RESEARCH ARTICLES

Derivation of Protein-Specific Pair Potentials Based on Weak Sequence Fragment Similarity

Jeffrey Skolnick,[1]* Andrzej Kolinski,[1,2] and Angel Ortiz[1]
[1]*Laboratory of Computational Genomics, Danforth Plant Science Center, St. Louis, Missouri*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **A method is presented for the derivation of knowledge-based pair potentials that corrects for the various compositions of different proteins. The resulting statistical pair potential is more specific than that derived from previous approaches as assessed by gapless threading results. Additionally, a methodology is presented that interpolates between statistical potentials when no homologous examples to the protein of interest are in the structural database used to derive the potential, to a Go-like potential (in which native interactions are favorable and all nonnative interactions are not) when homologous proteins are present. For cases in which no protein exceeds 30% sequence identity, pairs of weakly homologous interacting fragments are employed to enhance the specificity of the potential. In gapless threading, the mean *z* score increases from −10.4 for the best statistical pair potential to −12.8 when the local sequence similarity, fragment-based pair potentials are used. Examination of the ab initio structure prediction of four representative globular proteins consistently reveals a qualitative improvement in the yield of structures in the 4 to 6 Å rmsd from native range when the fragment-based pair potential is used relative to that when the quasichemical pair potential is employed. This suggests that such protein-specific potentials provide a significant advantage relative to generic quasichemical potentials. Proteins 2000;38:3–16.** © 2000 Wiley-Liss, Inc.

Key words: Knowledge-based potentials; tertiary structure prediction; potential derivation; sequence analysis

INTRODUCTION

One of the key problems in the prediction of a protein's tertiary structure from its amino acid sequence is the development of potentials that can recognize the native conformation among the myriad of alternatives.[1–3] Such potentials might be at atomic level of detail[4–6] or might describe interactions in a reduced protein model.[7,8] Whatever the representation, for protein structure prediction, one would like to have a Go-like pair potential in which all native interactions are attractive and all nonnative interactions are repulsive.[9,10] If one could obtain such a potential *without* knowledge of the native structure, this would greatly facilitate the prediction of the native conformation. Obviously, Go-like potentials are protein specific. For example, in a Go-like pair potential, some Leu pairs would be attractive (because they involve native contacts), whereas other Leu pairs would be repulsive because they would involve nonnative contacts. In contrast, in standard knowledge-based pair potentials, all pairs of Leu are attractive.[11–14] A key question is how to derive Go-like potentials without a priori knowledge of the native structure. One way to proceed is suggested by sequence-based approaches that detect evolutionary relationships among proteins.[15–17] These methods accurately predict protein structure provided that the level of sequence identity is sufficiently high, i.e., roughly 35%. If the sequence identity drops into the so-called twilight zone (below 30%), then the identification of homologous proteins becomes much less certain and structure prediction becomes less reliable.[18] This observation suggests it might be profitable to develop a methodology for potential derivation that can span the range from high to nonexistent sequence identity to proteins in a representative structural library. In what follows, we focus on a method that accomplishes this goal and focus specifically on the derivation of pair potentials, although the formalism could be applied to any type of potential.

Potentials can be at the level of atomic detail or can describe interactions at a reduced level of description of the protein. The advantage of atomic resolution models is that their relevant potentials can, in principle, be directly derived from the laws of physics.[4–6] The disadvantage is that such models are computationally very expensive.

Grant sponsor: National Institutes of Health; Grant Number: GM-48835.

*Correspondence to: Jeffrey Skolnick, Computational and Structural Biology, Danforth Plant Science Center, CET, 4041 Forest Park Avenue, St. Louis, MO 63108. E-mail: skolnick@danforthcenter.org or Andrzej Kolinski, Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02–093 Warsaw, Poland.

Received 30 April 1999; Accepted 11 August 1999

Nevertheless, there have been some encouraging developments of late in the ab initio folding of a small protein at full atomic detail.[19–21] When one reduces the level of description to provide for computational tractability, one needs potentials of mean force.[22] To date, no one has succeeded in developing a practical prescription for the calculation of mean force potentials from detailed atomic models; thus, alternative approaches have been developed. These come in two basic flavors. One formulation derives so-called knowledge-based potentials based on observed properties in proteins of known structure.[11,12,14,23–26] Most recently, transformations of such knowledge-based potentials have been proposed with the aim of improving the quality of such potentials.[27] Alternatively, potentials are obtained by searching for a parameter set that discriminates the native state from a collection of decoys.[28–31]

In what follows, because we shall extend knowledge-based approaches to the derivation of pair potentials, a brief overview of the fundamental ideas is appropriate. Consider for definiteness side chain contact-based pair potentials, i.e., we consider square well functional forms. In a representative structural database, let $\rho_{\gamma\delta}$ be the observed fraction of the total number of contacts occurring between amino acids of types γ and δ and let $p^{o}_{\gamma\delta}$ be the expected fraction of γ-δ contacts if there were no preferential interactions between amino acids γ and δ. The potential of mean force between residues γ and δ is then given by

$$\epsilon_{\gamma\delta} = -k_B T \ln(p_{\gamma\delta}/p^{o}_{\gamma\delta}) \quad (1)$$

Here, k_B is Boltzmann's constant, and T is the absolute temperature. Embedded in the calculation of $p^{o}_{\gamma\delta}$ is the choice of reference state, and to a great extent the differences in the various potentials derived to date arise because of differences in this term.[11] When the quasichemical approximation to $\rho^{o}_{\gamma\delta}$ is made,

$$p^{o}_{\gamma\delta} = x_\gamma x_\delta \quad (2)$$

where x_γ is the mole fraction of residues of type γ. Although initially it was believed that equation 2 neglected the effects of chain connectivity, recently it has been shown that this is not the case; the only approximation inherent in Equation (2) (or more precisely, a specific realization of this equation) is the neglect of side chain repacking because of different sizes of amino acids.[14] In fact, more recent work further suggests that the magnitude of this repacking term is also quite small (Skolnick, unpublished data).

In the types of potentials described above, one simply considers a complete structural library with no selection based on knowledge of local sequence similarity. Thus, in the formulation of Equation (1), the goal has been to derive potentials that are generic and can be applied to any protein. So, if, for example, in a very large structural library, one happened to have a protein with 50% identity to the target sequence, the potential would be insensitive to such a relationship, even though it is very likely that the target sequence adopts this structure. Although this may be advantageous if one wants to use knowledge-based potentials to investigate general questions of protein folding and folding mechanisms, this might not be the best strategy if structure prediction is the goal.

Of course, the idea of exploiting sequence similarity for the prediction of protein structure is not new. There are very powerful sequence-based approaches, such as BLAST[17,32] and FASTA,[33] which attempt to detect evolutionary relationships between proteins and thereby provide insights into protein structure and function. This type of approach works very well if one of the proteins has an experimentally determined structure, but it cannot be directly used for structure prediction when none of the homologous proteins has a solved structure. This is not to say that sequence information cannot be exploited in structure prediction. For example, a most promising approach to the prediction of local secondary structure is based on the extraction of secondary structural propensities from fragments chosen on the basis of their local sequence similarity to the probe sequence of interest.[34–37] This idea is motivated in part by the recent work of Salamov and Solovyev,[34] who demonstrated that by using local sequence alignments and multiple sequence information, a secondary structure prediction accuracy of 73.5% is obtained. Similarly, Baker et al.[35,38] report encouraging results for the assembly of tertiary structures using fragments having locally similar sequences to the probe sequence. Although these approaches use a criterion of local sequence similarity to select representative conformations that are then used in folding or for secondary structure prediction, alternatively, such a conformational subset could define a structural library from which knowledge-based potentials are extracted. This approach was followed in a recent article by Kolinski and coworkers.[39] When such local preferences are modulated by a protein model that includes protein-like features, then the accuracy of secondary structure prediction increases from 69% to 72% in a small test set.

Recently, Finkelstein[40] and coworkers have suggested a means of using homologous sequence information to enhance the specificity of pair potentials. The idea behind this approach is that the set of calculated potentials differs from the true potentials by random errors that can be reduced by averaging over homologous protein sequences. They then demonstrated that this conjecture holds in a model system constructed to satisfy this assumption. Recently, Reva et al.[41] showed that this result also holds when it is applied to statistical pair potentials extracted from a database describing interactions between Cα carbons. The basic idea is as follows: in a set of S-aligned sequences, in the *s*th such sequence, let the pair interaction between the ith and jth residues be ϵ_{ij}. Note that i and j refer to a position in the sequence and not residue types. The effective pair potential describing the average interaction between the ijth pair is given by

$$\overline{\epsilon_{ij}} = \sum_{s=1}^{S} \epsilon_{ij}(s) \Big/ S \quad (3)$$

One might imagine that the averaging process works by decreasing the contribution of $\overline{\epsilon_{ij}}$ in those regions of the sequence that are not conserved and enhancing its contribution in those regions that are. Using a recently developed Cα-based pair potential and a test set of 20 proteins, each having from 20 to 70 homologs taken from the HSSP sequence-structure base,[42] they showed that averaging of protein energies over homologs reduces the average z score (the energy in standard deviation units from the mean value) from ~ -6.1 to ~ -8.1. Such an increase in selectivity is important for protein fold recognition.

The organization of this article is as follows. In the first part of the Methods section, we explore various technical improvements to our quasichemical-based pair potential because this will provide the background potential against which all protein sequence–specific potentials will have to compete. Next, we describe a novel means of deriving protein-specific pair potentials (i.e., different potentials for each protein) for proteins whose native conformation is *unknown*. Such a treatment is a step towards deriving Go-like pair potentials. Then, in Results, we present a number of potentials and examine their performance in a set of gapless threading tests. Finally, in Discussion and Conclusions, we examine the significance of this work, its limitations, and the prospects for future developments.

METHODS

Improvements in Quasichemical-Based Pair Potentials

An essential question in the derivation of knowledge-based potentials is the choice of the reference state. In fact, because Equation (1) consists of the ratio of two quantities, the observed and expected random contact probabilities, in practice, there are a number of ways that these quantities can be obtained. Consider a library of L structures. Let $n_{\gamma\delta}(\ell)$ be the observed number of contacts between residue types γ and δ in the ℓth structure, with $N_c(\ell)$ being the total number of contacts in that structure. Then, the simplest way of calculating the pair potential between these two residues is to simply pool all the observed and expected number of contacts. This gives the following pair potential, termed the *composition independent scale*,

$$\epsilon_{\gamma\delta}^{\text{pooled}} = -k_B T \ln\left(\sum_{\ell=1}^{L} n_{\gamma\delta}(\ell) \Big/ \overline{n_{\gamma\delta}^{o}}\right) \tag{4a}$$

where the mean expected number of contacts $\overline{n_{\gamma\delta}^{o}}$ is calculated from

$$\overline{n_{\gamma\delta}^{o}} = N_c \, \overline{x_\gamma \, x_\delta} \tag{4b}$$

The total number of contacts is given by

$$N_c = \sum_{\gamma=1}^{20} \sum_{\delta=1}^{20} \sum_{\ell=1}^{L} n_{\gamma\delta}(\ell) \tag{4c}$$

and the average mole fraction of residue type γ is

$$\overline{x_m} = \sum_{\ell=1}^{L} a_\gamma(\ell) \Big/ \sum_{\gamma=1}^{20} \sum_{\ell=1}^{L} a_\gamma(\ell) \tag{4d}$$

with $a_\gamma(\ell)$ the number of residues of type γ in the ℓth protein.

The second means of calculating the pair potential between residues γand δ is to calculate the expected number of contacts not by counting the entire number in the structural database but rather by counting the expected number of contacts per protein, $n^{o}_{\gamma\delta}$. We term this the *partial composition corrected scale*. This is the approach we followed previously.[14] Here, rather than using Equations (4b) through (4d), we consider

$$\tilde{\epsilon}_{\gamma\delta} = -k_B T \ln\left(\sum_{\ell=1}^{L} n_{\gamma\delta}(\ell) \Big/ \sum_{\ell=1}^{L} N^{o}_{\gamma\delta}(\ell)\right) \tag{5a}$$

where the expected number of contacts of residues γ and γδ in the ℓth structure, $n^{o}_{\gamma\delta}(\ell)$, is

$$n^{o}_{\gamma\delta}(\ell) = N_c(\ell) \, x_\gamma(\ell) \, x_\delta(\ell) \tag{5b}$$

with $N_c(\ell)$ the total number of contacts and $x_\gamma(\ell)$ is the mole fraction of residues of type γ in the ℓth structure.

$$N_c(\ell) = \sum_{\gamma=1}^{20} \sum_{\delta=1}^{20} n_{\gamma\delta}(\ell) \tag{5c}$$

and

$$x_\gamma(\ell) = a_\gamma(\ell) \Big/ \sum_{\delta=1}^{20} a_\delta(\ell). \tag{5d}$$

The problem with Equations (4a) and (5a) is that by pooling residues from different proteins with different compositions, the expected and/or observed number of contacts can be incorrectly accounted for. For example, consider contacts involving relatively rare residues such as Trp. In those proteins, entirely devoid of such contacts, Equation (4a) would still assert that the expected number would be nonzero. Equation (5a) is somewhat better, in that the expected contact number would be zero for the kth such protein. Nevertheless, by pooling the observed number of contacts over all proteins, this observed number is spread out among those proteins that have contacts and those that do not.

This incorrect accounting for composition can be readily fixed, as follows:

$$\langle \epsilon_{\gamma\delta} \rangle = \sum_{\ell=1}^{L} \varphi_{\gamma\delta}(\ell) \, \epsilon_{\gamma\delta}(\ell) \Big/ \sum_{\ell=1}^{L} \varphi_{\gamma\delta}(\ell) \tag{6a}$$

with

$$\varphi_{\gamma\delta}(\ell) = \min\,(a_\gamma(\ell)\, a_\delta(\ell), 1) \tag{6b}$$

$$\epsilon_{\gamma\delta}(\ell) = = -k_B T \ln(n_{\gamma\delta}(\ell)/n^o_{\gamma\delta}(\ell) \tag{6c}$$

with $n^o_{\gamma\delta}(\ell)$ given by Equation (5b). Equation (6a) is simply the pair energy between residues of types γ and δ averaged over all the structures in the database. The averaging is over all structures where a γ–δ contact is found. We term this the *composition corrected, quasichemical scale*.

Derivation of protein specific pair potentials

Consider the sequence of interest, for which we propose building a protein-specific pair potential. Let each residue be located at the center of a window of $2w+1$ residues. Empirically, we have found that $w=6$ gives the best results. In a database of L protein structures, the local sequence similarity score for the ith such residue with respect to the kth fragment in the ℓth structure is given by

$$s_{ik}(\ell) = \sum_{\substack{iv=-w;\\ a_i = A_k((\ell))}}^{W} B\,(a_{i+iv}, A_{k+iv}(\ell)) \tag{7a}$$

where B is the Blosum62 mutation matrix,[43] a_n is the amino acid identity of the nth residue in the sequence of interest, and $A_k(\ell)$ is the identity of the kth amino acid in the ℓth structure. Note that the identity of amino acids at the center of both fragments must be the same. We then calculate the average sequence similarity score $<s_i>$ and standard deviation σ_i of the similarity between the fragment centered about the ith residue and all other fragments in the structural database. Namely,

$$\langle s_i\rangle = \sum_\ell \sum_k s_{ik}\,(\ell) \Big/ M_i(L) \tag{7b}$$

$$\sigma_i = \sqrt{\sum_\ell \sum_k (s_{ik}\,(\ell) - \langle s_i\rangle)^2 \Big/ M_i(L)} \tag{7c}$$

with $M_i(L)$ the number of fragments whose sequence similarity is compared with the fragment centered about the ith residue in the entire database.[44] It is trivially obtained by replacing s_{ik} in the numerator of Equation (7b) by 1 and evaluating the resulting expression.

Next, for residues i and j, we consider all contacting fragments, such that $s_{ik}(\ell) > f\sigma_i$, $s_{jm}(\ell) > f\sigma_m$ and $a_i = A_k(\ell)$; $a_j = A_m(\ell)$. Empirically, f = 3.0 gives the best results. Then, we count the weight of all contacts within ± 5 residues of the contacting central pair as

$$q_{i+\Delta_1,\,j+\Delta_2}\,(k,l,\ell)$$

$$= C_{k+\Delta_1,l+\Delta_2}(\ell)\theta(B(a_{i+\Delta_1}, A_{k+\Delta_1}))\theta(C(a_{j+\Delta_2}, A_{m+\Delta_2}))\omega_{ij,kl} \tag{8a}$$

where $C_{rs}=1(0)$ if residues r and s are (not) in contact, $|\Delta_1|\leq 5$ and $|\Delta_2 \leq 5$

$$\theta(x) = \begin{cases} 1 \text{ if } x > 0 \\ 0 \text{ otherwise} \end{cases} \tag{8b}$$

The idea of choosing all residues whose amino acid substitution matrix element is favorable is that such contacts are more likely to be retained in the native conformation of the sequence of interest. Finally, we wish to give those contacts involving fragments that are more homologous to the sequence of interest greater weight than we give those that lie just at the threshold of sequence identity. Empirically, we have found a functional form for $\omega_{ij,\,kl}$ of the following type to work quite well

$$\omega_{ij,km}(\ell) = 0.2 + (s_{ik}(\ell) - f\sigma_i)(s_{jm}(\ell) - f\sigma_m) \Big/ (s^o_{ii} - f\sigma_i)(s^o_{jj} - f\sigma_m) \tag{8c}$$

where $s_{ii}{}^o$ is the corresponding sequence similarity score for matching the ith fragment with itself. That is,

$$s^o_{ii} = \sum_{kk=i-w}^{i+w} B(a_{kk}, a_{kk}) \tag{8d}$$

The idea behind Equation (8) is that we consider not only the central contacting residue of the fragment of interest but also the contacts of the neighbors. For sets of overlapping homologous fragments, this will generate a coherent side chain contact potential that stretches over interacting regions of the protein. In practice, it is found to substantially improve the specificity of the pair potential.

Using Equation (8a), the total weight of observed contacts for the i–jth pair in the sequence of interest is

$$Q_{ij} = \sum_\ell \sum_k \sum_l q_{ij}\,(k,l,\ell). \tag{9a}$$

In general, in a library of nonhomologous structures (chosen so that the global sequence identity is less than 30%), Q_{ij} is nonzero for only a small fraction of the possible number of contacting pairs. For those pairs, we calculate the pair potential between residues as

$$E_{ij} = -k_B T \ln\left(Q_{ij} \Big/ Q^o_{ij}\right) \quad Q_{ij} > 0. \tag{9b}$$

Here, the expected number of contacts, Q^o_{ij}, in a protein containing N residues is given by

$$Q^o_{ij} = \sum_{i=j}^{N}\sum_{j=1}^{N} Q_{ij} \Big/ N^2 \tag{9c}$$

For those regions of the molecule lacking such contacts (because there are no contacting sequentially homologous fragments in the structural database), we simply employ the best statistical pair potential that we have derived. In practice, this is given by Equation (6a), which, as shown

TABLE IA. Partially Composition-Corrected Pair Scale

	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP
GLY	1.7	1.4	.9	.9	.8	.6	.5	1.1	.6	.8	.6	.7	.7	1.1	.8	.3	.7	.4	.3	.2
ALA	1.4	1.0	.9	.6	−.1	.6	−.3	.8	−.2	1.1	.8	−.1	1.0	1.1	.6	.4	.5	−.2	−.2	−.5
SER	.9	.9	.5	.4	.4	.2	.4	.6	.3	.3	.4	.4	.7	.4	.2	.3	.1	.1	.1	−.2
CYS	.9	.6	.4	−1.7	−.5	.1	−.5	.4	−.5	.5	.5	−.5	.9	.8	.2	.2	−.2	−.8	−.3	−.7
VAL	.8	−.1	.4	−.5	−.8	−.1	−1.0	.2	−.8	.8	.5	−.9	.5	.5	.3	.1	−.1	−.9	−.7	−.9
THR	.6	.6	.2	.1	−.1	.2	−.2	.5	−.1	.1	.2	.0	.5	.2	.1	.0	.0	−.2	−.2	−.3
ILE	.5	−.3	.4	−.5	−1.0	−.2	−1.1	.1	−.8	.6	.4	−1.2	.4	.3	.0	−.2	.0	−1.1	−.8	−1.1
PRO	1.1	.8	.6	.4	.2	.5	.1	.9	.0	.9	.6	.1	.8	.5	.4	.1	.3	−.2	−.5	−.7
MET	.6	−.2	.3	−.5	−.8	−.1	−.8	.0	−1.0	.4	.1	−1.0	.3	.4	−.1	.0	−.3	−1.1	−.8	−1.2
ASP	.8	1.1	.3	.5	.8	.1	.6	.9	.4	.6	.0	.6	−.2	.7	.2	−.6	−.2	.4	−.1	.0
ASN	.6	.8	.4	.5	.5	.2	.4	.6	.1	.0	.1	.3	.3	.3	.0	.0	.0	.0	−.2	−.2
LEU	.7	−.1	.4	−.5	−.9	.0	−1.2	.1	−1.0	.6	.3	−1.1	.3	.4	.0	−.2	.0	−1.1	−.9	−1.1
LYS	.7	1.0	.7	.9	.5	.5	.4	.8	.3	−.2	.3	.3	1.6	−.4	.1	.6	.6	.3	−.2	−.1
GLU	1.1	1.1	.4	.8	.5	.2	.3	.5	.4	.7	.3	.4	−.4	1.0	.2	−.5	−.1	.2	−.2	−.2
GLN	.8	.6	.2	.2	.3	.1	.0	.4	−.1	.2	.0	.0	.1	.2	.0	.0	−.1	−.1	−.3	−.5
ARG	.3	.4	.3	.2	.1	.0	−.2	.1	.0	−.6	.0	−.2	.6	−.5	.0	−.1	−.1	−.4	−.7	−.6
HIS	.7	.5	.1	−.2	−.1	.0	.0	.3	−.3	−.2	.0	.0	.6	−.1	−.1	−.1	−.6	−.3	−.7	−.7
PHE	.4	−.2	.1	−.8	−.9	−.2	−1.1	−.2	−1.1	.4	.0	−1.1	.3	.2	−.1	−.4	−.3	−1.2	−.9	−1.3
TYR	.3	−.2	.1	−.3	−.7	−.2	−.8	−.5	−.8	−.1	−.2	−.9	−.2	−.2	−.3	−.7	−.7	−.9	−.7	−1.1
TRP	.2	−.5	−.2	−.7	−.9	−.3	−1.1	−.7	−1.2	.0	−.2	−1.1	−.1	−.2	−.5	−.6	−.7	−1.3	−1.1	−1.1

TABLE IB. Composition-Corrected Pair Scale

	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP
GLY	1.1	1.2	.8	.1	.7	.6	.4	.8	.2	.7	.4	.6	.6	.9	.5	.2	.3	.3	.2	−.2
ALA	1.2	.8	.8	.0	.0	.5	−.3	.7	−.3	.9	.7	−.1	.8	1.0	.4	.3	.2	−.2	−.3	−.6
SER	.8	.8	.2	−.4	.4	.2	.3	.4	−.1	.2	.2	.4	.5	.3	.1	.2	−.2	.0	.0	−.5
CYS	.1	.0	−.4	−2.4	−.7	−.5	−.8	−.4	−1.3	−.3	−.5	−.6	−.3	−.2	−.7	−.6	−1.4	−1.1	−1.0	−1.5
VAL	.7	.0	.4	−.7	−.7	−.1	−1.0	.2	−.8	.6	.3	−.9	.4	.4	.1	.0	−.3	−.9	−.7	−1.0
THR	.6	.5	.2	−.5	−.1	.0	−.2	.3	−.3	.1	.1	.0	.4	.1	.0	.0	−.2	−.3	−.2	−.6
ILE	.4	−.3	.3	−.8	−1.0	−.2	−1.1	.0	−.9	.5	.2	−1.2	.3	.3	−.1	−.2	−.3	−1.1	−.9	−1.2
PRO	.8	.7	.4	−.4	.2	.3	.0	.3	−.4	.6	.3	.1	.5	.4	.0	.0	−.2	−.3	−.6	−.9
MET	.2	−.3	−.1	−1.3	−.8	−.3	−.9	−.4	−1.4	.0	−.3	−1.0	−.1	.0	−.5	−.4	−.9	−1.2	−1.0	−1.6
ASP	.7	.9	.2	−.3	.6	.1	.5	.6	.0	.2	.0	.5	−.2	.5	.0	−.6	−.4	.2	−.2	−.4
ASN	.4	.7	.2	−.5	.3	.1	.2	.3	−.3	.0	−.2	.3	.2	.2	−.2	−.2	−.4	−.2	−.3	−.6
LEU	.6	−.1	.4	−.6	−.9	.0	−1.2	.1	−1.0	.5	.3	−1.1	.3	.4	.0	−.2	−.2	−1.1	−.9	−1.2
LYS	.6	.8	.5	−.3	.4	.4	.3	.5	−.1	−.2	.2	.3	.6	−.4	.0	.4	.1	.1	−.3	−.5
GLU	.9	1.0	.3	−.2	.4	.1	.3	.4	.0	.5	.2	.4	−.4	.5	.1	−.5	−.3	.1	−.2	−.5
GLN	.5	.4	.1	−.7	.1	.0	−.1	.0	−.5	.0	−.2	.0	.0	.1	−.5	−.2	−.5	−.3	−.4	−.9
ARG	.2	.3	.2	−.6	.0	.0	−.2	.0	−.4	−.6	−.2	−.2	.4	−.5	−.2	−.3	−.4	−.4	−.7	−.9
HIS	.3	.2	−.2	−1.4	−.3	−.2	−.3	−.2	−.9	−.4	−.4	−.2	.1	−.3	−.5	−.4	−1.2	−.6	−.9	−1.2
PHE	.3	−.2	.0	−1.1	−.9	−.3	−1.1	−.3	−1.2	.2	−.2	−1.1	.1	.1	−.3	−.4	−.6	−1.3	−1.0	−1.4
TYR	.2	−.3	.0	−1.0	−.7	−.2	−.9	−.6	−1.0	−.2	−.3	−.9	−.3	−.2	−.4	−.7	−.9	−1.0	−.9	−1.3
TRP	−.2	−.6	−.5	−1.5	−1.0	−.6	−1.2	−.9	−1.6	−.4	−.6	−1.2	−.5	−.5	−.9	−.9	−1.2	−1.4	−1.3	−1.7

below, is found to be the most specific as assessed by gapless threading. That is,

$$E_{ij} = \langle \epsilon_{a_1 a_j} \rangle \qquad \text{when } Q_{ij} = 0. \tag{9d}$$

The net result of such an analysis will be a protein-specific, pseudopair potential, which does not require knowledge of the global protein structure; rather, it attempts to build the protein-specific pair potential on the basis of local sequence similarity of contacting fragments. The advantage of this formulation is that if there were a highly homologous protein in the structural library, essentially a target potential for the corresponding structure would result. This stands in contrast to more standard techniques for the derivation of statistical pair potentials, in which such valuable information would be washed out and essentially ignored. Indeed, in threading without gaps, z scores of −40 to −50 result for the native sequence in its structure if the native structure is (*incorrectly*) included in the structure library used to construct the potential. In the validation phase, to avoid this effect, one has to be very careful to insure that no structures of moderate to weakly homologous sequences are included in the structural library (this is the origin of the 30% sequence identity cutoff described below in the selection of structures). However, suppose that in a genuine structural

TABLE II. Comparison of Various Quasichemical-Based Pair Potentials as a Function of the Number of Structures Used to Derive the Potential

Quasichemical composition-independent scale			Quasichemical partial composition-corrected scale			Quasichemical composition-corrected scale		
Number of structures used to derive potential	Number of structures correctly assigned	Mean z score of correctly assigned structures	Number of structures used to derive potential	Number of structures correctly assigned	Mean z score of correctly assigned structures	Number of structures used to derive potential	Number of structures correctly assigned	Mean z score of correctly assigned structures
25	43/45	−7.50	25	44/45 (43/44)	−7.26 (−7.34)	25	43/45	−8.99
25[a]	43/45	−8.09	25[a]	44/45 (43/44)	−7.99 (−8.07)	25[a]	43/45	−9.36
50	43/45	−8.02	50	44/45 (43/44)	−7.76 (−7.85)	50	43/45	−9.37
496	43/45	−7.92	496	44/45 (43/44)	−7.69 (−7.78)	496	43/45	−9.78

[a]A different set of 25 randomly chosen structures is chosen. The number in parenthesis is the average z score for the set of structures correctly identified by either the composition-independent or the composition-corrected scale.

prediction, such a weak sequence similarity exists between the probe sequence and one of the proteins used to build the fragment library. It may well be detected by this approach so that a Go-like potential is obtained. This is a real advantage that could be profitably exploited by this method. On the other hand, if there are no contacting pairs of fragments of significant sequence similarity, then one simply recovers the statistical pair potential of Equation (6a). In other words, this method interpolates between a standard statistical potential and a predicted Go-like potential. As suggested by the results described below, in practice, one is in this intermediate limit when realistic cases are considered.

Next, for the test proteins, a list of homologous sequences was generated by scanning the EMBL/SWISSPROT database with FASTA[45] and filtering the sequences using MAXHOM.[46] We calculate the effective pair potentials by the method of Finkelstein.[47] If one has a set of S homologous protein sequences, then

$$\overline{E_{ij}} = \sum_{s=1}^{S} Eij(s) \,/\, S \qquad (10)$$

$s = 1$ is the sequence of interest and all sequences $s > 1$ are aligned to the first sequence. All sequences have a sequence identity greater than 30% with respect to the first sequence. For sequence s, if either (or both) of the partners have a gap in their alignment with the first sequence, then $E_{ij}(S) = 0$. This has the effect of weighting positions that are aligned more than those that are not. In principle, this should enhance the specificity of the potential.

Structural Database

A set of 496 structures was randomly selected from the May 1996 PDB select library[48] for the generation of the protein-specific pair potentials, with a further set of 45 test proteins also randomly selected. This list of proteins is available for downloading from our Web site (http://bioinformatics.danforthcenter.org/).

For gapless threading tests, each sequence is threaded through the set of 496 structures used to generate the pair potentials, plus the additional 45 proteins that constitute the testing set. This is done to insure that the protein discriminates not only against proteins in the testing set but also against proteins that make up the training set.

RESULTS

Quasichemical Type of Pair Potentials

Tables IA and IB present the quasichemical partial composition corrected scale (derived on the basis of Equation [5a]) and the composition corrected scale (derived on the basis of Equation [6a]), respectively. These scales, and the composition-independent scale, may be downloaded from our Web site (http://bioinformatics.danforthcenter.org/).

In Table II, we present the results for the various quasichemical-based scales in gapless threading as a function of the number of protein structures used to generate the potentials. A total of 543 structures were used. Each protein is threaded into the set of proteins used to derive the potentials, as well as the other 44 structures of the test set. Both the composition-independent potentials and the scale that is partially corrected for composition perform comparably; they do not become more specific with increasing size of the structural database and give very similar results. This is a somewhat disconcerting result, in that one might have expected the specificity of the potentials to improve with an increasing number of structures. In contrast, the composition-corrected scale becomes more specific with increasing numbers of structures (from an average of −9.18 for 25 structures to −9.78 when 496 structures are used to derive the scale). Furthermore, the mean z score is about two standard deviation units lower when composition is accounted for correctly. In other words, this scale performs substantially better than do alternative approaches.

Table III shows the set of proteins used, the z score of the native state, and the protein structure whose energy is lowest and its z score obtained in gapless threading using

TABLE III. Summary of Gapless Threading Results for Quasichemical-Based Scales†

	Quasichemical composition-independent scale			Quasichemical partially composition-corrected scale			Quasichemical composition-corrected scale		
Protein	Z_{native}[a]	Best[b]	Z_{best}[c]	Z_{native}	Best	Z_{best}	Z_{native}	Best	Z_{best}
1aak_	−8.60	1aak_	−8.60	−8.45	1aak_	−8.45	−10.84	1aak_	−10.84
1amm_	−12.08	1amm_	−12.08	−12.15	1amm_	−12.15	−10.50	1amm_	−10.50
1c2rA	−3.48	1c2rA	−3.48	−3.46	1c2rA	−3.46	−4.66	1c2rA	−4.66
1cewI	−5.31	1cewI	−5.31	−5.15	1cewI	−5.15	−6.39	1cewI	−6.39
1cid_	−9.41	1cid_	−9.41	−9.32	1cid_	−9.32	−12.34	1cid_	−12.34
1cof_	−8.05	1cof_	−8.05	−8.03	1cof_	−8.03	−10.17	1cof_	−10.17
1csn_	−11.30	1csn_	−11.30	−11.20	1csn_	−11.20	−13.51	1csn_	−13.51
1dja_	−7.13	1dja_	−7.13	−7.00	1dja_	−7.00	−9.63	1dja_	−9.63
1dyr_	−12.06	1dyr_	−12.06	−12.01	1dyr_	−12.01	−13.77	1dyr_	−13.77
1eaf_	−7.20	1eaf_	−7.20	−7.12	1eaf_	−7.12	−8.29	1eaf_	−8.29
1eca_	−7.83	1eca_	−7.83	−7.66	1eca_	−7.66	−8.99	1eca_	−8.99
1erw_	−7.61	1erw_	−7.61	−7.48	1erw_	−7.48	−9.84	1erw_	−9.84
1esl_	−8.62	1esl_	−8.62	−8.17	1esl_	−8.17	−11.30	1esl_	−11.30
1fkj_	−5.75	1fkj_	−5.75	−5.73	1fkj_	−5.73	−7.73	1fkj_	−7.73
1gdy_	−6.82	1gdy_	−6.82	−6.78	1gdy_	−6.78	−8.08	1gdy_	−8.08
1gen_	−11.44	1gen_	−11.44	−11.18	1gen_	−11.18	−14.22	1gen_	−14.22
1ghr_	−9.99	1ghr_	−9.99	−9.87	1ghr_	−9.87	−13.04	1ghr_	−13.04
1gky_	*−1.09*	*1psdA*	*−3.50*	*−1.10*	*1psdA*	*−3.49*	*−1.19*	*1psdA*	*−3.96*
1gpr_	−7.07	1gpr_	−7.07	−6.90	1gpr_	−6.90	−9.31	1gpr_	−9.31
1hcp_	−7.31	1hcp_	−7.31	−6.70	1hcp_	−6.70	−7.25	1hcp_	−7.25
1hfh_	−4.70	1hfh_	−4.70	−4.46	1hfh_	−4.46	−7.31	1hfh_	−7.31
1iae_	−8.73	1iae_	−8.73	−8.46	1iae_	−8.46	−9.62	1iae_	−9.62
1icn_	−7.43	1icn_	−7.43	−7.24	1icn_	−7.24	−9.54	1icn_	−9.54
1jcv_	−5.60	1jcv_	−5.60	−5.40	1jcv_	−5.40	−7.84	1jcv_	−7.84
1kaz_	−8.88	1kaz_	−8.88	−8.77	1kaz_	−8.77	−11.15	1kaz_	−11.15
1lbd_	−7.51	1lbd_	−7.51	−7.45	1lbd_	−7.45	−8.28	1lbd_	−8.28
1ltsD	−7.92	1ltsD	−7.92	−7.71	1ltsD	−7.71	−9.85	1ltsD	−9.85
1mls_	−7.55	1mls_	−7.55	−7.54	1mls_	−7.54	−9.75	1mls_	−9.75
1onc_	−7.20	1onc_	−7.20	−7.01	1onc_	−7.01	−8.87	1onc_	−8.87
1pkp_	−4.90	1pkp_	−4.90	−4.77	1pkp_	−4.77	−6.01	1pkp_	−6.01
1pou_	−7.98	1pou_	−7.98	−7.83	1pou_	−7.83	−9.43	1pou_	−9.43
1put_	−5.42	1put_	−5.42	−5.48	1put_	−5.48	−6.97	1put_	−6.97
1rci_	−6.72	1rci_	−6.72	−6.69	1rci_	−6.69	−8.72	1rci_	−8.72
1rgs_	−9.43	1rgs_	−9.43	−9.44	1rgs_	−9.44	−11.07	1rgs_	−11.07
1rie_	−5.66	1rie_	−5.66	−5.46	1rie_	−5.46	−7.98	1rie_	−7.98
1rsy_	−9.28	1rsy_	−9.28	−9.08	1rsy_	−9.08	−10.93	1rsy_	−10.93
1thx_	−8.25	1thx_	−8.25	−8.15	1thx_	−8.15	−10.43	1thx_	−10.43
1tlk_	−5.29	1tlk_	−5.29	−5.18	1tlk_	−5.18	−8.00	1tlk_	−8.00
1vin_	−8.71	1vin_	−8.71	−8.77	1vin_	−8.77	−9.71	1vin_	−9.71
1xel_	−9.89	1xel_	−9.89	−9.82	1xel_	−9.82	−13.19	1xel_	−13.19
2hvm_	−11.91	2hvm_	−11.91	−11.73	2hvm_	−11.73	−15.66	2hvm_	−15.66
2pcy_	−6.47	2pcy_	−6.47	−6.32	2pcy_	−6.32	−9.06	2pcy_	−9.06
2sas_	−9.38	2sas_	−9.38	−9.19	2sas_	−9.19	−10.80	2sas_	−10.80
5p21_	−8.50	5p21_	−8.50	−8.47	5p21_	−8.47	−10.48	5p21_	−10.48
1ptx_	*−4.20*	*linp_*	*−4.22*	−3.66	1ptx_	−3.66	*−5.30*	*1bmfA*	*−6.19*

†Italicized structures are those incorrectly assigned as having the best *z* score.
[a]Z_{native} is the *z* score of the native state.
[b]Best is the name of the lowest-energy structure.
[c]Z_{best} is the *z* score of the lowest-energy structure.

the quasichemical composition-independent, partially composition-corrected, and composition-corrected scales. Failures in gapless threading are italicized. Interestingly, in 43 cases in which the native structure is selected, all scales find the same set of native structures. Only in the case of 1ptx does the partially corrected composition scale identify the native fold, whereas the other two scales do not. Interestingly, its *z* score is worse for the partially corrected composition scale (−3.66) than when the full composition correction is used (−5.30). 1ptx is a scorpion toxin whose native structure contains four disulfide bonds. This is perhaps the reason why this set of scales fails to recognize it as the native state. Furthermore, all scales fail to identify 1gky. This is a guanylate kinase that was solved in complex with guanosine monophosphate. Here, there is no apparent reason why all scales (including protein-specific pair potentials) fail. How-

TABLE IV. Correlation Coefficient Between Various Quasichemical Scales

	Composition-independent scale	Partially corrected scale	Composition-corrected scale
Composition-independent scale	1.00	0.99	0.92
Partially corrected scale	0.99	1.00	0.91
Composition-corrected scale	0.92	0.91	1.00

TABLE V. Comparison of Mean *z* scores of Correctly Identified Structures Using the Different Scales

Scale type	Number correct	Mean *z* score
Quasichemical composition independent	43/45	−7.92
Quasichemical partial composition correction	44/45	−7.69
Quasichemical composition corrected	43/45	−9.78
Sequence similarity averaged quasichemical	43/45	−10.43
Protein-specific pair potential	42/45	−9.92
Sequence similarity averaged protein-specific pair potential	43/45	−12.75

[a]Averaged over the 43 correctly identified structures given in Table III but does not include 1ptx.

ever, these results suggest that if one uses a *z* score of−7 as a conservative cutoff, one can be reasonably certain that the native fold will be found. This conclusion was also found when a subset of 383 proteins of the PDB select set were used and the composition-corrected pair potential was derived for each protein in the set. Here, we used a jackknife procedure to derive the composition-corrected potential. Of these, 370 are correctly selected, but no protein with a *z* score below −7.0 is incorrectly assigned to an alternative structure.

The fact that different reference states detect the same set of native structures reflects the fact that the correlation coefficient between the various scales is high (see Table IV); nevertheless, substantial differences in performance are seen. The question is why. As shown in Table I, the biggest differences involve interactions with aromatic residues. In addition, interactions among the rarer residues (e.g., MET-MET from −1.0 to −1.4) are the most modified. It is this accounting for less typical interactions that the composition-based correction is designed to fix, and, based on the results of Table II, it has at least partially succeeded.

The situation can be further improved by using multiple sequence averaging as in Equation (3). As shown in Table V, which summarizes the results of all scales, the mean *z* score improves from −9.78 to −10.43 when the composition-corrected pair potential is averaged over a set of aligned sequences. Once again, 43 of 45 structures are correctly identified as being native. The two incorrectly identified folds, 1gky and 1ptx, remain. In Table VI, we present the *z* score of the native structure and the lowest energy structure and its *z* score, respectively, for the multiple-sequence averaged potentials. Failures in gapless threading are italicized. Again, if a *z* score threshold of −7 is used, then no protein is incorrectly matched to its native state.

Protein Specific Pair Potentials

Empirically, we have found that for a database in which no two structures have a greater than 30% sequence identity, typically about 5% of the contacts involve fragments selected on the basis of weak local sequence similarity. The remaining pair contacts was described by our best statistical pair potential, namely the composition-corrected quasichemical scale. As shown in Table V, the protein-specific scale now recognizes only 42/45 structures with a mean *z* score of −9.92. This is marginally better than the composition-corrected quasichemical scale itself. The one additional protein missed by this class of potentials, 1hcp, has a marginal *z* score of −5.5.

The real utility of this method of pair potential derivation only becomes apparent from Table V when we consider the sequence similarity averaged protein-specific pair potential. With the addition of multiple sequence averaging, 1hcp is once again recognized with a *z* score of −7.92. All proteins that are not recognized (1ptx and 1gky) have *z* scores above −7.0. Over the testing set, the mean *z* score of correctly identified folds is −12.75. This is an improvement of 2.2 standard deviation units from the best sequence similarity averaged quasichemical potential. In addition, with respect to a single sequence, the improvement in mean *z* score is 2.83σ. In contrast, the composition-corrected quasichemical scale improves by 0.95σ. Similarly, the pair potential derived by Reva and coworkers[41] improves by about 2σ . However, there the mean *z* score is about −8.1, whereas it is −12.75 in the present work.

What multiple sequence averaging does is enforce the contribution of those regions in the structures that have a very consistent pattern of side chain contacts across the protein family. Those regions that do not consistently select a pair of interacting fragments are replaced by a weighted average of quasichemical scale and the local fragment scale. If a pair of putative interacting regions is spuriously chosen for a few sequences, but not for all members of the protein family, then the quasichemical scale will essentially describe the interaction between the selected pairs. Regions that contain gaps are treated as being neutral (pair potential of zero). This helps to eliminate highly variable regions that should not dictate the fold.

Examination of Table VI reveals that a number of proteins have a native state *z* score less than −15. The presence of such a protein-specific pair potential should facilitate folding as well as refinement of low-resolution models. Indeed, about a 1-Å improvement was found in a series of refinements of threading models when protein-specific pair potentials were used compared with results using the composition-corrected quasichemical scales.[49] One might also imagine that such potentials should also be useful in threading where gaps are allowed. However, use of a residue-specific pair potential precludes dynamic programming because the scoring function is inherently nonlocal.

Finally, for proteins up to 400 residues in length, we note that protein-specific pair potentials can be obtained from our Web site (http://bioinformatics.danforthcenter.org/).

TABLE VI. Compilation of Gapless Threading Results for Sequence Similarity Averaged Pair Potentials and Protein-Specific Pair Potentials†

	Quasichemical composition-corrected scale sequence similarity averaged			Protein-specific pair potential			Sequence similarity averaged protein-specific pair potential		
Protein	Z_{native}[a]	Best[b]	Z_{best}[c]	Z_{native}	Best	Z_{best}	Z_{native}	Best	Z_{best}
1aak_	−13.04	1aak_	−13.04	−8.77	1aak_	−8.77	−12.85	1aak_	−12.85
1amm_	−12.32	1amm_	−12.32	−9.24	1amm_	−9.24	−13.59	1amm_	−13.59
1c2rA	−4.66	1c2rA	−4.66	−13.24	1c2rA	−13.24	−13.24	1c2rA	−13.24
1cewI	−7.38	1cewI	−7.38	−5.73	1cewI	−5.73	−7.89	1cewI	−7.89
1cid_	−12.29	1cid_	−12.29	−9.92	1cid_	−9.92	−13.16	1cid_	−13.16
1cof_	−12.07	1cof_	−12.07	−8.18	1cof_	−8.18	−12.08	1cof_	−12.08
1csn_	−13.34	1csn_	−13.34	−14.47	1csn_	−14.47	−16.20	1csn_	−16.20
1dja_	−8.25	1dja_	−8.25	−11.97	1dja_	−11.97	−10.31	1dja_	−10.31
1dyr_	−15.30	1dyr_	−15.30	−13.07	1dyr_	−13.07	−20.76	1dyr_	−20.76
1eaf_	−8.83	1eaf_	−8.83	−8.70	1eaf_	−8.70	−11.30	1eaf_	−11.30
1eca_	−9.30	1eca_	−9.30	−8.40	1eca_	−8.40	−10.40	1eca_	−10.40
1erw_	−11.43	1erw_	−11.43	−9.45	1erw_	−9.45	−17.13	1erw_	−17.13
1es1_	−12.05	1es1_	−12.05	−11.97	1esl_	−11.97	−15.81	1esl_	−15.81
1fkj_	−9.13	1fkj_	−9.13	−7.88	1fkj_	−7.88	−11.77	1fkj_	−11.77
1gdy_	−8.53	1gdy_	−8.53	−6.52	1gdy_	−6.52	−8.12	1gdy_	−8.12
1gen_	−15.68	1gen_	−15.68	−15.30	1gen_	−15.30	−16.70	1gen_	−16.70
1ghr_	−15.81	1ghr_	−15.81	−12.04	1ghr_	−12.04	−17.03	1ghr_	−17.03
1gky_	*−2.69*	*3pga1*	*−4.20*	*−2.00*	*2olbA*	*−5.13*	*−3.10*	*3pga1*	*−5.16*
1gpr_	−8.61	1gpr_	−8.61	−7.75	1gpr_	−7.75	−12.95	1gpr_	−12.95
1hcp_	−8.88	1hcp_	−8.88	*−5.45*	*1pkp_*	*−6.38*	−7.92	1hcp_	−7.92
1hfh_	−8.01	1hfh_	−8.01	*−5.50*	*2fcr_*	*−5.60*	−7.70	1hfh_	−7.70
1iae_	−9.95	1iae_	−9.95	−7.06	1iae_	−7.06	−9.86	1iae_	−9.86
1icn_	−9.61	1icn_	−9.61	−12.58	1icn_	−12.58	−17.58	1icn_	−17.58
1jcv_	−8.26	1jcv_	−8.26	−9.86	1jcv_	−9.86	−10.56	1jcv_	−10.56
1kaz_	−11.70	1kaz_	−11.70	−10.61	1kaz_	−10.61	−14.81	1kaz_	−14.81
1lbd_	−8.80	1lbd_	−8.80	−7.03	1lbd_	−7.03	−8.50	1lbd_	−8.50
1ltsD	−9.85	1ltsD	−9.85	−8.37	1ltsD	−8.37	−8.37	1ltsD	−8.37
1mls_	−9.91	1mls_	−9.91	−10.62	1mls_	−10.62	−12.71	1mls_	−12.71
1onc_	−9.37	1onc_	−9.37	−8.12	1onc_	−8.12	−10.96	1onc_	−10.96
1pkp_	−7.02	1pkp_	−7.02	−7.62	1pkp_	−7.62	−10.12	1pkp_	−10.12
1pou_	−9.53	1pou_	−9.53	−8.44	1pou_	−8.44	−9.16	1pou_	−9.16
1put_	−7.62	1put_	−7.62	−7.78	1put_	−7.78	−8.62	1put_	−8.62
1rci_	−8.85	1rci_	−8.85	−7.32	1rci_	−7.32	−8.95	1rci_	−8.95
1rgs_	−12.37	1rgs_	−12.37	−12.15	1rgs_	−12.15	−15.00	1rgs_	−15.00
1rie_	−8.31	1rie_	−8.31	−6.53	1rie_	−6.53	−9.35	1rie_	−9.35
1rsy_	−14.45	1rsy_	−14.45	−12.10	1rsy_	−12.10	−17.34	1rsy_	−17.34
1thx_	−8.75	1thx_	−8.75	−10.18	1thx_	−10.18	−15.07	1thx_	−15.07
1tlk_	−8.11	1tlk_	−8.11	−10.93	1tlk_	−10.93	−13.07	1tlk_	−13.07
1vin_	−9.30	1vin_	−9.30	−7.27	1vin_	−7.27	−9.13	1vin_	−9.13
1xel_	−14.80	1xel_	−14.80	−12.20	1xel_	−12.20	−16.51	1xel_	−16.51
2hvm_	−15.37	2hvm_	−15.37	−16.09	2hvm_	−16.09	−22.92	2hvm_	−22.92
2pcy_	−8.87	2pcy_	−8.87	−14.54	2pcy_	−14.54	−16.93	2pcy_	−16.93
2sas_	−10.85	2sas_	−10.85	−8.59	2sas_	−8.59	−8.65	2sas_	−8.65
5p21_	−12.04	5p21_	−12.04	−11.89	5p21_	−11.89	−17.34	5p21_	−17.34
1ptx_	*−6.04*	*1iow_*	*−6.38*	−6.34	1ptx_	−6.34	*−6.18*	*1mla_*	*−6.30*

†Italicized structures are those incorrectly assigned as having the best *z* score.
[a]Z_{native} is the *z* score of the native state.
[b]Best is the name of the lowest-energy structure.
[c]Z_{best} is the *z* score of the lowest-energy structure.

Application to Ab Initio Folding

In reality, the actual quality of a given potential is determined by its selectivity and specificity. The gapless threading results presented here indicate that the two classes of potentials have quite high selectivity. Both recognize most of the proper folds for the set of the test proteins. However, the fact that a number of additional native structures are recognized by the homology-based potential indicates its higher selectivity. Furthermore, the significantly higher *z* scores for the recognized structures in the case of the homology-based potentials indicate that its specificity is also significantly greater. However, based

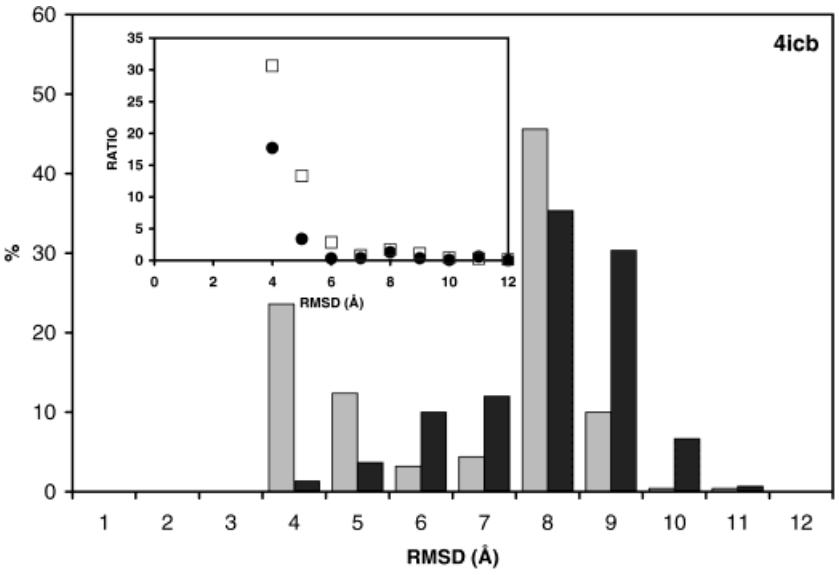

Fig. 1. For 4icb, histograms of the percentage of the simulations that led to structures with a given rmsd from the native structure (for Cα atoms) for both the quasichemical (dark gray) and protein-specific potentials (light gray). **Inset:** for structures obtained at the end of the simulated annealing trajectory, the black circles indicate the ratio of good structures of various qualities obtained from the homology-based potential to those obtained from the purely statistical potential. The open squares correspond to the data from the entire annealing trajectory.

on gapless threading alone, it is unclear how the enhanced specificity and selectivity found in this relatively straightforward test apply to the most rigorous of situations, the relative performance of such potentials in ab initio folding simulations. To address this essential question, we have performed a large number of ab initio Monte Carlo folding experiments on a few small proteins of different structural classes: 1gb1 which is an α/β protein, 1cis, which is a mainly β-protein, 1ctf which is an α+β protein, and 4icb which is an α-helical protein.

Reduced Model and Monte Carlo Simulations

For each sequence, simulations on a reduced protein representation were undertaken for two types of pairwise long-range potentials, given by Equations (6) and (9), respectively, with exactly the same other contributions to the model force field. Because this high-coordination lattice model has been previously discussed,[50] here we briefly outline its essential features. The model chain is built from virtual bonds connecting centers of consecutive interaction units, each comprising the average position of a group of atoms that include the side chain and the corresponding α-carbon atom. The model force field, besides the tested pairwise potentials, contains potentials representing short-range interactions, main chain hydrogen bonds, and two types of hydrophobic burial potentials. In contrast to the previously published application of this model, here the folding simulations are purely ab initio and use a very conservative secondary structure prediction scheme. For very strongly predicted helix or extended regions, (5 and higher scores in both PHD[51] and PSIpred methods [http://globin.bio.warwick.ac.uk/psipred/]), statistical potentials describing short-range interactions have been derived from a structural database consisting of all-β and all-α proteins, respectively.

About 600 simulated annealing Monte Carlo simulations have been performed for each protein-potential pair. The purpose of these simulations was to compare the specificity of the two potentials and their ability to discriminate between native-like conformations and a manifold of random coil states in ab initio simulations. Also, long isothermal simulations near the folding temperature were performed from which the relation between rmsd from native and conformational energy could be extracted and analyzed.

Results of Ab Initio Folding Simulations

In Figures 1 through 4 we illustrate the results of simulated annealing Monte Carlo experiments. The histograms show the percentage of the simulations that led to structures with a given rmsd from the native structure (for the Cα atoms) for both potentials of interest from the end of the simulated annealing trajectories. From the point of view of ab initio folding, helical proteins are the easiest. Because 4icb is rather irregular helical protein, it is not a trivial fold. In gapless threading, the mean z scores are −8.0 and −11.5, respectively, for the quasichemical and protein-specific pair potentials. For ab initio folding, in nearly 40% of the cases, simulated annealing led to the correct fold (with an rmsd from native of 4 to 5 Å) when the homology-based potential was employed. The large peak around 8 Å corresponds to various misfolded structures, with a large contribution from topological mirror image folds. For this molecule, the purely statistical potential performed significantly worse. Only about 5% of simulations lead to similar quality native-like structures. The inset in Figure 1 shows the ratio of good structures of various qualities obtained from the homology-based potential to those of the purely statistical potential. The black circles give another representation of the data presented in the histogram, and the open squares correspond to the data from entire annealing trajectory. Comparison of the two shows that the homology-based potential performs better at all temperatures

The more complex topology of protein G (1gb1) is perhaps a more typical case. We note that in gapless thread-

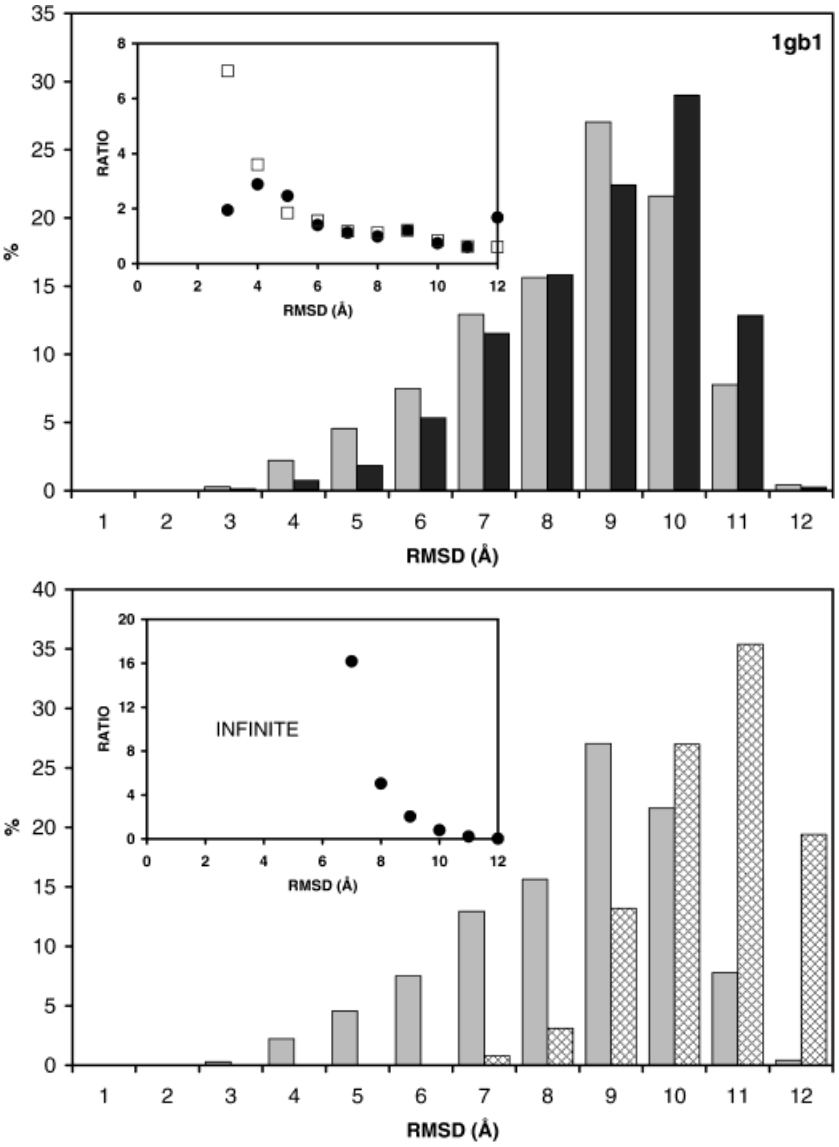

Fig. 2. **A:** For 1gb1, histograms of the percentage of the simulations that led to structures with a given rmsd from the native structure (for Cα atoms) for both the quasichemical and the protein-specific potentials. **Inset:** for structures obtained at the end of the simulated annealing trajectory, the black circles indicate the ratio of good structures of various qualities obtained from the homology-based potential (light gray) to those obtained from the purely statistical potential (dark gray), whereas the open squares correspond to the data from the entire annealing trajectory. **B:** Comparison of the results for the homology-based potential (gray) with simulations of hypothetical compact random structures of 1gb1 chains (hatched). The compactness was enforced by a strong, residue-dependent, centrosymmetric potential.

ing, the z score of the native fold is -6.1 and -6.5 for the quasichemical and protein-specific potentials, respectively. Again, as shown in Figure 2A for ab initio folding, the homology-based potential performs qualitatively better. In about 7% of the runs, the correct fold could be recovered. Is this result significant for protein structure prediction? In Figure 2B, we compared the results for the homology-based potential with simulations of hypothetical compact random structures of 1gb1 chains. The compactness was enforced by a strong, residue-dependent, centrosymmetric potential. Other interactions were neglected. As a result, the obtained structures had a density (and volume) very close to the native structure, with the fraction of buried hydrophobic residues very similar to that observed in the same size native globular protein. In 1000 simulations, no structure below 7 Å rmsd from native was observed. Thus, the results of our folding simulations with more realistic protein-like potentials are very far from random. The potential nicely discriminates not only against other protein folds but also against the enormous number of random conformations. It has to be mentioned that the correlation between rmsd and the system's conformational energy for both potentials is not very strong. Whether this reflects inadequate conformational sampling or defects in the potential and protein model demands further investigation.

Overall, the homology-based potential performs on average much better in ab initio folding than does the simple statistical pair potential. A qualitative difference could be seen for 1gb1, 4icb, and to a lesser extent for the 1cis case, shown in Figure 3, in which the mean z score of the native fold increases from -7.6 to -9.1 in gapless threading. For 1ctf, the homology-based potential seems to favor a topological mirror image of the native structure and generates a

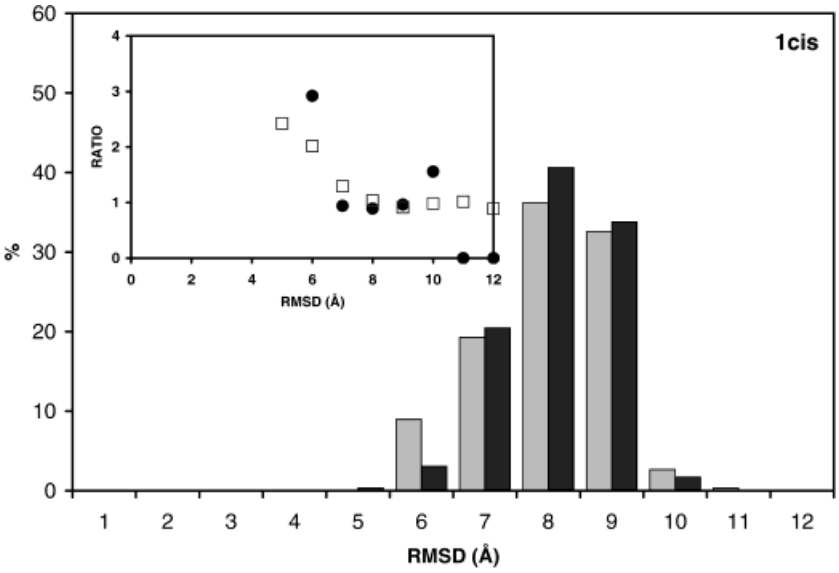

Fig. 3. For 1cis, histograms of the percentage of the simulations that led to structures with a given rmsd from the native structure (for Cα atoms) for both the quasichemical (dark gray) and protein-specific (light gray) potentials. **Inset:** for structures obtained at the end of the simulated annealing trajectory, the black circles indicate the ratio of good structures of various qualities obtained from the homology-based potential to those obtained from the purely statistical potential, whereas the open squares correspond to the data from the entire annealing trajectory.

somewhat smaller number of the correct folds. The difference between the two potentials for this case is rather insignificant when assessed on the basis of rmsd alone (see Fig. 4A), but if the distance rmsd from native is employed (see Fig. 4B), then the same trend is very strongly observed. We further note that this result is consistent with the trend observed in gapless threading in which the z score of the native state increases from -4.7 to -10.1 when the protein-specific potential is employed.

Overall, these results suggest a tendency to generate better structures at higher yield with more negative z scores on gapless threading. Thus, gapless threading can be used as a crude and fast indicator of the quality of a potential in ab initio folding.

Finally, clustering of the results of a large number of simulations enables a quite dependable identification of the proper native-like structures. Proper averaging within the "native-like" structures makes possible further improvement in the accuracy of predicted structures. Parenthetically, note that when the potentials presented here are updated to account for various modes of side chain contacts (three classes: parallel, antiparallel, and acute packing), the efficiency of ab initio folding could be significantly improved, but the trend of quasichemical versus protein-specific potentials remains the same. This will be discussed in detail elsewhere.

DISCUSSION AND CONCLUSIONS

In the present work, we have presented a methodology for the derivation of statistical pair potentials, which accounts for protein composition effects. By correctly accounting for composition, we can improve the sequence-native structure specificity as assessed by gapless threading. Although such a test by no means insures that the protein will be foldable using such a potential, empirically, we have found a good qualitative correlation between the gapless threading z score and improved foldability and quality of the resulting models. Compared with potentials that do not account for composition, the mean z score is about 2σ better, and the potential is found to systematically improve when more proteins are added to the structural database. Mainly, this results from the correct weighting of the expected contribution of rare contacts. Such potentials then provide an enhanced baseline for additional improvements.

One such improvement involved the generalization of these pair potentials to be protein specific. The formalism we developed nicely interpolates from a pure quasichemical statistical potential when there are no contacting pairs of fragments that have significant sequence similarity to the sequence of interest in the structural database to a Go-like potential when an example of a homologous protein is included in the structural database. Clearly, it is advantageous to have the largest structural database possible. In the case considered here, we examined what happens when no protein sequence in the structural database has more than 30% sequence identity. This regimen is the worst-case scenario. The resulting protein-specific pair potential, when multiple sequence averaging is included, performs substantially better than any of our previously derived pair potentials. This conclusion is based on the average gapless threading z scores, the behavior of the potential in protein structure refinement,[49] and, most importantly, on our ab initio folding results. For gapless threading, the observed spread of z scores reflects the anecdotal nature of finding interacting pairs of locally homologous fragments. For ab initio folding, a significantly greater yield of low rmsd structures is generated when protein-specific pair potentials are employed. This alone clearly justifies their use.

In a series of articles on the derivation and testing of pair potentials, we have demonstrated the following: First, the quasichemical approximation to statistical pair potentials in fact accounts for chain connectivity, but it does not account for side chain repacking when one sequence is mounted in the structure of another.[14] Second, for the

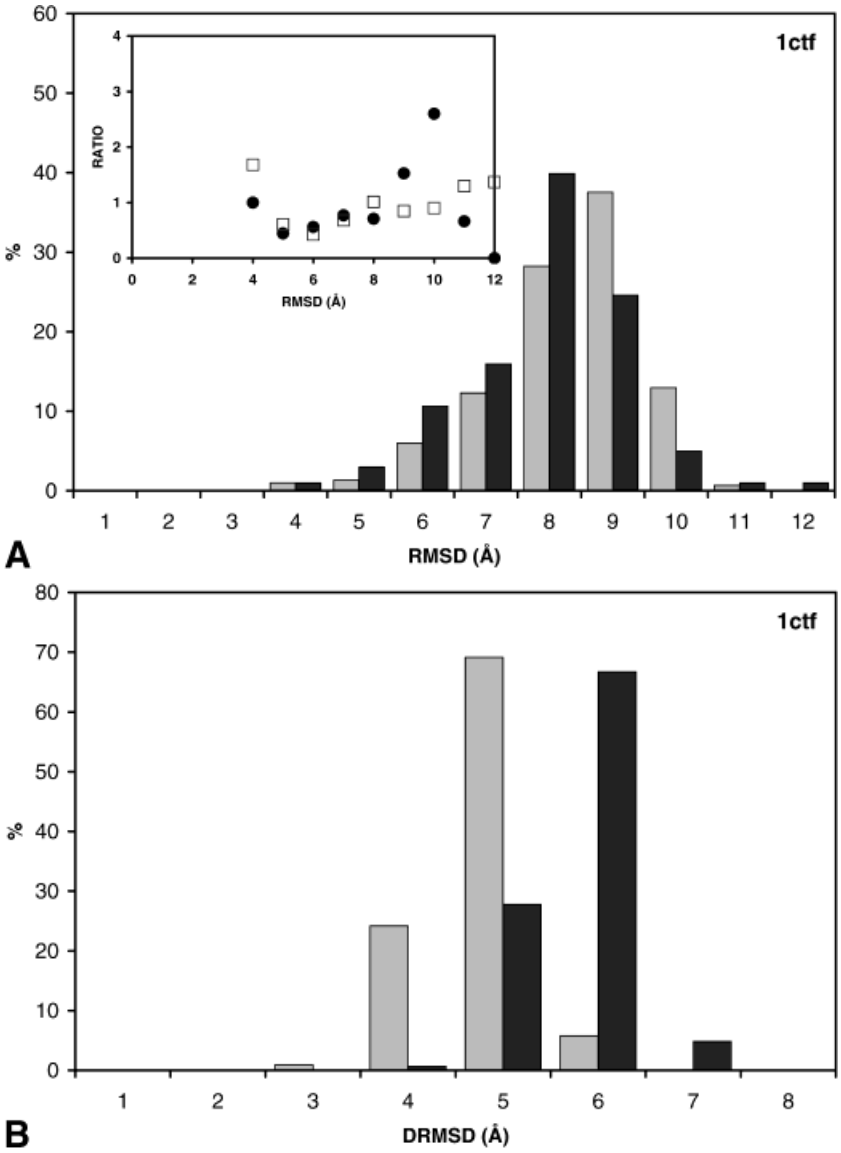

Fig. 4. **A:** For 1ctf, histograms of the percentage of the simulations that led to structures with a given coordinate rmsd from the native structure (for Cα atoms) for both the quasichemical (dark gray) and protein-specific potentials (light gray). **Inset:** for structures obtained at the end of the simulated annealing trajectory, the black circles indicate the ratio of good structures of various qualities obtained from the homology-based potential to those obtained from the purely statistical potential, whereas the open squares correspond to the data from the entire annealing trajectory. **B:** For 1ctf, histograms of the percentage of the simulations that led to structures with a given distance rmsd from the native structure (for Cα atoms) for both the quasichemical and protein-specific potentials.

range of parameters characteristic of real proteins, such knowledge-based scales can recover the "true" potential in model systems.[52] Third, such knowledge-based potentials are highly correlated with CHARMM plus GB/SA type potentials in unfolding simulations of the GCN4 leucine zipper.[53] Thus, they have a physical basis. Fourth, proper correction of composition effects can lead to enhanced specificity. Fifth, using contacting fragments that exhibit significant local sequence similarity to the protein of interest, it is possible to derive a protein-specific potential that exceeds, in some cases by a considerable margin, the specificity of quasichemical potentials.

Future work will involve investigation of the repacking contribution to effective pair potentials. Preliminary work suggests that this effect is quite small. Next, we are generalizing these pair potentials to include orientation effects; as Bahar and Jernigan have pointed out, these may be significant.[54] Finally, these protein-specific pair potentials can be improved by including information about known or predicted secondary structure and/or tertiary contacts such as disulfide bonds. Although more work in the field of potential derivation is clearly warranted, these results suggest that progress is being made.

ACKNOWLEDGMENTS

Stimulating discussions with Drs. A. Finkelstein and B. Reva are gratefully acknowledged. The assistance of P. Rotkiewicz in the preparation of the figures is most appreciated. Andrzej Kolinski is an International Scholar of the Howard Hughes Medical Institute.

REFERENCES

1. Skolnick J, Kolinski A, Ortiz AR. Reduced protein models and their application to the protein folding problem. J. Biomol Struct Dyn 1998;16:381–396.
2. Monge A, Lathrop EJP, Gunn JR, et al. Computer modeling of

protein folding: conformational and energetic analysis of reduced and detailed protein models. J Mol Biol 1995;247:995–1012.
3. Liwo A, Oldziej S, Kazmieckiewicz R, et al. Design of a knowledge-based force field for off-lattice simulations of protein structure. Acta Biochim Pol 1997;44:527–547.
4. Brooks BR, Brucculeri R, Olafson B, et al. CHARMM: a program for macromolecular energy minimization, and molecular dynamics. J Comp Chem 1983;4:187–217.
5. Roterman IK, et al. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. Phi-psi maps for *N*-acetyl alanine *N′*-methyl amide: comparisons, contrasts and simple experimental tests. J Biomol Struct Dyn 1989;7:421–453.
6. Pearlman DA, Case DA, Caldwell JC. *AMBER*. San Francisco: University of California, 1991.
7. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
8. Kolinski A, Skolnick J. Lattice models of protein folding, dynamics and thermodynamics. Austin, TX: R.G. Landes, 1996. p 200.
9. Go N, Abe H, Mnuno H, et al. Protein folding. N. Jaenicke, editor. Amsterdam: Elsevier/North Holland, 1980. pp 167–181.
10. Go N, Taketomi H. Respective roles of short- and long-range interactions in protein folding. Proc Natl Acad Sci USA 1978;75: 559–563.
11. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
12. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
13. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
14. Skolnick J, et al. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
15. Karlin S. Statistical significance of sequence patterns in proteins. Curr Opin Struct Biol 1995;5:360–371.
16. Henikoff S, Greene EA, Pietrovski S, et al. Gene families: the taxonomy of protein paralogs and chimeras. Science 1997;278:609–614.
17. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. Trends Biochem Sci 1998;23:444–447.
18. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
19. Duan Y, Wang L, Kollman PA. The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. Proc Natl Acad Sci USA 1998;95: 9897–9902.
20. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282:740–744.
21. Daura X, Jaun B, Seebach D, et al. Reversible peptide folding in solution by molecular dynamics simulation. J Mol Biol 1998;280: 925–932.
22. Barker JA, Henderson D. What is liquid? Understanding the states of matter. Rev Mod Phys 1976;48:587–671.
23. Tanaka S, Scheraga HA. Medium- and long-range interactions parameters between amino acids for prediction of three-dimensional structures of proteins. Macromolecules 1975;9:945–950.
24. Jernigan RL. Protein folds. Curr Opin Struct Biol 1992;2:248–256.
25. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. Proteins 1992;13:258–271.
26. Levitt M, Gerstein M, Huang E, et al. Protein folding: the endgame. Annu Rev Biochem 1997;66:549–579.
27. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–369.
28. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;277:876–888.
29. Maiorov VN, Crippen GM. Learning about protein folding via potential functions. Proteins 1994;20:167–173.
30. Hao MH, Scheraga HA. How optimization of potential functions affects protein folding. Proc Natl Acad Sci USA 1996;93:4984–4989.
31. Hao MH, Scheraga HA. Optimizing potential functions for protein folding. J Phys Chem 1996;100:14540–14548.
32. Altschul SF, Madden TF, Schaffer AP, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
33. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 1994;24:307–331.
34. Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. J Mol Biol 1997;268:31–36.
35. Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. J Mol Biol 1998;281:565–577.
36. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci USA 1998;95:11158–11162.
37. Shortle D. The state of the art. Curr Biol 1999;9:R205–R209.
38. Simons KT, Kooperberg C, Huang E, et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
39. Kolinski A, Kolinsk A, Jaraszewski L, Rotkiewicz P, et al. An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side groups centers of mass. J Phys Chem 1998;102:4628–4637.
40. Finkelstein AV. 3D protein folds: homologs against errors—an estimate based on the random energy model. Phys Rev Lett 1998;80:4823–4825.
41. Reva B, Skolnick J, Finkelstein A. Averaging interaction energies over homologs improves fold recognition in gapless threading. Proteins 1999;35:353–399.
42. Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 1998;26:313–315.
43. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49–61.
44. McLachlan AD. Test for comparing related amino acid sequences. J Mol Biol 1971;61:409–424.
45. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
46. Sander C, Schneider R. Database of homology derived protein structures and the structural meaning of sequence alignment. Proteins 1991; 9:56–68.
47. Badretdinov A, Finkelstein AV. How homologs can help to predict protein folds even though they cannot be predicted for individual sequences. J Comput Biol 1998;5:369–376.
48. Hobohm U, Sander C. Selection of a representative set of structures from the Brookhaven Protein Databank. Protein Sci 1992;1: 409–417.
49. Kolinski A, Rotkiewicz P, Ilkowski I, et al. A method for the improvement of threading based protein models. Proteins 1999, in press.
50. Kolinski A, Skolnick J. Assembly of protein structure from sparse experimental data. An efficient Monte Carlo Model. Proteins 1998;32:475–494.
51. Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
52. Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.
53. Mohanty D, et al. Correlation between knowledge based and detailed atomic potentials: application to the unfolding of the GCN4 leucine zipper. Proteins 1999;35:447–452.
54. Bahar I, Jernigan RL. Coordination geometry of nonbonded residues in globular proteins. Fold Des 1996;1:357–370.