Protein Science (1998), 7:112–122. Cambridge University Press. Printed in the USA.

How do potentials derived from structural databases relate to "true" potentials?

LI ZHANG AND JEFFREY SKOLNICK

Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

(RECEIVED May 13, 1997; ACCEPTED September 3, 1997)

Abstract

Knowledge-based potentials are used widely in protein folding and inverse folding algorithms. Two kinds of derivation methods are used. (1) The interactions in a database of known protein structures are assumed to obey a Boltzmann distribution. (2) The stability of the native folds relative to a manifold of misfolded structures is optimized. Here, a set of previously derived contact and secondary structure propensity potentials, taken as the "true" potentials, are employed to construct an artificial protein structural database from protein fragments. Then, new sets of potentials are derived to see how they are related to the true potentials. Using the Boltzmann distribution method, when the stability of the structures in the database lies within a certain range, both contact potentials and secondary structure propensities can be derived separately with remarkable accuracy. In general, the optimization method was found to be less accurate due to errors in the "excess energy" contribution. When the excess energy terms are kept as a constraint, the true potentials are recovered exactly.

Keywords: contact potentials; potential of mean force; protein folding; secondary structure propensity

Recently, simplified, coarse-grained models of proteins have attracted much interest because they may offer a practical approach to solving the protein folding problem (Wilson & Doniach, 1989; Sippl, 1995; Finkelstein & Reva, 1996; Kolinski & Skolnick, 1996; Miyazawa & Jernigan, 1996). The advantage of these protein models is their greatly reduced computational costs. However, because many detailed features of real proteins are omitted, there is a great deal of uncertainty as to their range of validity. Many models have been proposed (Levitt, 1976; Ueda et al., 1978; Maiorov & Crippen, 1992; Sippl, 1995; Mirny & Domany, 1996; Miyazawa & Jernigan, 1996; Park & Levitt, 1996; Liwo et al., 1997; Ulrich et al., 1997), but it is not clear which formulation is best or which set of potential parameters is more accurate (Kocher et al., 1994; Godzik et al., 1995; Jones & Thornton, 1996). Thus, the current situation calls for careful examination of the theoretical foundations of the potential derivation methods. In this study, we focus on the parameterization methods under the assumption that the formulation of the potential functions is correct.

One term used commonly in the simplified potentials is the amino acid pair-specific contact potentials. These are designed to approximate interactions between noncovalently bonded amino acid residues. Other terms are employed to reflect local propensities of amino acid sequences for secondary structures (Kolinski et al., 1995; Rooman et al., 1995). Such potentials have been shown to be a valuable tool in structure modeling, protein folding, and inverse folding studies (Kocher et al., 1994; Sippl, 1995; Kolinski & Skolnick, 1996). In general, the methods used to derive contact potentials from a database of known protein structures can be classified into the two categories described below.

Reprint requests to: Jeffrey Skolnick, Department of Molecular Biology, TPC-5, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037; e-mail: skolnick@scripps.edu.

Boltzmann distribution method

The residue–residue contact frequencies observed in the native protein structures are assumed to obey a Boltzmann distribution. Thus,

$$E_{ij} = -k_B T \log(N_{ij}^{observed}/N_{ij}^{expected}), \quad (1A)$$

where i and j denote the amino acid residue types; E_{ij} is the extracted contact energy for an i-j contact; k_B is the Boltzmann's constant; T has a unit of temperature whose value is undetermined; $N_{ij}^{observed}$ is the number of i-j residue contacts observed in a structural database; $N_{ij}^{expected}$ is the number of i-j contacts expected in a reference state where there are no preferential interactions, such as in an unfolded random coil or in a randomly collapsed, compact state of proteins.

It is unclear how to define the reference state precisely, because it is an imaginary state that has no available structural database. In practice, the reference state is often built on the basis of the quasi-chemical approximation. This treats amino acid residues in proteins as unconnected entities in a thermodynamic equilibrium. It was suspected that such a crude approximation could cause serious

bias in the derived potentials (Miyazawa & Jernigan, 1996). However, in a recently developed method (Skolnick et al., 1997), the effects of chain connectivity, secondary structure, and chain compactness were explicitly included in construction of the reference state. It was found that these effects do not make a difference in the derived potentials. Thus, in this study, the quasichemical reference state was chosen (Godzik et al., 1995). Hence,

$$N_{ij}^{expected} = x_i x_j N_0, \tag{1B}$$

where i and j denote the amino acid types, x_i and x_j are the mole fraction of amino acid residues i and j in the database (also called the amino acid composition),

$$\sum_i x_i = 1, \tag{1C}$$

and N_0 is the total number of observed contacts in the database, i.e.,

$$N_0 = \sum_{ij} N_{ij}^{observed}. \tag{1D}$$

One should note that the presumed Boltzmann distribution expressed in Equation 1 is unusual, because a Boltzmann distribution usually describes an equilibrium within a particular system. Here, a database of different structures is considered. To understand the physical basis of the Boltzmann distribution across a database of structures, Finkelstein et al. (1995) proposed a theory based on random energy models of proteins and showed that a Boltzmann-like distribution arises naturally from low-energy conformations of random heteropolymers. A serious concern was raised about the presumed Boltzmann distribution in a recent study of Thomas and Dill (1996). In their two-dimensional lattice models, there are two residue types (hydrophobic, H, and polar, P). Pairs of residues interact with each other when they are in contact, with an energy of -1 for HH contacts and zero for HP or PP contacts. Because HP and PP contacts have equal energy, then, according to a Boltzmann distribution, an equal number of HP and PP contacts would be expected in the database of calculated minimum energy structures. However, significantly more HP contacts than PP contacts were found. It was argued that the problem originated from the fact that different types of contacts are not independent of each other due to chain connectivity.

Optimization method

Here, the assumption is that the native fold of a protein is the conformation with the lowest free energy and the protein has evolved to nearly optimum thermodynamic stability. The potential parameters are obtained by optimizing the stability of the native folds relative to misfolded structures (Goldstein et al., 1992; Maiorov & Crippen, 1992; Mirny & Shakhnovich, 1996). It was shown that the optimization procedure can recover the "true" potentials with a high linear correlation coefficient (ranging from 0.84 to 0.91) (Mirny & Shakhnovich, 1996). However, the repulsive terms in the potentials were systematically underestimated when compared with the attractive terms. The potentials also seemed to be overoptimized because the derived potentials gave an even higher stability than the true potentials.

There are also doubts about the optimization principle itself. After all, the native proteins may not have been maximally optimized for stability through evolution because it is known that sometimes mutations introduced in the native proteins can yield more stable proteins (Lim et al., 1994). The unique stable fold of a protein is apparently an essential feature required by its biological function. However, once that feature is achieved, there may not be any evolutionary pressure to make the protein even more stable (Serrano et al., 1993).

In summary, problems apparently exist in both of the potential derivation methods. The derived potentials and the true potentials correlate to some extent, but it is not simply a linear relationship, as expected previously. Consequently, it is necessary to examine the relationship between the derived and the true potentials in greater detail to search for the factors that affect the derived potentials. Thus, we have designed a simple model in which the true potentials are known and have applied various derivation methods to the databases constructed from minimum energy structures according to such true potentials. To avoid confusion, we note here that the term true potentials represents the original potentials used to construct the library of model structures; these do not necessarily correspond to the potentials experienced by real proteins. The derived potentials are those extracted from the constructed models of proteins. Our approach is similar to the works mentioned above (Mirny & Shakhnovich, 1996; Thomas & Dill, 1996), but here different questions were asked. Our study stresses the importance of generating native protein-like features in the constructed databases. Special attention was paid to examining how the stability of the structures affects the derived potentials and to identifying the necessary conditions for recovery of the true potentials. In addition, we address the question of what happens when there are different kinds of contributions to the potentials. Specifically, we explored the case where contact potentials and secondary structure propensities are considered simultaneously.

Following Godzik et al. (1995), the contact potentials can be decomposed into "ideal" terms and "excess" terms, defined as

$$E_{ij}^{ideal} = (E_{ii} + E_{jj})/2, \tag{2A}$$

$$E_{ij}^{excess} = E_{ij} - (E_{ii} + E_{jj})/2, \tag{2B}$$

where i and j denote residue types, E_{ii}, E_{jj}, and E_{ij} are contact energies. The ideal term (E_{ij}^{ideal}) defines a contact distribution where the mixing energy is zero. Godzik et al. showed that the excess term (E_{ij}^{excess}) largely depends on $N_{ij}^{observed}$, regardless of the choice of the reference state ($N_{ij}^{expected}$). In terms of E_{ij}^{excess}, potentials derived from various different methods were shown to agree with each other rather well (correlation coefficients typically range from 50 to 70%) despite differences in the contact definitions and data sets. Thus, the excess terms can be viewed as an intrinsic characteristic of native proteins. In the artificially constructed database of protein structures, the excess terms should agree with those from the real native proteins. Otherwise, the database can be viewed as being nonnative protein-like.

Another useful term is the quantity η_{ij}, which is defined as

$$\eta_{ij} = \frac{E_{ij} - (E_{ii} + E_{jj})/2}{E_{ii} - E_{jj}}. \tag{3}$$

Comparing Equation 3 with Equations 2A and 2B, it is easy to see that the numerator in Equation 3 is the excess energy term, whereas

the denominator is the difference in ideal terms. This factor is used because it is an invariant when E_{ii}, E_{jj}, and E_{ij} are subject to a linear transformation. Thus, if the true potentials and the derived potentials correlate linearly, the corresponding η_{ij} factor should be the same. Otherwise, the relationship between the two sets of potentials must be nonlinear.

Results

Model of proteins and database construction

A library of fragments excised from known native protein structures constitutes our structural database. These fragments are compact (with radii of gyration calculated from C_α coordinates of less than 14 Å) and have the same size (90 residues), but have different conformations (the RMS deviation, RMSD, of the C_α trace between any pair of structures is greater than 7.5 Å). A total of 2,553 fragments were extracted from 190 nonhomologous proteins, which are listed in Table 1.

For any given sequence, the pseudo native structure of this sequence is defined to be the conformation that has the lowest energy among all the conformations in the library. Using this definition, databases of such pseudo native structures were generated for a large number of random sequences.

To study the effect of structural stability on derived potentials, the pseudo native structures were sorted into different databases according to their stability measured in terms of their Z-scores (Sippl, 1993; see Equation 8). Structures of high stability were obtained by random searches in sequence space. The choice of Z-score as a measure of stability seems to be appropriate because the energy spectrum of all conformations forms a well-defined Gaussian distribution.

Care was also taken to keep the amino acid composition approximately fixed in the constructed structural databases. Amino acid sequences were generated by randomly picking residues from a residue pool having pre-defined amino acid residue composition. Starting from random sequences, Figure 1 shows the limiting values of Z-scores arising from a search of the sequence space. The initial random sequences had Z-scores around -3.5, and then evolved to sequences having Z-scores between -10 and -15. At each step of the search process, two mutation sites were randomly chosen in a sequence, and the residues were swapped within the sequence. In this way, the amino acid composition remains the same during the search process.

The compactness of the constructed protein models was monitored by the radius of gyration (calculated from the coordinates of the C_α atoms). Figure 2 shows the histograms of the radius of gyration of the protein models. The distribution of radius of gyration in all constructed databases is almost the same. Therefore, the constructed model proteins are very similar to the natural proteins in terms of amino acid composition and compactness.

In the constructed structural databases, some conformations were seen to occur much more often than others. About 17% of the conformations defined in the library are taken by 50% of the sequences as their lowest-energy conformation. However, because those sequences sharing a common conformation are not homologous to each other, they are not considered redundant structures in our analysis. A somewhat similar phenomenon was observed in some simple lattice models (Li et al., 1996). It was argued that such preferred structures are more "designable" and thus should prevail in nature through evolution. The relevance of our results to fold "designability" is beyond the scope of this study, but it will be addressed in future work.

Deriving potentials with the Boltzmann distribution

HP model

First, we consider the simplest case, the HP model, which has two residue types: H and P. The potential parameter set consists of only three energies: E_{HH}, E_{HP}, and E_{PP} for HH, HP, and PP contacts, respectively. E_{HH}, E_{HP}, and E_{PP} are referred to as the true

Table 1. *Proteins selected from the Protein Data Bank used to derive the potentials*[a]

2pcy	1molA	1hbg	2apr	2msbB	8fabB	1mbc	1colA	3sdpA	4cpv
2scpA	1gmfA	1ycc	1prcC	1ifc	1rbp	4ptp	2rhe	3cd4	2fb4H
1acx	1cobA	1paz	2azaA	2pabA	2gcr	2tbvA	2cna	2er7E	5hvpA
1f3g	4fgf	1nsbA	7timA	1gox	1ald	1pii	1wsyA	6xia	5rubA
2taaA	4enl	5p21	4fxn	3chy	5cpa	2trxA	1gp1A	1cseE	4dfrA
3adk	1gky	3pgm	1rhd	4pfk	3pgk	2yhx	2gbp	2liv	3grs
1trb	6ldh	1ipd	2pgd	8adh	1gd1O	7aatA	2ts1	1phh	3lzm
1lz1	9rnt	2sarA	1fkf	1snc	1aps	2sicI	8atcB	2tscA	4cla
1pyp	9wgaA	9pap	3blm	2cpp	8atcA	1csc	1ace	3cox	2cpkE
1fbpA	1fnr	1gstA	3gly	1lap	1ovaA	1lfi	1wsyB	2cyp	2glsA
3pmgA	8acn	2reb	6tmnE	1hgeA	3gapA	1prcH	1shaA	1ads	2end
2cpl	1gof	1sbp	2bbkH	1pda	1nar	1tml	1cmbA	1ede	1gpb
1rec	2mnr	3tgl	1aozA	1sltA	1ltsA	2abk	2ayh	1btc	2glt
1poc	2sim	1alkA	1dsbA	2dnjA	1poxA	2hhmA	1ndk	1minB	1rcb
2madL	1apa	1gal	1avhA	1cauA	1tplA	1atr	1aak	1add	1mat
1gdhA	2mtaC	2tmdA	3ecaA	1cde	1hmy	1tie	2aaiB	1glaG	1tbpA
1ula	2baa	1cid	1atnA	1nipB	1mioC	1mypC	2bpa2	1hc6	2pia
1tca	1top	1chmA	1hslA	1sacA	1trkA	2bb2	1lba	1vmoA	3sc2A
3sc2B	1bnh	1bbrE	1hplA	3aahA	2polA	1afnA	1pkp	1rpa	1tahA

[a]The PDB naming convention is used in the following list. The fifth letter in a name denotes the chain label.

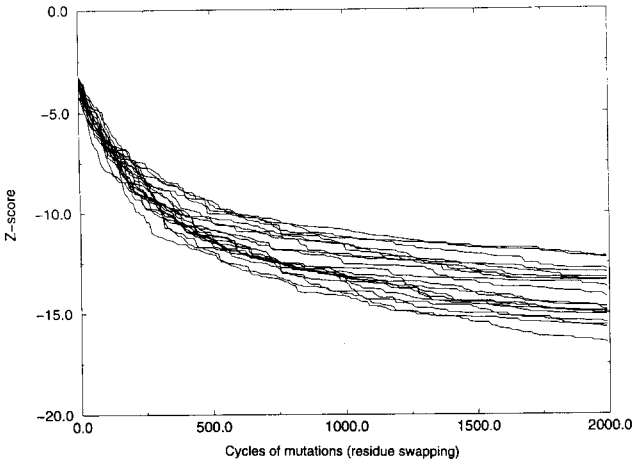

Fig. 1. Limit of Z-scores searched in sequence space. Random mutations were introduced to 20 random sequences to minimize the Z-scores of their lowest-energy conformations according to the contact potentials listed in Table 3A.

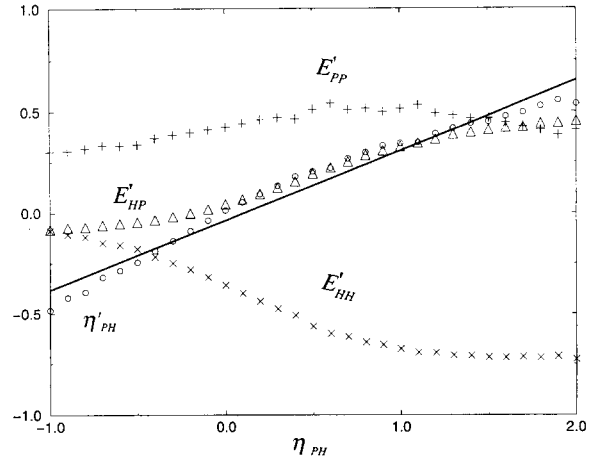

Fig. 3. Comparison of the derived versus the true contact potentials for HP model. For each η_{PH} value, 1,000 structures were generated for extracting the derived potentials.

potentials, whereas the derived potentials are denoted as E'_{HH}, E'_{HP}, and E'_{PP}, which were calculated from Equation 1.

Note that if the true potentials are multiplied by a constant or a constant is added to the potentials, the minimum energy structures remain the same. Therefore, a linear transformation operating on the true potentials does not change the derived potentials. This allowed us to use the factor η_{PH}, defined by Equation 3, to represent the true potentials. η_{PH} alone is an adequate descriptor for all the true potentials. If the derived potentials and true potentials correlate linearly, η_{PH} (calculated from E_{HH}, E_{HP}, and E_{PP}) should be equal to η'_{PH} (calculated from E'_{HH}, E'_{HP}, and E'_{PP}).

Figure 3 shows how the derived potentials are related to the true potentials in the HP model. The true potentials were $E_{HH} = -1$, $E_{PP} = 0$, and E_{HP} varied from -1.5 to 1.0. From Equation 3, it follows that $\eta_{PH} = E_{HP} + 0.5$. The derived potentials were calculated according to Equation 1, with $x_H = x_P = 1/2$. For each set of true potentials, 1,000 lowest-energy (nondegenerate) structures were calculated for 1,000 random sequences. These constitute the structural database for the derivation of E'_{HH}, E'_{HP}, E'_{PP}. The Z-scores of such structures are around -4.1.

From Figure 3, we see that as the true E_{HP} increases, the derived E'_{HP} and E'_{PP} increases, whereas the derived E'_{HH} decreases. The relationships are monotonic, but apparently nonlinear. However, the relationship between η_{PH} and η'_{PH} seems simpler. Linear fitting η_{PH} to η'_{PH} yields

$$\eta'_{PH} = 0.42\eta_{PH} - 0.02, \tag{4}$$

with a linear correlation coefficient of 0.994.

It is interesting to note that $\eta'_{PH}/\eta_{PH} \approx 0.42$ is a constant for any η_{PH} value. What then determines η'_{PH}/η_{PH}? Surprisingly, it was found that it depends on how stable the structures are. Equation 4 holds for structures having Z-scores around -4.1, whereas for structural databases with lower Z-score ranges, larger η'_{PH}/η_{PH} ratios are observed (see Table 2; note that each row represents the results obtained from a structural database characterized by Z-scores). The only case in which the derived potentials match the true potentials was in the database with structures having Z-scores around -9.6, where the η'_{PH}/η_{PH} ratio is close to 1.

It should be noted that the first row in Table 2 closely corresponds to the results obtained from the two-dimensional HP lattice models (Thomas & Dill, 1996). Thomas and Dill used the same "true" potentials (E_{HH}:E_{HP}:E_{PP} = -1:0:0), and found that the derived potentials rank as $E'_{HH} < E'_{HP} < E'_{PP}$. Because a complete set of short-chain lattice structures (excluding those that have degenerate ground state) was used in their study to extract the derived potentials, the Z-scores of such structures, on average, are not expected to be less negative than our databases. Thus, their results should correspond to our study in the case of random sequences, where the same ranking of the derived potentials can be found. But for databases with more negative Z-scores, the bias in the derived potentials diminishes.

These results suggest that the only way of recovering the true potentials is to take into account the stability of the structures.

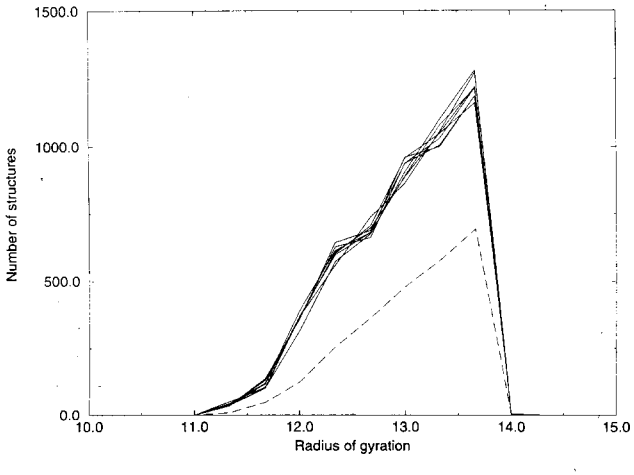

Fig. 2. Histograms of radius of gyration of proteins for proteins in the conformational library (dashed line) and proteins in constructed databases (solid lines). The constructed databases correspond to those shown in Table 5.

Table 2. *Ability to recover contact potentials in HP models using the Boltzmann distribution method*[a]

⟨Z⟩ ± RMSD	E'_{HH}	E'_{HP}	E'_{PP}	η'_{PH}/η_{PH}
−4.1 ± 0.4	−0.56	0.19	0.51	0.41
−4.3 ± 0.3	−0.59	0.22	0.54	0.44
−5.2 ± 0.2	−0.70	0.32	0.65	0.51
−6.2 ± 0.2	−0.80	0.43	0.77	0.56
−7.2 ± 0.2	−0.90	0.57	0.91	0.63
−8.1 ± 0.2	−0.99	0.75	1.03	0.73
−9.0 ± 0.4	−1.08	0.99	1.19	0.83
−9.6 ± 0.8	−1.14	1.21	1.27	0.95
−9.9 ± 1.0	−1.17	1.39	1.26	1.11

[a] E'_{HH}, E'_{HP}, and E'_{PP} are the derived contact potentials for HH, HP, and PP contacts, respectively. The "true" energies are: $E_{HH} = -1$; $E_{HP} = E_{PP} = 0$. Each database characterized by the range of Z-scores contains 1,000 structures.

When the structural database is composed of structures with "appropriate" stability, the distortion in the Boltzmann distribution is diminished so that the true potentials can be recovered accurately.

Models with 20 natural amino acid types

Next, proteins containing 20 natural amino acid types were considered. The amino acid composition in the constructed databases was kept the same as that in the native proteins listed in Table 1. Now the number of contact potential parameters is 210 (listed in Table 3), which is much more than in the HP model. Qualitatively, the same relationship was observed between the derived and the true potentials.

In the database constructed from random sequences without screening for stability, Z-scores of their lowest energy structures are around −3.5. As shown in the first row of Table 4, the potentials derived from this database linearly correlate with the true potentials (their correlation coefficient is 0.94). However, differences in potentials similar to those found for the HP models also exist here. Figure 4 shows the relationship between factor η_{ij} and η'_{ij}. Here, η_{ij} and η'_{ij} were calculated from the true potentials and the derived potentials, respectively (using Equation 3). The ratio $\eta'_{ij}/\eta_{ij} \approx 0.46$ is a constant largely independent of residue types. η'_{ij}/η_{ij} is thus simply denoted as η'/η. Because η'/η is less than 1.0, there is a systematic bias in the derived potentials, which is the same as that identified for the HP models.

Again, the stability of structures was found to affect η'/η. Table 4 shows the results obtained from databases of structures with various ranges of Z-scores. η'/η appears to increase as Z-scores decrease. Among the databases listed, of particular interest is the one with Z-scores around −8, where η'/η is close to 1.0. The contact potentials derived from a database of structures having Z-scores around −8 are shown in Figure 5. The true potentials and the derived potentials indeed correlate remarkably well.

Thus, it becomes important to have an estimate of the Z-scores of native protein structures with their native sequences. We took the sequences of the proteins listed in Table 1, whose size is less than 200 residues, and we threaded them (without gaps) through all the tertiary structures of proteins listed in Table 1. Using the potentials listed in Table 2, the Z-scores of native structures were calculated (shown in Fig. 6). The Z-scores of native sequences appear to be lower than those of random sequences. This is expected because the stability of native folds of proteins may have been acquired through long-time evolution. The average Z-scores of the native sequences is −8.0. Most of the Z-scores fall within the range of −10 to −7. Most proteins with Z-scores less negative than this range were found to be nonglobular and to form multimeric complexes. This result suggests that the stability of real native proteins is within a range that does not cause a bias in the Boltzmann distribution described in Equation 1.

However, it should be noted that there is a great deal of uncertainty in our estimated range of Z-scores of real native proteins because the actual potential energy function of a real protein is not known. Our so-called true potentials are true for our artificial proteins, but may not be true for real proteins. Nevertheless, for the protein models with 20 amino acid types and a wide range of Z-scores, the correlation between the derived and "true" potentials is always high. Thus, substantial errors are unlikely to arise in the derived potentials for real proteins.

Deriving contact potentials along with secondary structure propensities

Next, we examine how contact potentials and secondary structure propensities influence each other in potential derivations. Now, the conformational energy of the protein model is defined by

$$H = \sum_{ij} C_{ij} E_{ij} + \sum_{i} E^{i}_{\mu}, \tag{5}$$

where the summations are over all the residues of a sequence; E_{ij} is the contact potential, $C_{ij} = 1$ if residue i and residue j are in contact; otherwise, $C_{ij} = 0$; E^{i}_{ij} is the secondary structure propensity of residue j in the secondary structure type μ (for the derivation of E^{i}_{ij}, see Methods). Here μ can be "α" (3_{10}-helix and α-helix), "β" (β-strand), or "γ" (coil and turns).

The assumed true contact potentials and secondary structure propensities are listed in Table 3. Structural databases, each with 5,000 minimum energy structures, were constructed. The structures were sorted into different databases according to their range of Z-scores.

In Table 5, each row represents the results obtained from a structural database. Contact potentials were derived using Equation 1, whereas the secondary structure propensities were derived using Equation 7. In general, both kinds of potentials were recovered with high correlation. Thus, introducing secondary structure preferences does not seem to hinder the recovery of both potentials.

However, it should be noted that, in the constructed structural databases, the relative weight of the contact preferences to secondary structure propensities is not the same as that observed in the native proteins listed in Table 1. Here, the weight of the potentials was reflected in the "slopes" of the linear fit of derived potentials to the true potentials (Table 5). Relatively stronger secondary structure preferences were found in more stable structures.

For structures having Z-scores around −8, it was found that η'/η is close to 1.0. Along with the high correlation coefficient, the potentials derived from this database are considered most accurate. This result suggests that we can expect the derived potentials to be reliable, provided that the Z-scores of native proteins lie in an appropriate range and the functional form being used here to calculate conformational energies is correct.

Table 3. *The true potentials*

A. Contact potentials[a]

	Gly	Ala	Ser	Cys	Val	Thr	Ile	Pro	Met	Asp	Asn	Leu	Lys	Glu	Gln	Arg	His	Phe	Tyr	Trp
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. Secondary structure propensities[b]

AA	E^i_α	E^i_β
Gly	1.104	0.863
Ala	−0.583	0.122
Ser	0.308	0.269
Cys	0.213	−0.317
Val	−0.287	−1.023
Thr	0.425	−0.082
Ile	−0.574	−1.104
Pro	1.099	1.092
Met	−0.796	−0.645
Asp	0.373	0.878
Asn	0.626	1.031
Leu	−0.630	−0.588
Lys	−0.079	0.231
Glu	−0.553	0.149
Gln	−0.195	0.221
Arg	−0.205	0.090
His	0.070	0.067
Phe	−0.451	−0.708
Tyr	−0.152	−0.529
Trp	−0.200	−0.299

[a] Values listed in Table 3A are from Skolnick et al. (1997).
[b] See Equation 7 for definition of E^i_α and E^i_β.

Table 4. *Effect of structural stability on derived potentials in the protein models with 20 amino acid types*[a]

$\langle Z \rangle \pm$ RMSD	r	A	B	η'/η
−3.5 ± 0.4	0.94	0.81	−0.02	0.46
−5.1 ± 0.1	0.96	1.18	0.06	0.64
−6.1 ± 0.1	0.96	1.46	0.16	0.73
−7.1 ± 0.2	0.96	1.75	0.29	0.85
−7.9 ± 0.4	0.96	2.03	0.45	1.01
−8.9 ± 0.4	0.95	2.42	0.70	1.21
−9.9 ± 0.7	0.94	2.68	0.94	1.34

[a] r is the linear correlation coefficient between the derived and the true contact potentials; A and B are the slope and intercept, respectively, obtained from the linear least-square fitting of the derived potentials to the true potentials; η'/η was obtained from linear fitting η'_{ij} to η_{ij} for contact types that satisfies the condition that $|E_{ii} - E_{jj}| > 1.0$. Note that errors in η_{ij} become substantial when $|E_{ii} - E_{jj}|$ is small. Each database characterized by the range of Z-scores contains 5,000 structures.

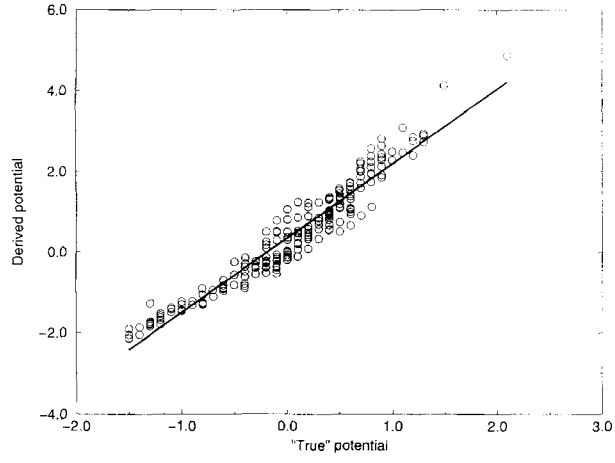

Fig. 5. Comparison between the true contact potentials and potentials derived from a database that has Z-scores around −8.

Deriving contact potentials by optimizing Z-scores

For the optimization method, it was found that whether or not secondary structure propensity is included makes no difference in the derived potentials; thus, we report here only the results that were obtained with known secondary structure propensities. The optimization method we used was adopted from that of Mirny and Shakhnovich (1996) and was applied to two sets of constructed structures. Each set contains 100 lowest-energy structures, but has different ranges of Z-scores. The first set, training set 1, was composed of random sequences with Z-scores around −3.5. The second set, training set 2, has Z-scores around −8, which is thought to be the most native protein-like.

Starting from random values, the derived contact potentials were put through a Monte Carlo procedure to optimize the harmonic mean of the Z-scores ($\langle Z \rangle_{harm} = 100/\{\Sigma 1/Z_i\}$) for each set of 100 structures (see Table 6). As shown in Figure 7, for proteins in training set 2, the optimization procedure drove $\langle Z \rangle_{harm}$ to converge to −10.3 after 4,500 cycles of optimization, whereas the derived potentials became increasingly correlated with the true potential, approaching a correlation coefficient of 0.74. The final derived potential is plotted against the true potential in Figure 8. Clearly, the repulsive terms in the true potentials tend to be systematically underestimated. This tendency may be intrinsic to the optimization procedure rather than to the details of the model of proteins because the same trend was also observed by Mirny and Shakhnovich (1996) in their lattice models.

Because the derived potentials were optimized with respect to their training sets, it is necessary to check how the derived potentials perform on a test set of other sequences. As shown in Table 6, we calculated the Z-scores using the derived potentials for two additional testing sets of structures. Little difference was observed between the test set proteins and the training set proteins. It is striking to see that, even for the testing sets of proteins, the Z-scores calculated from derived potentials are lower than those calculated

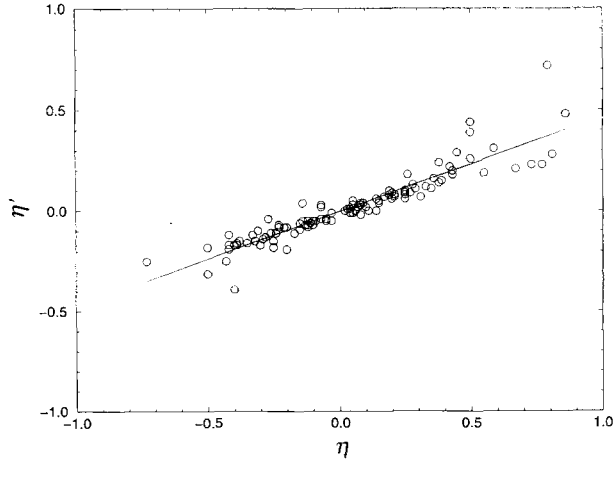

Fig. 4. Plot η' versus η. The derived potentials were computed from the database of random sequences without screening for stability (Z-scores of the structures are around −3.2). Note that data were plotted only for contact types that satisfy $|E_{ii} - E_{jj}| > 1.0$ and $|\eta| < 1.0$.

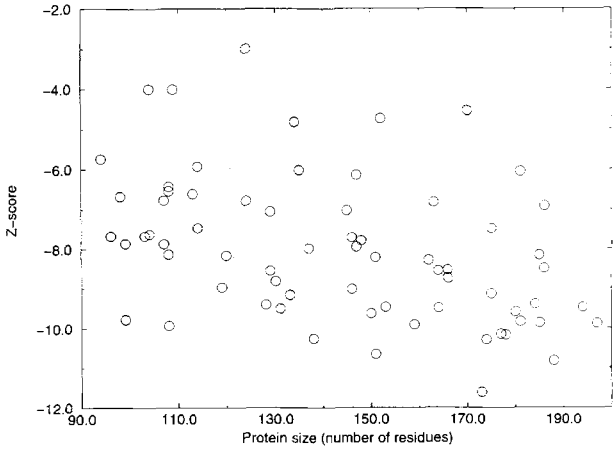

Fig. 6. Z-score of native protein structure. Z-scores were calculated by the potentials listed in Table 3.

Table 5. *Deriving contact potentials along with secondary structure propensity*[a]

	Contact potential				α-Propensity			β-Propensity		
⟨Z⟩ ± RMSD	*r*	*A*	*B*	η'/η	*r*	*A*	*B*	*r*	*A*	*B*
−3.5 ± 0.3	0.94	0.77	−0.02	0.60	0.95	0.37	−0.01	0.95	0.54	−0.02
−4.1 ± 0.1	0.95	0.90	−0.01	0.61	0.98	0.48	−0.02	0.96	0.63	−0.02
−5.1 ± 0.1	0.96	1.11	0.04	0.74	0.99	0.70	−0.01	0.98	0.81	−0.02
−6.1 ± 0.1	0.96	1.34	0.12	0.85	0.99	0.93	−0.01	0.98	0.98	−0.02
−7.1 ± 0.1	0.97	1.58	0.22	0.90	0.99	1.14	0.01	0.99	1.15	−0.01
−8.1 ± 0.1	0.96	1.85	0.34	1.01	0.98	1.39	0.04	0.97	1.29	0.01
−9.0 ± 0.2	0.96	2.10	0.51	1.08	0.97	1.66	0.08	0.95	1.44	0.03
−10.4 ± 1.0	0.95	2.50	0.80	1.37	0.96	2.09	0.16	0.90	1.66	0.11

[a] *r* denotes the linear correlation coefficients between the derived and the true potentials; *A* and *B* are the slope and intercept obtained from the linear least-square fitting of the derived potentials to the true potentials; η'/η is defined as in Table 4. Each database characterized by the range of *Z*-scores contains 5,000 structures.

from the true potentials. This means that the derived potentials, in general, are able to place the "native folds" at even more stable positions than the true potentials. Thus, the derived potentials appear to be overoptimized. Evidently, this is because the repulsive terms in the potentials were diminished. The difficult dilemma is that, before the derived potentials are overoptimized, they have not converged, but, once convergence has been reached, the potentials are overoptimized.

Further investigation revealed that these problems are particularly associated with the *excess* contributions to the potentials (see Equation 2B). The optimization procedure was altered such that the *excess* terms were kept as a constraint and only the *ideal* terms need to be parameterized. Starting from random, the derived ideal terms eventually converged to the "true" *ideal* terms with a linear correlation coefficient of 0.95 (Fig. 9). However, when the ideal terms were kept as a constraint, the derived excess terms only correlated with the "true" *excess* term with a correlation coefficient of 0.26.

Discussion

First of all, it is reassuring to see that the derivation methods in many cases can recover the true potentials quite accurately. Considering that database-derived potentials have been shown to perform well in inverse folding (Sippl, 1995) and that the derived potentials make good physical sense [they can be rationalized in terms of hydrophobicity, electrostatic interactions, etc. (Godzik, 1996; Miyazawa & Jernigan, 1996)], this provides us with greater confidence in the accuracy of the derived potentials. However, one has to bear in mind that the conclusions reached from our study are based on the assumption that the formulation of the potentials is correct. The key to successful force field development may lie in their formulation rather than in the parameterization. For structural prediction purposes, it may prove necessary to modify the definition of contact potentials and include additional terms in the energy functions to better assess packing, hydrogen bonding, etc., in proteins.

Second, it was found that different types of potentials can be derived separately using the Boltzmann distribution method. Contact potentials and secondary structure propensities were recovered separately from our constructed structural databases with remark-

Table 6. *Average Z-scores of training and testing set of proteins*[a]

Structure set	$\langle Z \rangle_{true}$		$\langle Z \rangle_{derived}$
Training set-1	−3.5 ± 0.4		−4.5 ± 0.6
Testing set-1	−3.5 ± 0.4		−4.5 ± 0.6
r		0.74	
Training set-2	−8.1 ± 0.1		−10.8 ± 0.8
Testing set-2	−8.1 ± 0.1		−10.6 ± 0.8
r		0.74	

[a] $\langle Z \rangle_{true}$ is the average Z-score calculated from the true potential; $\langle Z \rangle_{derived}$ is calculated from the potentials derived though the optimization procedure; *r* is the linear correlation coefficient between the derived potential and the true potentials. Set 1 structures are from random sequences, whereas set 2 structures are "designed" sequences, so they have lower Z-scores.

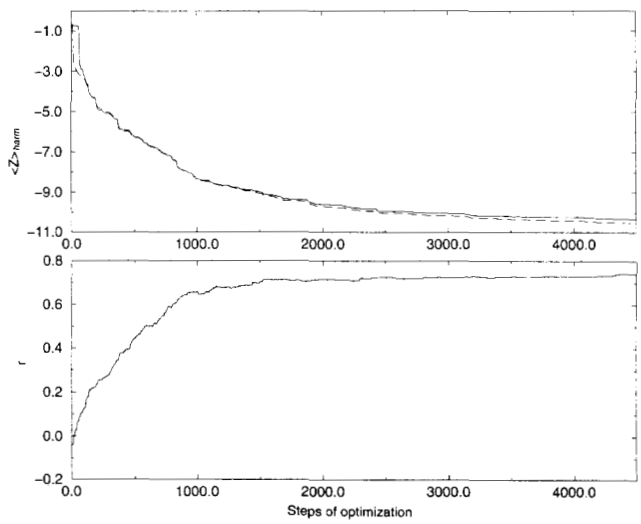

Fig. 7. Contact potentials derived by optimizing $\langle Z \rangle_{harm}$. The bottom line denotes the correlation coefficient between the derived and true potentials; the top solid line denotes the harmonic mean of the Z-scores of proteins in training set 2 and the dashed line denotes that of proteins in test set 2.

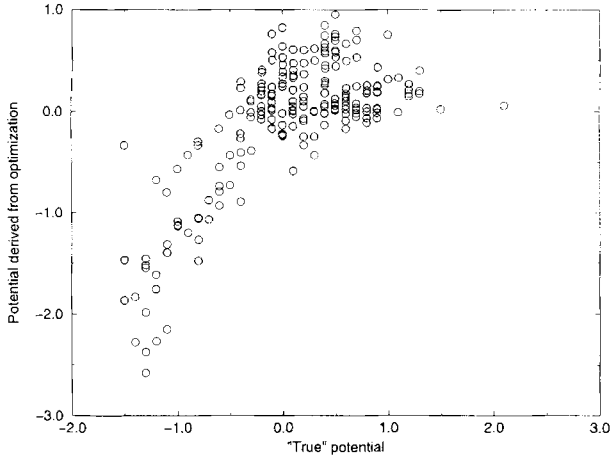

Fig. 8. Comparison of the true contact potentials and the potentials derived by optimizing the Z-scores.

able accuracy. However, because the contact potentials derived from α-helical proteins are different from those derived from β-sheet proteins (Godzik et al., 1995), the contact energy in proteins must be truly backbone dependent to account for this effect.

An important conclusion reached from this study is that stability of the structures in a database can affect the derived potentials. The effect of stability was seen in (1) the relative magnitude of excess versus ideal terms (defined in terms of η factor); and (2) the relative weighting of contact potentials versus secondary structure propensities.

The Boltzmann distribution method can sometimes fail. The errors in the derived potentials are evident in the HP models [both in our model and in the models of Thomas & Dill (1996)], but are less apparent in the models with 20 amino acid types. Our results thus do not contradict the results of Thomas and Dill, but depict a more general relationship between the derived and the "true" potentials. Our results of HP models and models with 20 amino acid types elucidate a consistent picture: the bias in the presumed Boltzmann distribution is best described by the ratio of η'/η, which increases with the stability of the structures in the database. However, the contact potentials derived from real proteins may happen to be very accurate because the stability of real native proteins is within a range that introduces little bias in the presumed Boltzmann distribution. In fact, from thermodynamic experimental data, the Z-scores of small proteins are estimated to be around -10 (L. Zhang & J. Skolnick, unpubl. work).

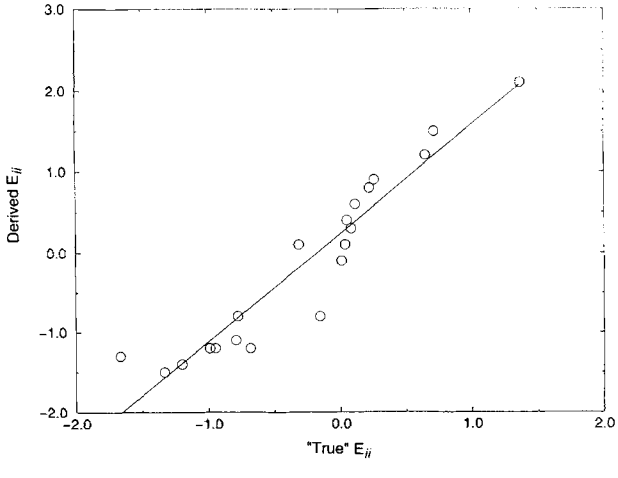

Fig. 9. Comparison of the *ideal* terms of the true contact potentials and that of potentials derived by optimizing the Z-scores. The *excess* terms are set as constraints in the optimization procedure.

In general, the stability optimization method may be less accurate than the Boltzmann method of deriving potentials. Our results are similar to Mirny and Shakhnovich (1996), who also found that the derived potentials systematically underestimate the repulsive terms. Because this phenomenon has been observed in various model systems, we believe the problem is rooted in the fact that native proteins are not maximally optimized in terms of contact potentials. Simply driving the Z-scores to their minima only results in overoptimization, which distorts the derived potentials. A new insight from our study is that the errors in the derived potentials are mostly in the *excess*, but not in the *ideal*, contribution. In contrast, in the Boltzmann distribution method, the excess terms (regardless of sign) are underestimated relative to the ideal terms.

Because of the simplicity of the protein models used, the relevance of this study to real proteins may be questionable. One of the concerns is that the size of the studied model's conformational space may be too small. However, this limited size does not seem to affect the results obtained in this study. We observed that even taking a subset of these 2,553 conformations made no difference. We also used lattice models of proteins (Kolinski & Skolnick, 1994) to define the conformational library and obtained similar results (data not shown). It remains to be seen whether the results obtained here still hold for the cases where the conformational space is much larger. However, given that similar results can be obtained from very different models of proteins, the conclusions reached from this study are unlikely to be artifacts of the models we used.

Methods

Conformational library

The names of the proteins used to construct a library of peptide fragments are listed in Table 1. A sliding window of 90 residues, each step sliding by 1 residue, was used to excise the fragments for all these proteins. To exclude noncompact and redundant structures, fragments having a radius of gyration larger than 14 Å (only C_α atoms are counted) were discarded, as were the fragments that can be superimposed onto each other within 7.5 Å RMSD (only C_α atoms are counted). This resulted in 2,553 fragments, which were then used as the conformational library. It was assumed that any given sequence with a length of 90 residues could take only a conformation within this library.

The contact maps are frozen, i.e., when a new sequence is threaded in one of the 2,553 fragments, the contact map remains the same as in the original fragment. This is equivalent to assuming that all amino acids have the same size and there is no repacking of the side-chain groups upon mutation. The definition of a contact is that there are two heavy atoms within 4.5 Å belonging to side chains of two residues and there is no covalent bond between the two residues.

Calculation of secondary structure propensity

In this study, we used a three-state classification of protein secondary structure types: "α" (α-helix, 3_{10}-helix), "β" (β strand), and "γ" (coil and turn). The secondary propensity of a residue is denoted as E_μ^i, which is the energy of a residue of type i in secondary structure μ (μ can be "α," "β," or "γ"). The definition of E_μ^i follows that of the method of Chou and Fasman (1974):

$$E_\mu^i = -k_B T \log\left(\frac{f_\mu^i}{\langle f_\mu \rangle}\right), \tag{6A}$$

where f_μ^i is the observed frequency of residue i in state μ; $\langle f_\mu \rangle$ is the observed frequency of state μ averaged over all residues. Using N_μ^i as the observed number of residues of type i in state μ, we have

$$f_\mu^i = \frac{N_\mu^i}{N_\alpha^i + N_\beta^i + N_\gamma^i} \tag{6B}$$

and

$$\langle f_\mu^i \rangle = \frac{\sum_k N_\alpha^k + \sum_k N_\beta^k + \sum_k N_\gamma^k}{\sum_k N_\mu^k}, \tag{6C}$$

where the summation is over all 20 amino acid types (k). Because it is only the relative differences between E_α^i, E_β^i, and E_γ^i, that have any real meaning, we take the "γ" state (coil and turns) as the reference state so that E_γ^i is set to zero and the helix propensity of an amino acid residue becomes

$$E_\alpha^i = -k_B T \log\left(\frac{N_\alpha^i}{N_\gamma^i} \cdot \frac{\sum_k N_\gamma^k}{\sum_k N_\alpha^k}\right), \tag{7A}$$

and the β-strand propensity of an amino acid residue becomes

$$E_\beta^i = -k_B T \log\left(\frac{N_\beta^i}{N_\gamma^i} \cdot \frac{\sum_k N_\gamma^k}{\sum_k N_\beta^k}\right). \tag{7B}$$

Construction of the structural database

The protocol used to construct a structural database is as follows:

1. Assume a set of "true" potentials.
2. Generate a random amino acid sequence by randomly picking residues from a pool of amino acid residues with a pre-defined amino acid residue composition.
3. Use the true potentials to calculate the conformational energies for all conformations in the constructed library and pick the lowest-energy conformation. Sequences that have two lowest-energy conformations with equal energy (these cases were rare) are discarded.
4. If it is required that the lowest energy of a sequence should have a Z-score in a specified range, then mutations are introduced in the sequence by swapping residues within the sequence (mutation sites were selected randomly), and step 3 is repeated until the Z-score of the lowest-energy conformation is appropriate.
5. Repeat steps 2, 3, and 4 until a desired number of lowest-energy conformations are collected.

Definition of Z-scores

We measure the stability of a conformation in terms of Z-scores, defined as follows:

$$Z = \frac{E_0 - \langle E \rangle}{\sigma_E}, \tag{8}$$

where $\langle E \rangle$ is the conformational energy averaged over the conformational library for a given sequence, σ_E is the standard deviation of this ensemble, and E_0 is the conformational energy of interest. Note that the Z-scores were used in two different contexts in this study: (1) for constructing databases of lowest-energy structures, the Z-scores of lowest-energy conformation were calculated by true potentials as a criterion for selecting sequences; and (2) for deriving the potentials through optimization of Z-scores, the Z-scores of given folds were calculated using "guessed" potential parameters. The two cases should not be confused: the Z-scores were optimized by introducing mutations in the sequences in the first case, whereas in the second case, the Z-scores were optimized by adjusting "guessed" potentials.

Correlation coefficient

The correlation coefficient between two sets of numbers ($\{x_i\}$, $\{y_i\}$, $i = 1, 2, 3, \ldots, N$) is defined as follows:

$$r = \frac{\sum x_i y_i - \sum x_1 \sum y_i / N}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}, \tag{9}$$

where $\bar{x} = \sum x_i / N$, $\bar{y} = \sum y_i / N$ and the summations are over i from 1 to N.

Acknowledgments

This research was supported in part by NIH grant GM48835. L.Z. is an NIH Postdoctoral Fellow. Useful discussions with Drs. Angel R. Ortiz, Wei-Ping Hu, Leszek Rychlewski, and Adam Godzik are gratefully acknowledged.

References

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

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

Finkelstein AV, Reva BA. 1996. Search for the most stable folds of protein chains: I. Application of a self-consistent molecular force field theory to a problem of protein three dimensional structure prediction. *Protein Eng 9*:387–397.

Godzik A. 1996. Knowledge-based potentials for protein folding: What can we learn from protein structures? *Curr Biol 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.

Goldstein R, Luthey-Schelten ZA, Wolynes P. 1992. Optimal protein-folding codes from spinglass theory. *Proc Natl Acad Sci USA 89*:1282–1286.

Jones DT, Thornton JM. 1996. Potential energy functions for threading. *Curr Opin Struct Biol 6*:210–216.

Kocher JPA, Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol 235*:1598–1613.

Kolinski A, Milik M, Skolnick J. 1995. A reduced model of short range interactions in polypeptide chains. *J Chem Phys 103*:4312–4323.

Kolinski A, Skolnick J. 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins Struct Funct Genet 18*:338–352.

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

Levitt M. 1976. A simplified representation of protein conformation for rapid simulation of protein folding. *J Mol Biol 227*:876–888.

Li H, Heling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science 273*:666–669.

Lim WA, Hodel A, Sauer RT, Richards FM. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA 91*:423–427.

Liwo A, Pincus MR, Wawak RJ, Rockovsky S, Oldziej S, Scheraga HA. 1997. A united-residue force field for off lattice protein-structure simulations. II. Parameterization of short-range interactions and determination of weights of energy terms by Z-score optimization. *J Comput Chem 18*:874–887.

Maiorov VN, Crippen GD. 1992. Contact potential that recognizes correct folding of globular proteins. *J Mol Biol 227*:867–888.

Mirny L, Domany E. 1996. Protein fold recognition and dynamics in the space of contact maps. *Proteins Struct Funct Genet 26*:391–410.

Mirny LA, Shakhnovich EI. 1996. How to derive protein folding potential? A new approach to an old problem. *J Mol Biol 264*:1164–1179.

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.

Rooman MJ, Kocher JP, Wodak SJ. 1995. Prediction of protein backbone conformation based on 7 structure assignments: Influence of local interactions. *J Mol Biol 221*:961–979.

Serrano L, Day AG, Fersht AR. 1993. Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J Mol Biol 233*:305–312.

Sippl MJ. 1993. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des 7*:473–501.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Skolnick J, Jaroszewski L, Kolinski A, Godzik A. 1997. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci 6*:676–688.

Thomas PD, Dill KA. 1996. Statistical potentials extracted from protein structures: How accurate are they? *J Mol Biol 257*:457–469.

Ueda Y, Taketomi H, Go N. 1978. Studies of protein folding, unfolding and fluctuations by computer simulations. II. A three dimensional lattice model of lysozyme. *Biopolymers 17*:1351–1548.

Ulrich P, Scott W, van Gunsteren WF, Torda AE. 1997. Protein structure prediction force fields: Parameterization with quasi-Newtonian dynamics. *Proteins Struct Funct Genet 27*:367–384.

Wilson C, Doniach S. 1989. Computer model to dynamically simulate protein folding: Studies with crambin. *Proteins Struct Funct Genet 6*:193–209.