.1	1.0	0.1	0.1	0.8	0.1	0.9	0.2	1.2	0.8	0.1	1.8	1.4	0.9	0.7	0.3	0.0	0.3	0.1
Ala	1.1	0.8	0.7	−0.2	−0.2	0.5	−0.3	0.6	−0.2	0.9	0.5	−0.2	1.4	1.0	0.6	0.3	0.0	−0.3	0.0	−0.2
Ser	1.0	0.7	0.6	−0.3	−0.3	0.4	−0.4	0.5	−0.2	0.8	0.4	−0.3	1.3	0.9	0.5	0.2	−0.1	−0.4	−0.1	−0.3
Cys	0.1	−0.2	−0.3	−1.3	−1.2	−0.6	−1.3	−0.4	−1.2	−0.2	−0.6	−1.2	0.4	0.0	−0.5	−0.7	−1.0	−1.4	−1.0	−1.2
Val	0.1	−0.2	−0.3	−1.2	−1.2	−0.6	−1.3	−0.4	−1.2	−0.2	−0.6	−1.2	0.4	0.0	−0.5	−0.7	−1.0	−1.4	−1.0	−1.2
Thr	0.8	0.5	0.4	−0.6	−0.6	0.1	−0.6	0.2	−0.5	0.5	0.1	−0.6	1.1	0.7	0.2	0.0	−0.3	−0.7	−0.3	−0.6
Ile	0.1	−0.3	−0.4	−1.3	−1.3	−0.6	−1.4	−0.5	−1.2	−0.2	−0.6	−1.3	0.3	−0.1	−0.5	−0.8	−1.1	−1.5	−1.1	−1.3
Pro	0.9	0.6	0.5	−0.4	−0.4	0.2	−0.5	0.4	−0.4	0.6	0.2	−0.4	1.2	0.8	0.4	0.2	−0.2	−0.6	−0.2	−0.4
Met	0.2	−0.2	−0.2	−1.2	−1.2	−0.5	−1.2	−0.4	−1.1	−0.1	−0.5	−1.2	0.5	0.1	−0.4	−0.6	−1.0	−1.3	−1.0	−1.2

Asp	1.2	0.9	0.8	−0.2	−0.2	0.5	−0.2	0.6	−0.1	0.9	0.5	−0.2	1.5	1.0	0.6	0.4	0.0	−0.3	0.0	−0.2
Asn	0.8	0.5	0.4	−0.6	−0.6	0.1	−0.6	0.2	−0.5	0.5	0.1	−0.6	1.1	0.7	0.2	0.0	−0.3	−0.7	−0.3	−0.6
Leu	0.1	−0.2	−0.3	−1.2	−1.2	−0.6	−1.3	−0.4	−1.2	−0.2	−0.6	−1.2	0.4	0.0	−0.5	−0.7	−1.0	−1.4	−1.0	−1.2
Lys	1.8	1.4	1.3	0.4	0.4	1.1	0.3	1.2	0.5	1.5	1.1	0.4	2.1	1.6	1.2	1.0	0.6	0.3	0.6	0.4
Glu	1.4	1.0	0.9	0.0	0.0	0.7	−0.1	0.8	0.1	1.0	0.7	0.0	1.6	1.2	0.8	0.6	0.2	−0.1	0.2	0.0
Gln	0.9	0.6	0.5	−0.5	−0.5	0.2	−0.5	0.4	−0.4	0.6	0.2	−0.5	1.2	0.8	0.3	0.1	−0.2	−0.6	−0.2	−0.5
Arg	0.7	0.3	0.2	−0.7	−0.7	0.0	−0.8	0.2	−0.6	0.4	0.0	−0.7	1.0	0.6	0.1	−0.1	−0.5	−0.8	−0.5	−0.7
His	0.3	0.0	−0.1	−1.0	−1.0	−0.3	−1.1	−0.2	−1.0	0.0	−0.3	−1.0	0.6	0.2	−0.2	−0.5	−0.8	−1.1	−0.8	−1.0
Phe	0.0	−0.3	−0.4	−1.4	−1.4	−0.7	−1.5	−0.6	−1.3	−0.3	−0.7	−1.4	0.3	−0.1	−0.6	−0.8	−1.1	−1.5	−1.1	−1.4
Tyr	0.3	0.0	−0.1	−1.0	−1.0	−0.3	−1.1	−0.2	−1.0	0.0	−0.3	−1.0	0.6	0.2	−0.2	−0.5	−0.8	−1.1	−0.8	−1.0
Trp	0.1	−0.2	−0.3	−1.2	−1.2	−0.6	−1.3	−0.4	−1.2	−0.2	−0.6	−1.2	0.4	0.0	−0.5	−0.7	−1.0	−1.4	−1.0	−1.2
C. Excess energy contribution to the pair interaction scale																				
Gly	0.0	0.1	−0.2	1.2	0.4	−0.3	0.3	−0.1	0.3	−0.4	−0.1	0.3	−0.9	−0.2	−0.1	−0.5	0.3	0.3	−0.2	−0.1
Ala	0.1	0.0	0.2	0.9	−0.2	0.0	−0.3	0.0	−0.2	0.4	0.4	−0.1	−0.1	0.2	0.1	0.3	0.4	0.1	−0.4	−0.4
Ser	−0.2	0.2	0.0	0.9	0.7	0.0	0.7	0.1	0.6	−0.1	0.3	0.7	−0.4	−0.3	0.2	0.0	0.1	0.6	0.2	0.2
Cys	1.2	0.9	0.9	0.0	0.6	1.1	0.8	0.9	0.9	1.0	1.2	0.9	0.9	0.9	0.8	1.4	1.1	0.8	0.9	0.6
Val	0.4	−0.2	0.7	0.6	0.0	0.4	0.1	0.3	0.4	1.2	1.0	0.0	0.3	0.5	0.7	0.7	1.0	0.2	0.2	0.1
Thr	−0.3	0.0	0.0	1.1	0.4	0.0	0.3	0.0	0.5	0.0	0.2	0.6	−0.3	−0.3	0.2	0.0	0.4	0.5	0.1	0.4
Ile	0.3	−0.3	0.7	0.8	0.1	0.3	0.0	0.5	0.2	0.8	1.1	0.0	0.2	0.5	0.6	0.6	1.0	0.2	0.1	0.0
Pro	−0.1	0.0	0.1	0.9	0.3	0.0	0.5	0.0	0.2	0.2	0.2	0.4	−0.4	−0.1	−0.2	−0.1	0.2	0.5	−0.3	−0.4
Met	0.3	−0.2	0.6	0.9	0.4	0.5	0.2	0.2	0.0	0.7	0.8	0.2	0.1	0.3	0.5	0.8	0.6	0.0	−0.1	−0.3
Asp	−0.4	0.4	−0.1	1.0	1.2	0.0	0.8	0.2	0.7	0.0	−0.1	0.9	−1.3	−0.1	0.0	−0.9	−0.1	0.8	−0.2	0.2
Asn	−0.1	0.4	0.3	1.2	1.0	0.2	1.1	0.2	0.8	−0.1	0.0	1.0	−0.4	−0.2	0.0	0.0	0.6	0.8	0.1	0.6
Leu	0.3	−0.1	0.7	0.9	0.0	0.6	0.0	0.4	0.2	0.9	1.0	0.0	0.1	0.6	0.7	0.6	0.9	0.1	0.1	−0.2
Lys	−0.9	−0.1	−0.4	0.9	0.3	−0.3	0.2	−0.4	0.1	−1.3	−0.4	0.1	0.0	−1.6	−0.7	0.1	0.4	0.1	−0.8	−0.5
Glu	−0.2	0.2	−0.3	0.9	0.5	−0.3	0.5	−0.1	0.3	−0.1	−0.2	0.6	−1.6	0.0	0.1	−1.0	−0.1	0.5	−0.4	−0.1
Gln	−0.1	0.1	0.2	0.8	0.7	0.2	0.6	−0.2	0.5	0.0	0.0	0.7	−0.7	0.1	0.0	0.1	0.3	0.5	−0.1	0.1
Arg	−0.5	0.3	0.0	1.4	0.7	0.0	0.6	−0.1	0.8	−0.9	0.0	0.6	0.1	−1.0	0.1	0.0	0.4	0.4	−0.2	0.1
His	0.3	0.4	0.1	1.1	1.0	0.4	1.0	0.2	0.6	−0.1	0.6	0.9	0.4	−0.1	0.3	0.4	0.0	0.8	0.0	0.1
Phe	0.3	0.1	0.6	0.8	0.2	0.5	0.2	0.5	0.0	0.8	0.8	0.1	0.1	0.5	0.5	0.4	0.8	0.0	0.1	−0.1
Tyr	−0.2	−0.4	0.2	0.9	0.2	0.1	0.1	−0.3	−0.1	−0.2	0.1	0.1	−0.8	−0.4	−0.1	−0.2	0.0	0.1	0.0	−0.2
Trp	−0.1	−0.4	0.2	0.6	0.1	0.4	0.0	−0.4	−0.3	0.2	0.6	−0.2	−0.5	−0.1	0.1	0.1	0.1	−0.1	−0.2	0.0

the amino acids can be clustered into groups. Various criteria for clustering these scales are explored in a series of follow-up papers that examine not only the parameter sets described here, but other parameter sets as well (Godzik et al., 1997).

In what follows, we discuss scales based on a total of four different reference states, whose assumptions are summarized in Table 1A. There are the two quasichemical-based scales that use either mole fraction- or contact fraction-based units. Then, there is the Gaussian chain reference state, which includes chain connectivity. Finally, there is the native reference state, in which effects of chain connectivity, compactness, and secondary structure are all included. Furthermore, given a reference state, various ways of actually implementing the calculation of the expected number of contacts exist. Table 1B summarizes the six scales that are based on various approximations to the reference states presented in Table 1A. In Table 4, the correlation coefficients between six scales, native, native-filtered, Gaussian, contact fraction-averaged, GKS (Godzik et al., 1992), and native-contact fraction, which is defined below, are presented. Each scale takes its name from the reference state upon which it is based. The Gaussian scale and the native scale are highly correlated, with a correlation coefficient, r, of 0.96. This makes sense because the reference states in both scales are well approximated by a quasichemical-mole fraction-based reference state. They are even more correlated than the native and native-filtered scales, which share the native reference state, but where the latter is a truncated version of the scale. Similarly, the quasichemical scales that employ contact fractions for the calculation of the contact frequency without specific interaction preferences, GKS and contact fraction-averaged, are highly correlated, with a correlation coefficient $r = 0.73$, but they are basically uncorrelated with all scales derived with a mole fraction-based reference state. The origin of the differences between the two kinds of scales based on mole fraction and contact fraction is addressed in the next section.

Difference between the mole fraction- and contact fraction-based energy scales

We consider here a simplified derivation that demonstrates the essential difference between energy scales derived on the basis of the mole fraction- and contact fraction-based reference states. Consider a large system containing C total contacts. Then, the mole fraction-based pair interaction energies is given by

$$\epsilon_x(\gamma,\mu) = -k_B T \ln \frac{N_{obs}(\gamma,\mu)}{C x_\gamma x_\mu}, \tag{17}$$

with x_γ defined in Equation 2b. Similarly, the contact fraction-based pair interaction energy is given by

$$\epsilon_\phi(\gamma,\mu) = -k_B T \ln \frac{N_{obs}(\gamma,\mu)}{C \phi_\gamma \phi_\mu}, \tag{18}$$

with ϕ_γ defined in Equation 2b.

Consider now the excess energy defined in Equation 3b. For both scales, it is straightforward to show that

$$\epsilon_{\gamma\mu,excess} = -k_B T \ln\left(\frac{N_{obs}(\gamma,\mu)}{\sqrt{N_{obs}(\gamma,\gamma)N_{obs}(\mu,\mu)}}\right). \tag{19}$$

That is, the excess energy is independent of the reference state. Averaging over a set of structures as in Equation 13b or 16b modifies this result only slightly.

Because the excess interaction energy does not depend on the choice of either a contact or mole fraction reference state, the difference in interaction energies between the two scales must reside in the ideal component of the pair potential defined in Equation 3a. Consider then, the difference in the ideal component between the contact and mole fraction reference states:

$$\epsilon_{\gamma\mu,ideal}(\phi) - \epsilon_{\gamma\mu,ideal}(x) = k_B T \ln\left(\frac{\phi_\gamma}{x_\gamma}\right) + k_B T \ln\left(\frac{\phi_\mu}{x_\mu}\right), \tag{20a}$$

which can be rewritten as

$$\epsilon_{\gamma\mu,ideal}(\phi) - \epsilon_{\gamma\mu,ideal}(x) = k_B T \ln\left(\frac{z_\gamma}{\bar{z}}\right) + k_B T \ln\left(\frac{z_\mu}{\bar{z}}\right). \tag{20b}$$

Here, the mean number of contacts per residue averaged over all residues is

$$\bar{z} = \frac{\sum n_\eta z_\eta}{\sum n_\eta}. \tag{20c}$$

Table 4. *Comparison of pair potentials derived with different reference states*

Reference state	Native	Native, filtered	Gaussian	Contact fraction-averaged	GKS	I/E[a]	Ideal[b]	Excess[c]
Native	—					1.74	0.80	0.16
Native, filtered	0.82	—				1.34	0.66	0.33
Gaussian	0.96	0.31	—			1.71	0.78	0.12
Contact fraction-averaged	0.38	0.34	0.43	—		0.73	0.04	0.71
GKS	0.26	0.22	0.31	0.73	—	0.86	0.15	0.77
Native-contact	0.98	0.81	0.97	0.43	0.27	1.71	0.77	0.21

[a] I is the average magnitude of the ideal component of the pair potential to the excess component, calculated for the diagonal and upper triangular elements of the appropriate interaction matrix.
[b] Correlation coefficient between the ideal component of the pair interaction matrix and the full interaction matrix.
[c] Correlation coefficient between the excess component of the pair interaction matrix and the full interaction matrix.

Thus, if for attractive (repulsive) pairs the mean number of contacts for residue A_γ is larger (smaller) than $\bar{z}$, then the magnitude of the ideal component will be smaller for the contact fraction-based scale than for the mole fraction scale. Because the excess component is reference state-independent, the ratio of the average magnitude of the ideal, I, to the excess component, E, will be smaller in the contact fraction-based scale. This observation is certainly true for the aromatic residues and, based on explicit calculation (see Table 4), it holds in general. Thus, the GKS and contact fraction-averaged scales have an I/E of 0.86 and 0.73, respectively. In contrast, all three scales, native, native-filtered, and Gaussian, which are basically equivalent to a quasichemical-mole fraction-based reference state, have an I/E of 1.74, 1.34, and 1.71, respectively.

Other differences exist between scales whose reference state is based on mole fraction and contact fraction, respectively. The spread of interactions is more substantial in all the mole fraction-based scales; under certain circumstances, this might impart enhanced resolution. The standard deviations of the parameters are 0.69 and 0.64 in the scales based on the native and Gaussian reference states, respectively. In the native scale, hydrophobic residues are attractive, like charged residues are mainly repulsive or at worst indifferent, and unlike charge pairs are attractive. In contrast, in the contact fraction-averaged and GKS scales, the standard deviations of the parameters are both 0.32. In the GKS scale, we find that Glu-Glu or Lys-Lys pairs are attractive, and Asp-Glu pairs are indifferent, but Glu-Lys pairs are attractive. In the native scale, Glu-Glu, Lys-Lys, and Asp-Glu pairs are repulsive, but surprisingly, Glu-Lys pairs are inert, and Asp-Lys pairs are repulsive. This is consistent with the work of Lumb and Kim (1995a, 1996) and appears to be in disagreement with the work of Hodges et al. (Lavigne et al., 1996).

Additional analysis of the ideal and excess contributions to the pair energy

The statement that mole fraction-based scales have a strong ideal component means that the interaction between many pairs of residues can be treated as the arithmetic average of an apparent single-residue dependent property. However, not all interactions are of the ideal type, so it is of interest to dissect the scales into their excess and ideal components. In the interest of brevity, we focus on the scale derived using the native reference state (see Table 3B,C), where the scale is decomposed into the ideal energy and excess energy contributions, respectively.

If one focuses on Table 3C, one of the more striking features is that pair interactions between the hydrophobic residues Val, Ile, Leu are ideal, i.e., the excess contribution is essentially zero. In addition, the excess interaction energies between these residues and Met, Phe, Tyr, and Trp are very small. The dominant deviations from ideality are associated with the interactions between hydrophobic and polar residues, which are repulsive, and the strong nonideal component associated with certain polar–polar interactions such as Glu-Lys, whose excess energy is attractive. Interestingly, the excess energy for like charge group interactions is essentially zero. Their repulsive interactions are reflected in the large ideal term. These combine in the full scale to give strong repulsions between like charges; in contrast, interactions between unlike charged groups, such as Glu-Lys or Asp-Lys, are either inert or weakly repulsive. In other words, the pair interaction specificity is not due to specific interactions between hydrophobic residues, but rather is due to pair interactions involving pairs of residues having at least one nonhydrophobic partner. The importance of polar interactions in destabilizing alternative structures has been suggested previously on the basis of a number of experiments (Lumb & Kim, 1995b); these pair scales are consistent with this conjecture. On the basis of Tables 2 and 3, this is not to say that pair interactions between hydrophobic partners are not strong, because they most certainly are; rather, they are not particularly specific on a per pair basis.

Native-contact fraction reference state

The contact fraction reference state derived in Equation 16 calculates the contact fraction from the actual structure of the sequence of interest. As shown below, this restriction in fact gives rise to the difference between the contact and mole fraction-based scales. To explore this point further, we introduce a final reference state, "native-contact," where the contact fraction-based reference state is used, but the expected contact frequency is obtained by threading the sequence of interest through all compact fragments in the structural library.

Suppose that we repeat the derivation as in Equation 16, but now calculate the contact probability as the product of the contact fractions averaged over all structures:

$$P^{\circ}_{contact}(\gamma\mu) = \frac{\sum_{\ell=1}^{S_{tot}} \sum_{m=1}^{S_{tot}} \sum_{k=0;\text{compact}}^{N(m)-N(\ell)} C(k,m,\ell)\,\phi(\gamma,k,m,l)\,\phi(\mu,k,m,l)}{\sum_{\ell=1}^{S_{tot}} \sum_{m=1}^{S_{tot}} \sum_{k=0;\text{compact}}^{N(m)-N(\ell)} C(k,m,\ell)}. \tag{21a}$$

Here, $\phi(k,m,\ell)$ is the contact fraction for the ℓth sequence threaded into the k,mth compact substructure. Thus, the native-contact reference state estimation for the energy is given by

$$\epsilon_{native-contact}(\gamma,\mu) = -k_B T \ln\left(\frac{P_{obs}(\gamma,\mu)}{P^{\circ}_{contact}(\gamma,\mu)}\right). \tag{21b}$$

As shown in Table 4, the correlation coefficient of this scale with the Gaussian reference state scale (defined in Equation 6b) is 0.97, and the scale based on the native reference state (defined in Equation 13c) is 0.98. Basically, as the sequence is threaded through a library of structures, the average number of contacts for a given residue type converges to the mean number of contacts averaged over all residues. Then, as indicated by Equation 20b, the difference between the mole fraction- and this contact fraction-based scale goes to zero. In other words, when the sequence is uncoupled from the structure, then the contact fraction- and mole fraction-based scales converge.

Performance of various scales in threading tests

Gapless threading

In Table 5, we summarize the results of a gapless threading test. In addition to the 87 proteins used as a testing set, we also include the original structures used to derive the statistical potentials. To assess the sensitivity of a given potential, unlike in more standard threading tests, we assign the structure based on its pair energy alone. That is, we do not randomize the sequence in the structure,

Table 5. *Comparison of various pair potentials in gapless threading tests*[a]

Scale	Number of correctly assigned sequences	Average Z-score of correctly assigned sequence
Native	82 of 87	−8.96
Native-filtered	81 of 87	−9.54
Gaussian	83 of 87	−9.71
Contact fraction-averaged	57 of 87	−4.10
GKS scale	24 of 87	−1.48

[a]Eighty-seven test sequences were threaded through a library of 311 structures, including the 224 sequences used to derive the pair potential.

subtract the average energy of the randomized sequence in the structure, and then rank the structures on this basis. There is a clear qualitative difference between the two kinds of scales. Those having a mole fraction-based reference state perform better consistently. Interestingly, on the basis of the average Z score, the Gaussian chain-based scale is the most specific. But in practice, the differences in performance of the three mole fraction-based scales in gapless threading are not significant. Note that the native-filtered scale performs marginally worse than the native scale, but it has a better Z score. This suggests that there may be a problem with using Z scores alone to assess the utility of a given interaction scheme.

Inverse folding with gaps

Next, we explored the sensitivity of a given potential as assessed by its ability to find a similar structure in a library of sequentially unrelated folds. We have found that the inclusion of an amino acid pair-specific, local secondary structure preference term enhances accuracy (Jaroszewski & Godzik, unpubl.). The weight of the local interaction energy term (W_{opt}) was optimized independently for each parameter set and equals 2.0 for all scales, except for the GKS scale, where $W_{opt} = 1$. As presented in Tables 6 and 7, the scale that detects the largest percentage of correctly predicted folds, 48%, uses the native reference state. Interestingly, the next best performing scale is the GKS scale, with a maximum of 44% structures recognized, which is followed by the Gaussian reference state scale with 40% recognized. The native-filtered scale performs marginally poorer than the other parameter sets. Thus, by including a richer variation in pair interactions some additional misfolded structures are eliminated. This set of results points out that simply using gapless inverse folding (and/or the mean Z score in gapless inverse folding) to assess the ability of a given potential to recognize topological cousins can be very misleading.

As illustrated by Table 6, the optimal values of gap penalties are different for each parameter set. The optimal gap penalties for the Gaussian and native parameter sets are higher than for the GKS parameter set. This is the consequence of the lower absolute values of the GKS interaction parameters.

As summarized in Tables 5, 6, and 7, based on this battery of tests, the native reference state-based scale is the best. It does very well in gapless inverse folding, as well as in detecting structural homology when gaps are permitted. The behavior of the filtered version of this scale is much more uneven. This points out that different scales may work well in one test, but not in another. Thus, a whole spectrum of tests must be employed before the relative utility of a given scale can be assessed. Of course, the final most important test is whether or not a given potential can fold a variety of proteins.

Table 6. *Sensitivity of detection of correct folds*[a]

Native parameter set			
gap\ext	0.3	0.6	1.2
2.0	36	44	32
4.0	36	44	48
8.0	24	28	32
Native-filtered parameter set			
gap\ext	0.1	0.3	0.6
1.0	16	24	24
2.0	32	32	32
4.0	32	36	40
Gaussian parameter set			
gap\ext	0.1	0.3	0.6
1.0	36	32	32
2.0	32	40	40
4.0	32	28	36
GKS parameter set			
gap\ext	0.3	0.6	1.2
2.0	36	36	36
4.0	36	44	36
8.0	28	24	20

[a]The sensitivity of detection by the threading method for various two-body interaction parameter sets, measured by the percentage of correctly predicted folds for 25 sequences, as a function of gap insertion and extension penalty (the sequences were threaded through a database of 380 structures).

Discussion

In this paper, we have presented a number of derivations of the pair potential describing effective interresidue interactions that explicitly account for chain connectivity. First, we considered a Gaussian chain reference state that incorporates the constraint of chain connectivity and nothing else. Next, we generalized the reference state to include the fact that the native state of proteins is compact and has regularities arising from the presence of secondary structural elements, but otherwise lacks specific side-chain interactions. Such a hypothetical native reference state has been constructed by calculating the contact probability between pairs of amino acids by

Table 7. *Comparison of inverse folding results with gaps obtained with various two-body interaction parameter sets*

Parameter set	Best sensitivity of detection in percent of correctly identified structures
Native	48
Native-filtered	40
Gaussian	40
Contact fraction-averaged	36
GKS	44

threading each sequence through a library of appropriately compact fragments of native proteins. The resulting scale has been shown to be the most sensitive of the scales as assessed by its performance in inverse folding both with and without gaps. However, in spite of the fact that the native scale is derived while maintaining chain connectivity and the presence of secondary structure, in fact it reduces to a scale based on the quasichemical-mole fraction-based approximation. The identical conclusion holds for the Gaussian chain reference state. The origin of these results is the factorization of the expected contact probability into sequence-dependent and structural-dependent terms. Because the amino acid sequence distribution is random, the resulting scale then reduces to a quasichemical scale based on mole fraction units. Indeed, even when contact fraction-based scales are used but sequence and structure are uncoupled, again, because of the randomness of protein sequences, the mole fraction-based, quasichemical reference state is recovered. Thus, we conclude that all scales which ignore the repacking of a sequence when it is threaded into a given structure will reduce to a quasichemical-mole fraction-based reference state.

Whether or not the quasichemical approximation is correct thus reduces to the question of the magnitude of the side-chain repacking term when a sequence experiencing just hard core interactions is threaded into a given structure. The importance of this contribution is, at present, unknown, and an estimation of this term is clearly necessary. Because the desired reference state is one having no preferential side-chain interactions, the most straightforward way to estimate this term is to assume a collection of hard-core side chains, rebuild the protein for the sequence of interest in each member of the structural library, and then calculate the modified side-chain contact map. The difference between the original contact map and the modified one would constitute the basis for determining the repacking energy correction to the quasichemical approximation. Such a series of calculations is now in progress.

Because there is renewed interest in interaction scales containing a reduced number of parameters, we also considered a filtered version of the native scale. Although such a scale worked acceptably well in gapless inverse folding, it performed somewhat worse when inverse folding with gaps (and the attendant exponential increase in alternative possible folds) was done. In other words, in the case of threading and perhaps for de novo folding as well, it is the variation in interactions that enhances the specificity, an intuitively reasonable result. When the scale is smoothed, it performs more poorly as the manifold of accessible conformations against which it must discriminate is increased. A fuller investigation of these effects is the topic of future work.

In this paper, pair potentials have been derived assuming a square well interaction. However, the basic prescription for the derivation of the potential is completely general and can be applied to any assumed form of the interaction. For example, if a distance-dependent potential is desired, because the reference state is a library of compact structures, the correct asymptotic behavior at large distances along with appropriate corrections for protein size follows immediately.

Of course, the ultimate test of any potential and protein representation is the ability to recognize the native conformation as being lowest in energy. At least as far as threading is concerned, we can point to improvements over our previously derived, quasichemical approximation-contact fraction-based potentials. More generally, it is unclear how well the new set of potentials will perform in folding proteins from the random coil state (Hao & Scheraga, 1994, 1995; Kolinski et al., 1995, 1996; Kolinski & Skolnick, 1996). Although the size of the existing structural database limits us to consider mostly pair and a few triplet interactions, the approach can be generalized to any set of interacting groups. Whether or not the resulting potentials can, in fact, fold numerous proteins, the range of validity of the quasichemical approximation that has formed the basis of the derivation of statistical potentials now has been established.

Materials and methods

Database preparation

The set of interaction parameters has been derived from a training library comprised of 224 highly refined, nonhomologous proteins extrinsic to the testing set of examined proteins. For gapless inverse folding, a testing set of 87 test proteins that contain at least 51 residues were examined. Both sets of proteins had a resolution better than 2.5 Å, a residual factor better than 20% and a homology threshold of 30%. The list of proteins used for parameter set development and testing is available via anonymous ftp from ftp.scripps.edu.

For each protein of interest, we construct a side-chain contact map using the definition that any pair of side chains are in contact whenever any of their heavy atoms is less than 4.5 Å apart. For parameter derivation, we assume a fixed contact map; that is, the contact map is excised from the crystal structure and is not allowed to readjust when a new sequence is threaded through the structure. Thus, the static contact map approximation is used. This approximation is equivalent to treating all amino acids as being of the same size. For gapless inverse folding, the identity of the partners may change, but the presence of the contact cannot.

Correlation coefficients

The correlation coefficient, r (Langley, 1970), is designed to test the hypothesis that two parameter sets $[x]$ and $[y]$ are linearly related and is defined by

$$r = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\sqrt{(\langle x^2\rangle - \langle x\rangle^2)(\langle y^2\rangle - \langle y\rangle^2)}}. \tag{22}$$

If the two sets are perfectly correlated, $r = 1$; if the two sets are completely anticorrelated, $r = -1$; and if the two sets are uncorrelated, $r = 0$.

Threading with gaps: Sensitivity of detection

The sensitivity of detection, i.e., the ability of a given parameter set to identify sequentially unrelated but topologically identical proteins was calculated for 25 sequences from 21 structural families. Each sequence was compared to topology fingerprints from a database containing 380 proteins. This database has been described elsewhere (Godzik et al., 1995). The sensitivity of detection by the threading method is measured as the percentage of correctly predicted folds for these 25 sequences. Correct prediction means that, for a given sequence, the best scoring protein is a member of the same structural family as the test sequence.

Optimal values of gap penalties and the weight of local interaction energy term

Because corresponding elements of similar proteins can differ in size, the analogous fragments of the sequences can shift. Thus, the

possibility of introduction of gaps in the aligned sequences is necessary (Godzik et al., 1992). By analogy to sequence analysis methods, two gap penalties are applied. The gap introduction penalty is the energetic cost of introducing a gap. The gap extension penalty is the cost of elongating an already existing gap. The best values of the gap penalties can be different for each form of the scoring function. Gap penalties should be optimized for each tested parameter set or new form of the scoring function. The calculation of the sensitivity of detection was determined for gap introduction penalties: 2.0, 4.0, 8.0; and for gap extension penalties: 0.3, 0.6, 1.2, which lie in the range of the optimal alignment of the sequence into the structure.

In the present version of our inverse folding algorithm, a local interaction energy function that reflects the statistical preference of consecutive pairs of amino acids for a given kind of secondary structure (Kolinski et al., 1995) was added. Elsewhere, we have shown that the inclusion of such a term improves the sensitivity of detection. The relative weight of this term influences the quality of the results. The weight of the local interaction energy term was optimized for each parameter set tested. The calculations were repeated for four values of this term (0.5, 1.0, 2.0, 4.0). The value giving optimal accuracy of the alignment was selected.

Definition of Z scores

One way of assessing the significance of the energy, E of a given sequence in a given structure is to express it in terms of the Z-score, defined as

$$Z = (E - \langle E \rangle)/\sigma, \tag{23a}$$

where $\langle E \rangle$ is the average value of the energy and σ is the standard deviation of the average energy,

$$\sigma = (\langle E^2 \rangle - \langle E \rangle^2)^{1/2}. \tag{23b}$$

In general, the lower the Z score is, the more specific the preference of a given sequence for a given structure.

Acknowledgments

This research was supported in part by grant GM-48835 of the Division of General Medical Sciences, the National Institutes of Health. Useful discussions with Dr. Angel Ramirez Ortiz are gratefully acknowledged. We also thank the reviewers for their very helpful comments.

References

Bryant SH, Lawrence CE. 1991. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential. A statistical model for nonbonded interactions. *Proteins Struct Funct Genet 9*:108–119.

Finkelstein AV, Badretdinov AY, Gutin AM. 1995. Why do protein architectures have Boltzmann-like statistics? *Proteins Struct Funct Genet 23*:142–150.

Flory PJ. 1953. *Principles of polymer chemistry*. Ithaca, New York: Cornell University Press. pp 402–413.

Godzik A. 1996. Knowledge-based potentials for protein folding: What can we learn from protein structures? *Structure 4*:363–366.

Godzik A, Kolinski A, Skolnick J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci 4*:2107–2117.

Godzik A, Kolinski A, Skolnick J, Jaroszewski L. 1997. Dominant effects in residue interaction in proteins. Analysis of energy parameter sets. *Proteins*. Submitted.

Godzik A, Skolnick J, Kolinski A. 1992. A topology fingerprint approach to the inverse folding problem. *J Mol Biol 227*:227–238.

Hansmann UHE, Okamoto Y. 1993. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple minima problem. *J Comput Chem 14*:1333–1338.

Hao MH, Scheraga HA. 1994. Statistical thermodynamics of protein folding: Sequence dependence. *J Phys Chem 98*:9882–9893.

Hao MH, Scheraga HA. 1995. Statistical thermodynamics of protein folding: Comparison of mean-field theory with Monte Carlo simulations. *J Chem Phys 102*:1334–1348.

Jernigan RL, Bahar I. 1996. Structure derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Kolinski A, Galazka W, Skolnick J. 1995. Computer design of idealized β-motifs. *J Chem Phys 103*:10286–10297.

Kolinski A, Galazka W, Skolnick J. 1996. On the origin of the cooperativity of protein folding. Implications from model simulations. *Proteins 26*:271–287.

Kolinski A, Godzik A, Skolnick J. 1993. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins. Application to designed helical proteins. *J Chem Phys 98*:7420–7433.

Kolinski A, Skolnick J. 1996. *Lattice models of protein folding, dynamics and thermodynamics*. Austin, Texas: R.G. Landes Company.

Langley R. 1970. *Practical statistics*. New York: Dover.

Lavigne P, Sonnichsen FD, Kay CM, Hodges RS. 1996. Interhelical salt bridges, coiled-coil stability and specificity of dimerization. *Science 271*:1136–1137.

Lumb KJ, Kim PS. 1995a. Measurement of interhelical electrostatic interactions in the GCN4 leucine zipper. *Science 268*:436–439.

Lumb KJ, Kim PS. 1995b. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry 34*:8642–8648.

Lumb KJ, Kim PS. 1996. Response to Lavigne et al. *Science 271*:1137–1138.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 277*:876–888.

Mattice WL, Suter UW. 1994. *Conformational theory of large molecules*. New York: Wiley.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Miyazawa S, Jernigan RL. 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J Mol Biol 256*:623–644.

Olszewski KA, Kolinski A, Skolnick J. 1996. Folding simulations and computer redesign of protein A three helix bundle motifs. *Proteins Struct Funct Genet 25*:286–299.

Park B, Levitt M. 1996. Energy functions that discriminate X-ray and near native folds from well constructed decoys. *J Mol Biol 258*:367–392.

Skolnick J, Kolinski A. 1996. Monte Carlo lattice dynamics and the prediction of protein folds. In: Gunsteren WF, Weiner PK, Wilkinson AJ, eds. *Computer simulations of biomolecular systems*. ESCOM, Amsterdam.

Sun S. 1993. Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci 2*:762–785.

Tanaka S, Scheraga HA. 1976. Medium and long range interaction parameters between amino acids for predicting three dimensional structures of proteins. *Macromolecules 9*:945–950.