J. Mol. Biol. (1996) **264**, 1164–1179

JMB

How to Derive a Protein Folding Potential? A New Approach to an Old Problem

Leonid A. Mirny and Eugene I. Shakhnovich*

Harvard University
Department of Chemistry
12 Oxford Street, Cambridge
MA 02138, USA

In this paper we introduce a novel method of deriving a pairwise potential for protein folding. The potential is obtained by an optimization procedure that simultaneously maximizes thermodynamic stability for all proteins in the database.

When applied to the representative dataset of proteins and with the energy function taken in pairwise contact approximation, our potential scored somewhat better than existing ones. However, the discrimination of the native structure from decoys is still not strong enough to make the potential useful for *ab initio* folding. Our results suggest that the problem lies with pairwise amino acid contact approximation and/or simplified presentation of proteins rather than with the derivation of potential. We argue that more detail of protein structure and energetics should be taken into account to achieve energy gaps. The suggested method is general enough to allow us to systematically derive parameters for more sophisticated energy functions. The internal control of validity for the potential derived by our method is convergence to a unique solution upon addition of new proteins to the database. The method is tested on simple model systems where sequences are designed, using the preset ''true'' potential, to have low energy in a dataset of structures. Our procedure is able to recover the potential with correlation $r \approx 91\%$ with the true one and we were able to fold all model structures using the recovered potential. Other statistical knowledge-based approaches were tested using this model and the results indicate that they also can recover the true potential with high degree of accuracy.

**Corresponding author*

Keywords: protein folding; protein folding potential; fold recognition

Introduction

The problem of how to determine the correct energetics is paramount to the complete solution of the protein folding problem.

Two avenues to determine energy functions (force-fields) for proteins have been pursued. The first uses more or less rigorous or semi-empirical classical or quantum mechanical calculations to determine, from the first principles, and/or fitting to spectroscopic experimental data, the forces acting between amino acids in a vacuum or in solution (Vasques *et al*., 1994). This approach is rigorous but it encounters formidable computational difficulties. Most importantly, it can be realized only within the framework of detailed, atomistic description of amino acids. However, detailed atom resolution models of proteins are not feasible for folding simulations due to obvious computational difficulties.

An alternative, more practical, approach is to introduce simplified, coarse-grained models of proteins where amino acids are represented in a simplified way, as one of few interacting centers which may have some internal degrees of freedom as well but which are generally much simpler than real amino acids (Levitt, 1976; Ueda *et al*., 1978; Miyazawa & Jernigan, 1985; Wilson & Doniach, 1989; Skolnick & Kolinsky, 1990; Shakhnovich *et al*., 1991). Such models are more tractable computationally in both threading approaches (Finkelstein & Reva, 1991; Jones *et al*., 1992) and *ab initio* simulations (Kolinsky & Skolnick, 1993, 1994). However, the serious problem with simplified representations of proteins is that it is difficult to describe protein energetics at the coarse-grained level of structure description. What ''force-fields'' should act between these simplified interacting centers, which remain identified with natural amino acids, such that native structures, for these

0022–2836/96/501164–16 $25.00/0

model proteins, still correspond to pronounced energy minima for their respective sequences? To address this problem Tanaka & Scheraga (1976) proposed an approach which was later developed by Miyazawa & Jernigan (1985) in their seminal contribution (the ''MJ'' method). The MJ method is based on a statistical analysis of protein structures and frequencies of contacts, defined in the realm of simplified protein representation. Frequencies of individual amino acid contacts were derived and compared with frequencies expected in the random mixture of amino acids and the solvent. Next, a quasichemical approximation was employed relating these properly normalized frequencies with ''potentials'' *via* the relation:

$$u_{ij} = -T \ln f_{ij}$$

where i and j denote amino acid types; f_{ij} are normalized frequencies of contacts between them extracted from the database of existing structures. The definition of the energy scale denoted as ''temperature'' (T) in the quasichemical approximation of MJ is a delicate problem. It was addressed in a recent work by Finkelstein *et al*. (1993, 1995) who also showed that quasichemical approximation may be a reasonable one under the assumption that protein sequences are random. In the recent study Mirny & Domany (1996) showed that quasichemical approximation is also valid if contacts are independent and uniformly distributed. The subsequent development of the knowledge-based approach based on quasichemical approximation included efforts to incorporate distance-dependent forces (Sippl, 1990), better representation of amino acid geometry and approximation of multiple-body interactions (Kolinsky & Skolnick, 1993, 1994), dihedral angles (Kolaskar & Prashanth, 1979; Nishikawa & Matsuo, 1993; Rooman *et al*., 1992; DeWitte & Shakhnovich, 1994), and better treatment of solvent–protein interactions (Park & Levitt, 1996). A detailed analysis of knowledge-based potentials and examples of their successful and unsuccessful application is given by Kocher *et al*. (1994). Approaches to derive potentials from quasichemical approximation, especially the most difficult issue of reference state, are discussed by (Godzik *et al*., 1995). Real potential is believed to distinguish the native structure by making its energy much lower than energy of all other conformations, i.e. it provides stability of the native structure. Protein sequences should also fold fast to their respective native conformations. It was shown, for simple models of proteins, that these two conditions, thermodynamic stability and kinetic accessibility, are met when the native state is a pronounced energy minimum for the native sequence, compared to the set of misfolded conformations (Sali *et al*., 1994; Shakhnovich, 1994; Gutin *et al*., 1995). Therefore it is reasonable to suggest that the essential property of the correct, ''true'', folding potential is that the energy of a native sequence folded into its respective native conformation should be much lower than the energy of this sequence in every alternative conformation.

An approach to derivation of protein folding potentials that takes this requirement explicitly into account was proposed by Goldstein *et al*. (1992; GSW) and Maiorov & Crippen (1992). Goldstein *et al*. maximized the quantity T_f/T_g, which is equivalent to the maximization of energy gap between the native state and bulk of decoys. They showed that, for each individual protein, the problem of potential optimization has a simple analytical solution; however their approach encountered a serious problem in averaging the potential over different structures in the database. Indeed, a potential optimized for one protein is not necessarily (and in fact never!) optimal for another protein, while the goal is to find a potential which is good, or optimal, simultaneously for many proteins. Goldstein *et al*. (1992) found an *ad hoc* procedure of averaging over a protein database which offered good results in their tests.

Here, we suggest a systematic method for deriving a potential which delivers a pronounced energy minimum to all proteins in their native conformations and hence should provide fast folding and stability of model proteins. The method is general and is not limited to any form of potential or any model of a protein. Another important feature of this approach is that it has internal criteria of self-consistency: when the derived potential does not change significantly upon addition of new proteins to the database, it corresponds to meaningful, nontrivial energetics.

The proposed new method of potential derivation should be rigorously tested and compared with existing approaches. However, there is a serious problem with testing parameters derived by any approach: the lack of objective rigorous criteria of success because true potentials are not known (and they are not likely to exist since real proteins differ from their simplified representations). A reasonable criterion is whether the derived potentials are useful for fold recognition and *ab initio* folding. The results of numerous tests by many groups (see the comprehensive analysis in the special issue of Proteins (Lattman, 1995)) show that, while existing knowledge-based potentials often do a decent job in fold recognition, they are not sufficiently accurate for *ab initio* folding. Strong evidence that the ''bottleneck'' in *ab initio* folding is in the energy function rather than in the search strategy is that *ab initio* procedures fail because decoys with energy lower than energy of the native conformations are found in test cases (Covell, 1994; Elofson *et al*., 1995). Conversely, in inverse folding tests, the native structure, in most cases (though not always), has lower energy than decoys (Wodak & Rooman, 1993; Lemer *et al*., 1995; Miyazawa & Jernigan, 1996).

The best way to assess different procedures of derivation of potentials is to use, as a test case, a

model system where the correct form of the Hamiltonian (say, pairwise contact potential) is given, and the true potential is known. Different procedures to extract potentials can be applied, and then the true and derived potentials can be compared. Further more, the derived potentials can be used in a model system for threading or *ab initio* folding to compare their performance with that of the true potentials and thus to close the circle.

Such a comprehensive analysis is possible in the realm of lattice models. Thomas & Dill (1996) took the first step in this direction when they considered two-dimensional short lattice chains composed of monomers of two types.

Here, we test our procedure for derivation of potentials as well as other approaches using both: (1) representative set of the native proteins obtained from the Brookhaven protein databank (PDB); and (2) three-dimensional lattice model proteins composed of 20 types of amino acids.

A sequence design algorithm has been recently developed which generates sequences with specified relative energy (*Z*-score, Bowie *et al*., 1991) in a given conformation (Shakhnovich & Gutin, 1993a,b; Abkevich *et al*., 1995a). We use this method to carry out the following rigorous procedure for testing our and alternative approaches to derive potential. (1) Select at dataset of the native protein conformations. We use a representative subset of the native structures obtained from PDB for real proteins and a set of random compact lattice conformations for the lattice model. (2) Using some potential to serve as the true potential for the model, design sequences to have selected conformations as native ones, thus creating a model protein databank. (3) Using different procedures, we extract ''knowledge-based potentials''. (4) Compare derived potentials with the true potential. Test the performance of the derived potentials in *ab initio* folding simulations (for the lattice model) and threading (for real proteins), using model proteins from the built dataset that has not been used to derive potentials.

This approach to test the potentials allows one to systematically test by different parameter derivation procedures how large the database size must be to successfully derive parameters. Furthermore, we can create databanks of model proteins with various levels of stability that make it possible to determine how well optimized protein sequences must be to allow successful parameter derivation.

We apply the procedure described above (except *ab initio* folding tests) not only to lattice model proteins but to also real proteins. With a set of true parameters, we design sequences for a representative set of protein conformations, derive the optimized potential, and evaluate the maximal value of *Z*-score for a given model Hamiltonian. This value is compared with scores for native sequences with the same optimized potential. This comparison sheds light not only on the advantages and pitfalls of the parameter derivation procedure but, more importantly, it determines which models can serve better for prediction of protein conformations.

In subsequent chapters of this paper we will carry out this program.

Results

Lattice model

We consider a conformation of a protein chain as a self-avoiding walk on a cubic lattice. Two amino acids that are not nearest neighbors in sequence and are located in the next vertices of the lattice are considered in contact. Energy of a conformation is given by equation (6).

Dataset of stable and folding proteins

The dataset of lattice proteins consists of 200 randomly chosen compact conformations of 27-mer on a $3 \times 3 \times 3$ cube (Shakhnovich & Gutin, 1991; Shakhnovich *et al*., 1991; Sali *et al*., 1994; Socci & Onuchic, 1994). We derive the potential using the first 100 of the lattice proteins and then test the derived potential for the remaining 100 lattice proteins from the dataset. We use the potential obtained by Miyazawa & Jernigan (1985) as the true one. Using the true potential for each native conformation in the dataset, we design a sequence which minimizes *Z*-score for this conformation (see above). The stability and folding of each designed sequence are tested by the Monte Carlo folding simulations. Each starts from a random coil (Shakhnovich *et al*., 1991; Sali *et al*., 1994) and continues until it reaches its respective native conformations.

Derivation of potential

To obtain the potential which minimizes *Z*-scores for model proteins, we use Monte Carlo procedure in the space of potentials (see Methods). Starting from different random potentials, the Monte Carlo search converges fast, and the resulting potential does not depend on the starting random potential. The procedure converges to a unique potential even at zero optimization temperature $T_{\mathrm{opt}} = 0$. This shows that there is only one minimum in the space of potentials in our model. This guarantees that the derived potential provides the global minimum to the target function $\langle Z \rangle_{\mathrm{harm}}$.

Derived *versus* true potential

The potentials obtained from this method are compared with the real one in several ways. Figure 1 presents 210 values of interactions in the derived potential *versus* the same values for the true potential. Correlation $r = 0.84$ shows that our method is able to reconstruct the true potential.

The values of energy for attractive interactions ($U(\xi, \eta) < 0$) are predicted much better than the

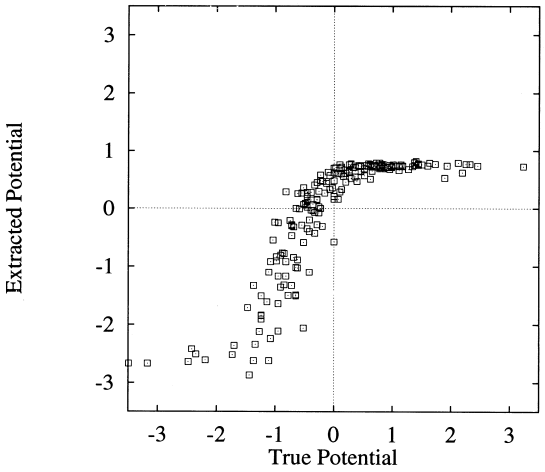

Figure 1. Derived potential *versus* true potential for the lattice model.

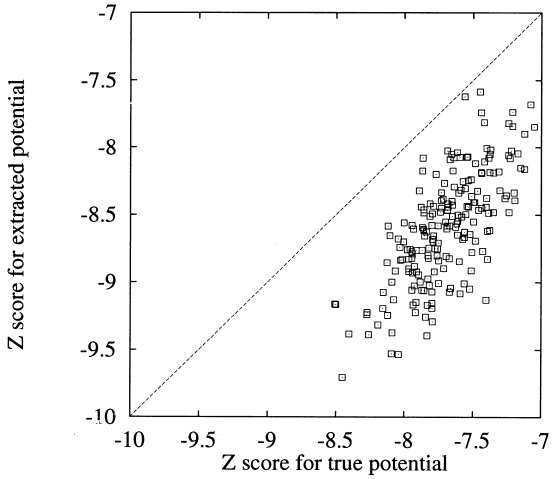

Figure 2. *Z*-score of model proteins with the derived potential *verus* *Z*-score with the true potential.

energy values of repulsive interactions ($U(\xi, \eta) > 0$). Attractive interactions stabilize the native conformations and appear much more frequently among native contacts. Repulsive interactions, in contrast, are very rare among native contacts and therefore the statistics is much poorer for them. Some repulsive contacts cannot be found in the dataset of model proteins. In contrast, contacts between all amino acids are present among native contacts in real proteins (see below). The absence of contacts between some types of amino acids in the model dataset is the result of a very strong sequence design. The design finds a sequence that provides very high stability of the native conformation in the given model, and by doing so it eliminates repulsive contacts which destabilize the native conformation. This observation is the first indicator that native sequences do not appear to be well designed for stability in terms of our model contact pairwise potential, 4.5 Å cutoff of residue-residue interactions etc.).

Ab initio folding with derived potentials

Ability of the derived potential to fold model proteins is tested by their *ab initio* folding. Folding simulations are carried out using the standard Monte Carlo method for polymers on a cubic lattice. A discussion of the lattice Monte Carlo simulation technique, and its advantages and caveats, is detailed in many publications (see e.g. Sali *et al*., 1994; Socci & Onuchic, 1994). Each simulation starts from a random coil conformation, proceeds at constant temperature and lasts about five times longer than the mean folding time.

All of the tests are performed for the proteins that were not used in the derivation procedure. First we compare $Z(\mathbf{U}_{der})$ values provided by derived potential with $Z(\mathbf{U}_{true})$ values for the true potential. Figure 2 presents $Z(\mathbf{U}_{der})$ as a function of $Z(\mathbf{U}_{true})$. Derived potential provides almost the same or even lower values of *Z*-score for all proteins.

We define folding time for each protein as a mean first passage time, i.e. time when the native conformation is first reached. Time is measured in MC steps. Forty runs are performed for each protein for both true and derived potentials. Simulations are run at temperature $T = 0.7$. Figure 3 presents the scatter plot of folding time obtained for the derived potential *vs* folding time for the true potential. All proteins that fold with the true potential fold also with the derived potential and exhibit approximately the same folding time.

The folding test proves that the derived potential is able to provide fast folding for all proteins with well-designed sequences.

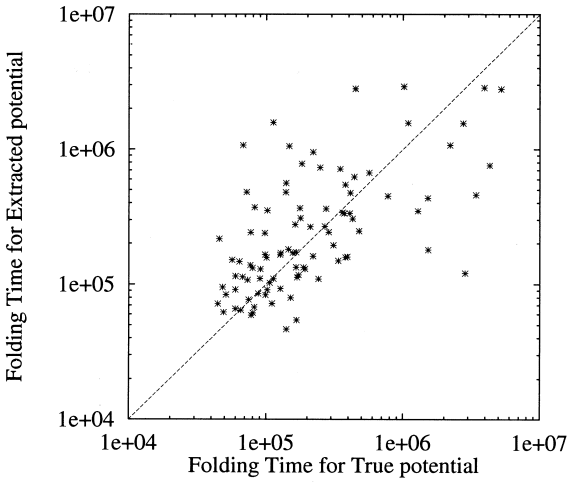

Figure 3. Folding time with derived potential *versus* folding time with the true potential for the lattice model. 100 lattice model proteins not used for derivation of potentials were taken for Monte-Carlo folding simulations. Folding ''time'' is measured in Monte-Carlo steps required to reach the native conformation.

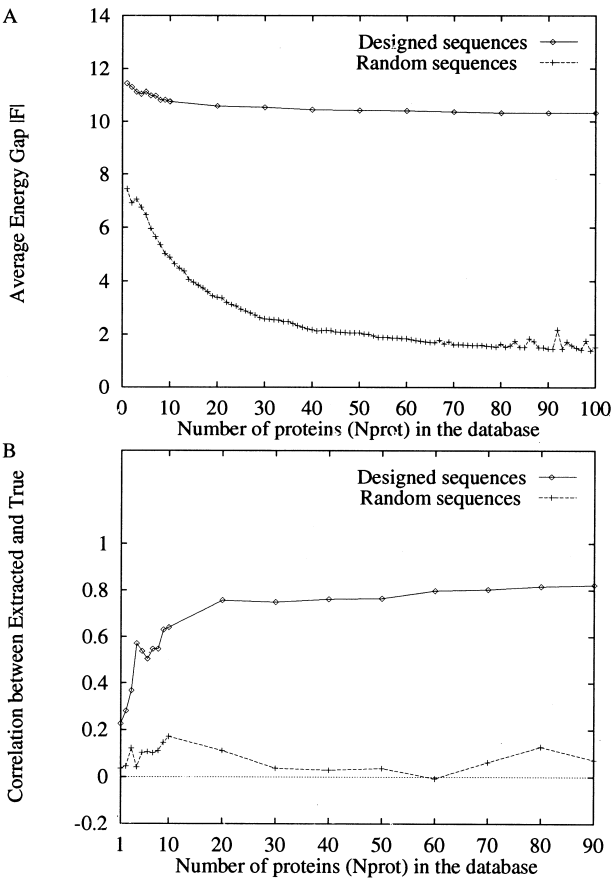

Figure 4. Effect of the database size, used for derivation of the potential, on average energy gap (A) and convergence test (B), for lattice model proteins.

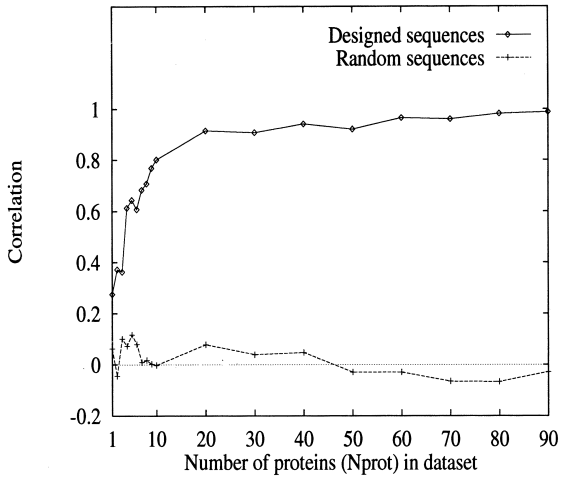

Figure 5. Convergence of potential for lattice model proteins. Correlation between potentials derived from all 100 model proteins and potential derived from $N_{prot} < 100$ proteins is shown as a function of N_{prot}.

Effect of the database size

How sensitive is the derived potential to the number of proteins used in the derivation? How many proteins are required to obtain a potential similar to the true potential? To address these questions we perform a derivation of potential for databases with various numbers of proteins.

For the database containing N_{prot} proteins we derive a potential using the technique described above and compute the average score (energy gap) $|F| = |\langle Z \rangle_{harm}|$ provided by the derived potential. (Figure 4A). We also compute the correlation of the true potential and the one obtained for N_{prot} proteins (see Figure 4B).

For few (1 . . . 5) proteins one can obtain a potential that provides a very large energy gap for these proteins. This potential, however, fails to provide a reasonable gap for other proteins and is not similar ($r = 0.2 \ldots 0.5$) to the true potential. As the number of proteins in the database increases, the average energy gap decreases approaching a rather high constant value of ($|F| = 1.6$). The correlation between the derived and true potentials approaches a constant value, $r = 0.85$. To ensure that the derived potential converges to a meaningful value as the number of proteins in the database increases, we compare the potential obtained for N_{prot} proteins with the one obtained for all 100 proteins. The correlation between these potentials as a function of N_{prot} is shown in Figure 5. Clearly, as the number of proteins in the database increases, the correlation between the derived potentials approaches 1 and, hence, the potential converges to a unique solution.

The results of this procedure clearly demonstrate the stability of our procedure. It is also important that the potential, which is highly correlated with the true potential ($r = 0.8$), can be obtained with only 40 . . . 50 proteins, which is of the order of the size of the database of non-homologous stable disulfide-free proteins available from the PDB. Convergence of the derivation procedure also guarantees that obtained potential does not depend on the number of specific properties of the proteins used for the derivation if the database is large enough.

Are there enough parameters? Are there too many parameters?

An important issue is whether the number of parameters adjusted in the potential is sufficient to provide a large enough gap for all proteins with designed sequences. For example, two-letter (HP) models are too non-specific to make the native structure unique: for any sequence, many conformations of three-dimensional HP heteropolymers have the same energy as the native conformation has. The native state in such models is not unique in most cases; correspondingly no sequence, random or designed, can have any energy gap in the HP models (Yue *et al*., 1995).

On the other hand, the number of parameters should not be too large. If the number of parameters is too large it is always possible to find a ''potential'' for which all members of the database used in the derivation have low energies, but the

resulting potential is unrelated to the true potential and does not provide low energy to proteins which are not members of the derivation database.

More specifically, the question is whether the problem of finding parameters is under-determined or over-determined, i.e. how the number of independent functions to minimize Z-scores of individual proteins is related to the number of independent parameters. For an over-determined problem the number of functions/constraints is greater than the number of parameters and, hence, there is no solution that minimizes all the functions well. This is not the case for our designed sequences, since there is true potential which provides a large enough gap for all proteins. Below we address this question for native sequences of real proteins. If the problem is under-determined, then the number of functions/constraints is less than the number of parameters and one can find an infinite number of solutions minimizing all functions. This is the case when the number of proteins in the database is small. As we have shown above, the potential derived for a few proteins provides an average energy gap greater than that provided by the true potential, but it shares no similarity with the true potential.

However, as the number of proteins in the database increases, an average gap approaches that for the true potential and the derived potential becomes very similar to the true one. To ensure that we do not have too many parameters we have devised a control procedure with randomly shuffled sequences.

Randomly shuffled sequences: an essential control

As a control we carried out the derivation procedure for our database of model proteins using shuffled sequences instead of the designed ones for each protein. In this case one should not expect that there exists any potential which makes all the native structures to be of low energy for randomly shuffled sequences, i.e. in this case our procedure should not lead to any meaningful solution. What happens in this case?

Again, for a few (1 . . . 5) proteins one can find a potential which provides a large enough energy gap ($|F| = 0.8 \ldots 1.2$) for randomly shuffled sequences (see Figure 4A). However, in contrast to the designed sequences, average energy gap drops substantially to a marginal level of $|F| = 0.2$ as the number of proteins in the database increases. Clearly, there is no correlation between the true potential and potential derived for a database with shuffled sequence (see Figure 5). There is also no correlation ($r = 0.0$) between potentials obtained for 100 proteins with shuffled sequences and potentials obtained for $N_{\text{prot}} < 100$ of these proteins. Hence, the procedure does not converge to any potential for proteins with randomly shuffled sequences.

Consequently, no pairwise potential can provide stability simultaneously to all of the native conformations with the shuffled sequences.

Comparison of the results for designed sequences with the control case of shuffled sequences suggests that the problem of finding a pairwise potential is not under-determined, i.e. 210 parameters of the potential are sufficient to provide a large gap for designed sequences and are not sufficient to provide a large gap for any pair of sequence and conformation.

Note that the procedure is able to distinguish between designed and randomly shuffled sequences without prior information about the potential used for the design. Designed sequences show the convergence of average energy gap $|F|$ to a level of $F = 1.4$ as the number of proteins increases. Potential is converging which is seen from high correlation between potentials obtained for different number of proteins in the dataset. In contrast, no convergence to a single potential is observed for proteins with randomly shuffled sequences (see Figure 5). The target function F approaches small values of 0.2 as the number of proteins in the data base increases. Where are the native proteins on this scale? Do they behave more like designed or like randomly shuffled sequences?

Native proteins

The model

We build a database of proteins with less than 25% of sequence homology, longer than 50 and shorter than 200 amino acids. The database contains 104 proteins listed below (in pdb-code names): 1hcr 1cad 1enh 1aap 1ovo 1fxd 1cse 1r69 1plf 2sn3 1bov 1mjc 1hst 1hyp 1ubq 4icb 1pk4 1poh 1aba 1lmb 1cyo 1brs 1fna 1mol 1stf 1gmp 1frd 1hsb 1ida 1plc 1aya 1onc 1sha 1fus 1psp 1fdd 256b 1acx 1bet 1fkb 1pal 2sic 1brn 2trx 1ccr 2msb 1dyn 1c2r 1etb 1gmf 2rsl 1paz 1rpg 1acf 2ccy 3chy 135l 1aiz 1rcb 1adl 1bbh 1slc 1eco 2end 4fxn 1ith 1cdl 1flp 2asr 1ilr 1lpe 1hbi 1bab 1lba 1mba 8atc 1ash 2fx2 2hbg 2mta 1f3g 1ndc 1aak 1cob 4i1b 1mbd 2rn2 1esl 1hfc 1hlb 1pnt 1hjr 4dfr 119l 3dfr 2cpl 5p2l 1rcf 9wga 2alp 1fha 1bbp 2gcr 1hbq. We use this database to derive the potential that maximizes the average energy gap $|\langle Z \rangle_{\text{harm}}|$. We define a contact between two amino acids as when the distance between their nearest heavy atoms is less than cutoff value 4.5 Å. In contrast to the lattice model, real proteins have a different length and a different number of contacts. These factors affect the value of Z-score. To account for the increase of Z with protein length we introduce the following normalization:

$$Z_{\text{norm}} = \frac{Z}{\sqrt{n_{\text{nat}}}}$$

where n_{nat} is the number of native contacts. Normalized values of Z_{norm} are used to compute $F_{\text{norm}} = \langle Z_{\text{norm}} \rangle_{\text{harm}}$ harmonic mean. Our criterion overemphasizes poor scores and therefore is very

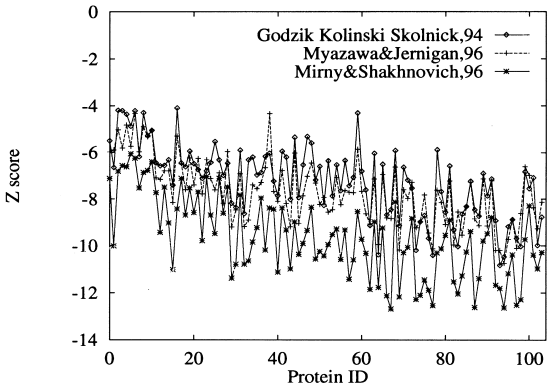

Figure 6. *Z*-score of native proteins with different potentials. Proteins are arranged in order of increasing length; this explains the systematic trend of decrease of *Z*-score as protein ID no. increases.

sensitive to proteins in the database that are more random-like; hence their presence in the dataset can distort the resulting potential. To avoid this difficulty we selected proteins for the dataset, which we believe are stabilized by similar physical forces (hydrophobic, electrostatic, H-bonds, etc.) and avoided proteins stabilized by other factors such as disulfides or coordinated metals, heme groups, etc.

Figure 6 presents values of *Z*-score obtained for the derived potential. Although our method finds the potential that maximizes the average energy gap ($|Z|$) simultaneously for all proteins, the values of the gap obtained for real proteins are rather small. This indicates that in the framework of the model we use (contact pairwise potential, 4.5 Å cutoff of residue–residue interactions etc) no pairwise potential can provide high stability simultaneously to all native proteins.

How good is the model for native proteins?

The potential derived from native proteins converges to a certain value of *Z*-scores. Is this value large or small? To answer this question we should compare this value with two limiting cases: (1) when the functional form of energy function is ''exact'', and sequences are well-designed for this energy function; and (2) with randomly shuffled sequences.

For each protein in the dataset of native proteins, we design a sequence using MJ potential as the true potential preserving amino acid composition of the native sequence. Then we derive a potential for a subset containing N_{prot} proteins from the database. The derivation is performed for the proteins built of: (1) the native structures with their native sequences, (2) the native structure with sequences designed for them; (3) and the native structures with randomly shuffled sequences.

Figure 7 presents average normalized energy gap values $|F_{norm}|$ obtained for all three sets of sequences as a function of the number of proteins in the database. Similar to the lattice model (see Figure 4A), designed sequences reach high values of the gap $|F| = 1.2$ whereas for random sequences $|F| = 0.2$. Derivation of potential for native sequences yields $|F| = 0.56$, which is considerably less than the gap provided for the same structures with designed sequences.

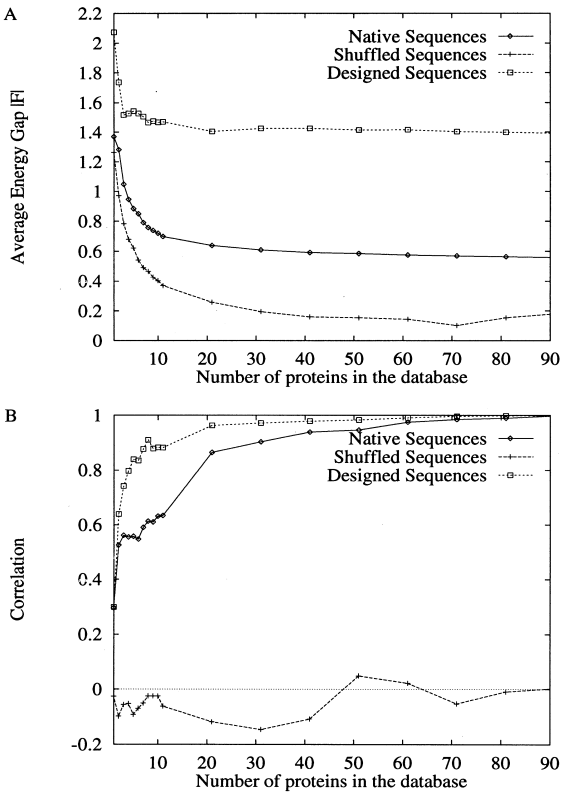

Figure 7. Native proteins. A, Effect of the database size on the energy gap; B, convergence test: correlation between the potential derived using smaller database and the potential derived using all 104 proteins in the database.

Another important property of designed sequences is that the derived potential converges to a single potential as the number of proteins in the database increases. Randomly shuffled sequences, in contrast, lack this convergence. The criterion of this convergence is the correlation between potentials derived using 100 proteins and potentials derived using $N_{prot} < 100$ proteins. The correlation between potentials obtained for the database of native proteins as a function of the number of proteins in the database is shown in Figure 7B.

In contrast to randomly shuffled sequences, native sequences as well as designed sequences provide convergence to a single potential as the number of proteins used for the derivation increases. Hence, we are able to find a potential which, for the given model, maximizes the energy gap for all native proteins simultaneously. This result clearly demonstrates that the model energy function used in this study of proteins is meaningful and reflects some essential interactions, but not all, since there is a pronounced difference

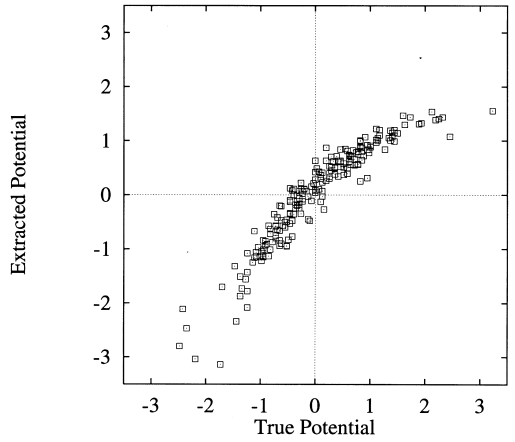

Figure 8. Derived potential *versus* true potential for the native proteins.

between F-values for designed and random sequences.

Since our method of derivation maximizes $|Z|$-scores for all proteins, no potential can provide a greater $|Z|$-score for the studied Hamiltonian (pairwise interaction potential) than this method. Our results demonstrate that very moderate $|Z|$-scores can be obtained using a pairwise potential for native proteins and no potential can increase values of $|Z|$ for them. However, other models utilizing different protein structure representations or different forms of potential can be more efficient in providing a large enough energy gap for native proteins. Using our procedure one can compare different models quantitatively and select the one which provides the largest energy gaps for native proteins.

Can we derive a potential from the dataset of poorly designed sequences?

Using the lattice model and native proteins, we demonstrated that our procedure is able to reconstruct sufficiently accurate true potential if sequences in the database are well designed. Is this requirement too restrictive? How well can we reconstruct potential for proteins with poorly designed sequences?

To mimic the poor design of native proteins observed in our model, for each protein structure, we design sequences that provide the same value of energy gap as the native sequence. Design is performed using MJ potential as the true one. Next we derive the potential for these poorly designed sequences and compare them with the true potential.

Figure 8 presents the scatter plot of interaction energies for the obtained potential plotted against the same values of the true potential. The result shows that our procedure reconstructs potential for poorly designed sequences very well providing a correlation of $r_{poor} = 0.91$ with the true potential. The poor design of native sequences in our model does not affect the quality of the reconstruction of the true potential.

Can poorly designed proteins fold?

As we have demonstrated above, native proteins are rather poorly designed in terms of the pairwise potential. The best possible pairwise potential provides a rather small energy gap to the native proteins, which is characterized by the typical value of $Z_{norm} = 0.5 \ldots 0.7$. Well designed proteins have, in contrast, $Z_{norm} = 1.2 \ldots 1.4$ and may be able to fold to their native conformation, as lattice model simulations suggest. The question is whether poorly designed sequences can fold as well.

To address this question we turn to the ''ideal'' lattice model and build a dataset of 200 poorly designed proteins. Proteins in this dataset are designed to have $Z_{norm} = 0.5 \ldots 0.7$. Then we derived potential for this dataset as described above and performed folding simulations for all sequences using true and derived potential. The result is that no protein was able to fold to its native conformation neither with derived nor with the true potential.

In all cases there was a conformation which has an energy below the energy of the native conformation. Hence: (1) the native conformation is not the global energy minimum for a poorly designed protein; (2) poorly designed proteins are unable to fold to their native conformations in *ab initio* folding simulations.

Fold recognition of poorly designed proteins

Sampling techniques that are more constrained to protein-like conformations (Finkelstein & Reva, 1991; Jones *et al*., 1992; Wodak & Rooman, 1993) can, however, recognize the native and native-like folds among a small enough pool of alternative conformations. The success of different pairwise potentials for the fold recognition shows that this sampling technique works quite well even for poorly designed proteins. Using our set of poorly designed sequences we perform fold recognition tests for all lattice model proteins by threading sequences of each protein through 200 alternative conformations. Only three out of 200 sequences recognize a non-native conformation as those of the lowest energy. This result is in contrast to previous observations that for every protein in this set native conformation is not the global energy minimum. Hence, the only reason why fold recognition works for 197 proteins is that a set of decoys is not too large and representative, so that the native conformation had the lowest energy.

Not surprisingly, the comparison of the results of *ab initio* folding simulations and fold recognition indicates that folding is a much more complicated problem than fold recognition, since a much larger energy gap is required for successful folding compared to fold recognition.

The question whether poorly designed proteins

Table 1. Comparison of different procedures for derivation of potential

Potential	$\langle Z \rangle$	Correlation with true potential	Fraction of proteins able to fold (%)
True potential	7.68	1.00	100
This work	8.61	0.83 (0.82)	96
Goldstein *et al*. (1992)	8.45	0.78 (0.71)	94
Hinds & Levitt (1994)	7.18	0.86 (0.84)	99
Miyazawa & Jernigan (1985)	7.09	0.75 (0.68)	95

Correlations are computed for potentials obtained using all 200 proteins. Correlations shown in brackets are for potentials obtained using 100 proteins.

can be used for recognition of the native fold in threading experiments has yet to be studied systematically.

Comparison with other potentials and techniques for extraction of potential

Several knowledge-based techniques for derivation of potentials from native protein structures have been suggested. It is important to compare our potential with other pairwise potentials, and our method of derivation with other methods. Figure 6 presents Z-scores computed for proteins of our database using our potential and two other potentials taken from the literature. Clearly our potential provides significantly lower values of Z-score for all proteins in the database. Two other potentials perform well as well as providing rather low Z-scores. Although potentials are obtained using different techniques, the overall profiles of Z-score for this dataset of proteins are very similar for all three potentials, i.e. when a protein has low Z-score with one potential it usually has low Z-score with another potential. High correlations between Z-score values provided by these three potentials for the same set of proteins ($r_{MJ,GKS} = 0.83$ $r_{GKS,MS} = 0.86$ and $r_{MJ,MS} = 0.83$) indicate that high or low value of Z-score is a property of a protein itself irrespective of potential used. Different proteins are known to have different stability, i.e. different quality of design, which is displayed by high or low values of Z-score.

Comparison with other techniques for extraction of potential

It is important to compare not only potentials themselves but also the techniques for derivation of potential. Our ideal lattice model is very useful for this purpose. We apply different techniques to the same set of lattice proteins and test obtained potentials in the same way as we did this for our technique.

Here we compare four techniques for derivation of potential. The first two are widely used statistical knowledge-based methods to derive energy of residue–residue and residue–solvent interactions. Knowledge-based techniques are reproduced following Miyazawa & Jernigan (1985, 1996) (MJ) and Hinds & Levitt (1994; HL). The third tested technique is the procedure suggested by Goldstein *et al*. (1992; GSW). This procedure is somewhat similar to our method since the potential is obtained to maximize ratio T_c/T_f, which is similar to the Z score we use. Goldstein *et al*. found analytic expression for potential which maximizes T_c/T_f for one protein. To find the potential for a set of proteins they used averaging, which is not justified but it yielded good results. We followed the procedure described by (Goldstein *et al*. (1992) to test their technique. Note that both the GSW procedure and ours are optimization techniques, whereas HL and MJ are statistical knowledge-based ones.

The results for different techniques are summarized in Table 1. Our potential is aimed to minimize harmonic mean Z score and, as expected, provides a lower value of $\langle Z \rangle_{harm}$ than other potentials. GSW procedure gives only slightly higher values of mean Z, which proves that both optimization techniques are powerful enough to provide a large energy gap for proteins of a dataset. Knowledge-based techniques provide a large energy gap as well. The drastic difference between knowledge-based techniques and optimization techniques becomes transparent when we compare Z-scores obtained for different derived potentials with Z-scores provided by the true potential (see Figure 9). Both optimization techniques provide Z-scores which are lower than Z for the true potentials. Knowledge-based techniques, in contrast, provide Z-scores higher than those for the true potential. Hence, knowledge-based potentials provide a smaller energy gap than the true potential does, whereas potentials obtained by optimization deliver an energy gap which is greater than those for the true potential. The decrease of energy gap by knowledge-based potentials can be crucial for *ab initio* folding, especially for weakly designed proteins which have rather small gap even with the true potential.

All tested techniques are also quite efficient in reconstruction of the true potential, exhibiting, however, different patterns of distortion of the original potential. Both optimization techniques tend to underestimate repulsive interactions (see Figure 1). Knowledge-based techniques, in contrast, provide good estimates of energies of repulsive interactions, suffering from underestimation of attractive interactions (see Figure 10). Attractive interactions are responsible for stabilization of the native conformation and underestimation of attrac-

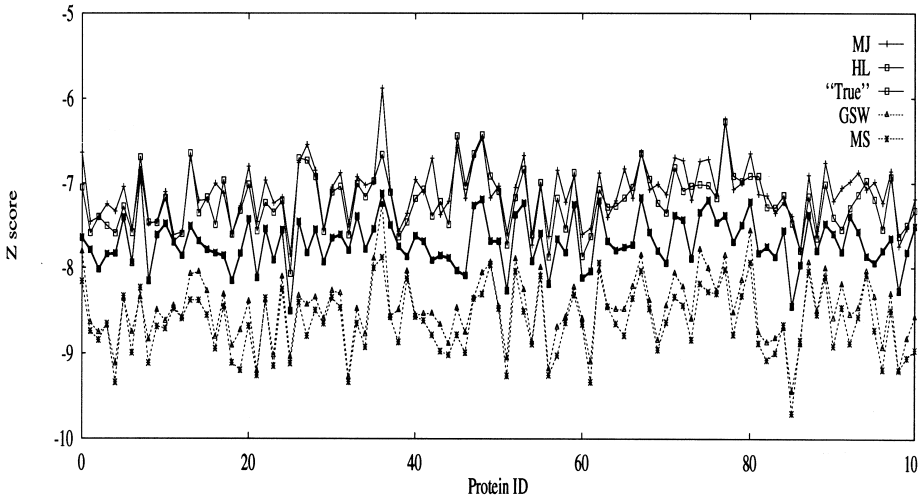

Figure 9. *Z*-score for 100 test lattice proteins with potentials derived by different techniques from 100 database lattice proteins. HL, Hinds & Levitt; MJ, Miyazawa & Jernigan, GSW, Goldstein *et al*.; MS, this work.

tive interactions leads to the observed (Figure 9) increase in *Z*-score for knowledge-based potentials.

Another deformation of the true potential by MJ technique is that it yields strong non-specific attraction between residues, which is seen as low negative average interaction between residues ($\langle U_{MJ} \rangle = -1.07$ when $\sigma(U)$ is set to 1). This non-specific attraction favors more compact conformation irrespective of amino acid sequence. This effect can mislead *ab initio* folding and fold recognition. The origin of this non-specific attraction is in residue–solvent interactions taken into account by MJ procedure. Estimate of the number of solvent–solvent interactions is responsible for the non-specific attraction.

Although all derivation procedures reconstruct the true potential with systematic deviations, all potentials are able to provide a large enough energy gap for well designed model sequences.

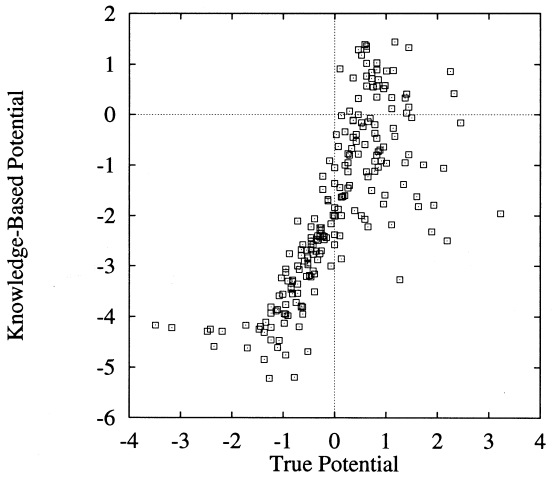

Figure 10. Potential obtained by statistical knowledge based technique *versus* true potential for the lattice model.

Discussion

In this work we proposed and tested a novel systematic approach to the long-standing problem of how to find the correct potential for protein folding.

In contrast to widely used knowledge-based statistical technique, which relies on hardly justifiable assumption of Boltzmann statistics, we use optimization in space of parameters to search for a potential which maximizes stability of all native proteins in the dataset.

The procedure was tested using the ideal model where sequences were designed with some known, true potential and the recovered potential turned out to be quite close to the true one. The key feature of ideal models (both lattice and off-lattice) is that the form of the energy function (two-body contact Hamiltonian) is ''exact'', and the goal of the parameter search is to determine 210 numbers, parameters of this Hamiltonian. We showed that our procedure recovers the parameters reliably and uniquely. It is important to note that it is not crucial for our method that sequences in the database are well-designed: in fact derivation of potentials using the database of weakly designed sequences (i.e. having relatively high *Z*-score) yielded potentials which were similarly quite close to the true potential. This is in contrast to the control case of assigning randomly shuffled sequences to structures: for them our procedure did not converge to any meaningful potential. In this case addition of any new ''protein'' (in fact a structure with a random sequence assigned to it) changed the potential dramatically, consistent with the notion that there is no potential which delivers low energy to all structure-random sequence pairs. In contrast, even for weakly designed sequences there is the potential for which these sequences have low (but perhaps not the lowest) energy in their corresponding native conformations, and such potential is readily recovered by our optimization technique.

The method has internal controls of self-consistency. First is that the optimization procedure in parameter space converges rapidly and at all algorithmic temperatures to a unique solution (no multiple-minima problem in space of parameters). This suggests that the obtained solution delivers a global minimum of Z-scores for studied proteins: no other potential can provide, on average, lower Z-scores for the same structures and sequences in the same model.

Another important test of self-consistency of the proposed method is convergence of potentials when the database size grows. This clearly points out that the problem is not under-determined as well as it indicates clearly that there indeed exists a potential with which all structures have low energy. This criterion is especially important and useful when we consider more complicated, than pairwise contact, energy functions (see below).

The ideal models provide an ideal opportunity to compare our new method with other approaches, in particular with the methods based on quasi-chemical approximation. Comparison of the true potential with the ones derived by the Miyazawa & Jernigan (1996) and Hinds & Levitt (1994) methods (including most demanding *ab initio* folding tests) shows that procedure based on the quasichemical approximation can extract potential with impressive accuracy. This conclusion is in contrast with the assertion of Thomas & Dill (1996), who also tested MJ procedure using lattice model and argued that the extracted potential is not an accurate approximation of the true potential. We believe that the most important criterion of success of extracted potential is how it performs in *ab initio* folding or threading tests. Thomas and Dill's test is similar in spirit to threading because they addressed the issue of how often the global energy minimum structure remains as such with extracted potentials, judged by exhaustive enumeration of conformations. They considered all sequences (having unique native state) for 14-mers and 16-mers with random sequences. However, the native conformations of random sequences (as well as other sequences having no or minimal energy gap) are extremely unstable with respect to any uncertainties in potentials (Bryngelson, 1993). This is in contrast with folding sequences (with energy gaps) which were shown to be much more robust with respect to uncertainties in potentials (Pande *et al*., 1995). Unfortunately, there are no sequences in HP model which have energy gaps (Thomas & Dill, 1996), and one cannot even design such sequences in HP model.

Therefore we believe that the major reason for the conclusion reached by Thomas & Dill (1996) is that they used the model where native conformation is unstable with respect to even minor uncertainties in potentials. It is important to note also that models where sequences do not have energy gaps are equally unstable with respect to point mutations (Shakhnovich & Gutin, 1991). The remarkable stability of proteins with respect to many point mutations is further strong evidence that real proteins should have a pronounced energy gap, a property absent in HP models.

Building a set of alternative conformations to compute Z-score is an important part of this work. All results discussed here have been obtained under the following assumptions regarding the presentation of alternative conformations: (1) all alternative conformations have the same compactness as the native conformation; (2) all contacts are equally probable; and (3) they are statistically independent in the set of alternative conformations. These assumptions allowed us to compute Z-score for a protein without building the set of alternative conformations explicitly. In fact, in order to compute the Z-score for pairwise potential one needs to calculate the average frequency of a contact and covariance of two contacts in the set of alternative conformations. The first assumption states that the number of contacts in alternative conformations is the same as in the native one. Assuming compactness of alternative conformations we eliminate the effect of non-specific attraction/repulsion in the recognition of the native conformation. Non-specific attraction introduced into potential favors most compact conformations irrespective of amino acid sequence, which can give rise to false positives: very compact low energy conformations for any protein in the fold recognition test. One should be careful about a non-specific term in a potential as it can substantially affect the results of *ab initio* folding or fold recognition. On the other hand, these non-specific terms can be readily eliminated by shifting the parameters by a given value (Shakhnovich, 1994; Gutin *et al*., 1995).

By assuming equal probability for all contacts we neglect the slight prevalence of contacts between amino acids close to each other along the polypeptide chain, which exists even in random coil. However, since alternative conformations have the same compactness as the native one local contacts are not expected to dominate in these conformations (Abkevich *et al*., 1995). Different probabilities of local and non-local contacts can be taken into account by assigning higher probabilities to local contacts in the set of unfolded conformations used in calculation of Z-score.

Our assumption of independent contacts is strictly valid only for point-size non-connected objects. Chain connectivity enforces positive correlation between contacts i, j and $i, j + l$ for small $l = 1, 2 \ldots$ On the other hand, excluded volume of amino acids leads to anti-correlation between contacts, i, j and i, k since amino acid i can have only a limited number of contacts due to excluded volume interactions. Several other factors can contribute to correlation between contacts in opposite ways and the final outcome of these effects has yet to be understood.

The set of alternative conformations built in this way turned out to be adequate for estimating the energy gap, in our model, since lattice proteins

which have low enough Z-scores are able to fold fast to their native conformations.

In general, while deriving a potential for a particular task and sampling procedure (fold recognition, design of an inhibitor, *ab initio* folding etc), one has to construct a set of alternative conformations which will be used as decoys during the sampling.

Alternative conformations used in this work correspond more closely (though not exactly) to sampling by folding under the condition of average attractive interaction between amino acids, while fold recognition is likely to have a different set of decoys. This set of decoys should be used in our procedure of derivation of potentials for fold recognition. It can be implemented by explicit generation of alternative conformations for a given protein by threading its sequence through other protein structures of the database. Frequency of contacts and contact correlations computed for alternative conformations built in this way are to be used for computing Z-scores and derivation of potential. While derived, the potential will provide the highest possible Z-score for fold recognition. This work is in progress.

Now we turn to the discussion of the results obtained for real proteins. First of all we see that pairwise contact approximation is not meaningless for real proteins, i.e. certain aspects of their energetics are captured by that model. The clear evidence for that is that our procedure converges to a unique potential, and the Z-scores of proteins with that potential are considerably lower than for randomly shuffled sequences. This suggests that such a simplified Hamiltonian still carries some ''signal''. In this sense the derived two-body potentials are useful since they are able to discriminate between native conformation and decoys, when the number of decoys is not too large.

However, the Z-scores obtained for proteins within the pairwise contact Hamiltonian approximation are not sufficiently low to provide high stability (or large energy gap) for all proteins simultaneously. Hence all knowledge-based potentials can have only limited success in folding or recognition of the native fold among alternative conformations. This result can help in the understanding of the origin of problems arising with various structure prediction techniques. Our results suggest that limited success in folding simulations in the simple model with pairwise potentials may be due not to incorrect potentials (i.e. 210 numbers) but rather due to the deficiency of the model itself, and no other potentials within the same model of pairwise contact interactions can provide better results uniformly for numerous tested proteins (of course there can be successes with potential which are optimized to fold just one protein; Hao & Sheraga, 1996a,b); however, as our analysis shows, such potential (speaking in our terms, derived by optimization from the database of one protein) will fail when used to fold another protein.

Several models have been suggested for protein folding which vary in accuracy of structure representation and in complexity of the energy function. What is the optimal number of parameters of the energy function? How does the number of parameters affect the results of the procedure to extract potentials? To address these questions we developed a convergence test which allows us to estimate stability of the obtained potential with respect to the dataset of proteins used. This test indicates whether the number of parameters used to maximize the energy gap (210 in the case of contact pairwise potential) is large enough to provide the gap for all protein simultaneously and is small enough not to over-fit the data and adopt any random sequence to a protein structure in the database. Our results indicate that 210 parameters of contact pairwise potential are not too many (potential converges as the size of the database increases), but the model itself is not sufficiently realistic to provide the large gap for real proteins. More accurate presentation of energy function (possibly including local conformational preferences, distance dependent interactions, multibody interactions etc.) is likely to be necessary to achieve better discrimination between the native structure and decoys. The presented method allows us to assess systematically the validity of different models and therefore can serve as a powerful tool in the search for the most adequate model for protein folding.

Methods

Derivation of potential

Energy function assigns a value of the energy to a given conformation for a given amino acid sequence:

$$E = E(\mathrm{Sequence}, \mathrm{Conformation}, \mathbf{U}) \qquad (1)$$

where $\mathbf{U}$ is the set of parameters of potential to be derived from known native protein structures.

We use the Z-score as a measure of how pronounced is the energy minimum corresponding to the native conformations (with respect to a set of alternative conformations; Bowie *et al*., 1991):

$$Z = \frac{E_N - \langle E \rangle_{\mathrm{conf}}}{\sigma_{\mathrm{conf}}(E)} \qquad (2)$$

This is the deviation of the energy of the native conformation from the average energy of alternative conformations measured in units of standard deviation. The average energy $\langle E \rangle$ and variance $\sigma(E)$ are computed for a set of alternative conformations (see below). The absolute value of the Z-score is the natural measure of the energy gap.

Our goal is to find a potential $\mathbf{U}$ that minimizes Z-scores (maximizes the energy gap) simultaneously for all proteins in the dataset. This is achieved by building a target function which is an appropriate combination of individual Z scores and then optimizing this function with respect to $\mathbf{U}$. One should carefully choose a combination of Z-scores to optimize. If the target function to be optimized is naively chosen, for example a sum of Z-scores, then low values of the target function can be obtained if Z is small enough for some proteins and large

for all others. To avoid this kind of a problem, one has to minimize $max_m(Z^{(m)})$, which is, however, very difficult to deal with because of its discontinuity. We chose a harmonic mean of $Z^{(m)}$ scores as a function to be minimized:

$$\langle Z\rangle_{\text{harm}} = \frac{M}{\sum_{m=1}^{M} 1/Z_m} \tag{3}$$

A harmonic mean is a smooth approximation of $max_m(Z^{(m)})$ since terms with the smallest absolute value of $Z^{(m)}$ scores contribute most to the harmonic mean.

To maximize the energy gap for all proteins in a dataset, we search for a potential **U** which maximizes the value of function $F(\mathbf{U}) = -\langle Z\rangle_{\text{harm}}$. The value F is directly related to the energy gap. Hence below we refer to F as an energy gap, with the understanding that it is not exactly identical to it, but that these two quantities have a monotonic interdependence.

We also apply some constraints to the potential **U**:

$$\langle \mathbf{U}\rangle = 0 \tag{4}$$

$$\sigma^2(\mathbf{U}) = \langle(\langle \mathbf{U}\rangle - \mathbf{U})^2\rangle = 1 \tag{5}$$

The first constraint sets an average interaction between amino acids to zero, i.e. eliminates non-specific attraction/repulsion between amino acids. Non-specific attraction/repulsion favors more/less compact conformations irrespective the amino acid sequence. We use the first constraint to avoid this kind of bias.

The second constraint sets the dispersion of interaction energies to one. If energy is a linear function of parameters, multiplication of **U** by an arbitrary constant does not change the values of Z-score. By setting $\sigma(\mathbf{U}) = 1$, we choose units of energy.

Potential **U** is obtained by maximizing of $F(\mathbf{U})$ using a procedure for non-linear optimization. The potential obtained in this way is (by procedure) the one which provides the largest energy gaps simultaneously to all proteins in the dataset as far as $\langle Z\rangle_{\text{harm}}$ is an accurate approximation of $max_m(Z^{(m)})$.

One of the most important parts of the method is the set of alternative conformations used to compute Z-scores. In general, one has to use the same set of conformations for sampling and for computing Z-scores. For example, to optimize the potential for threading, one has to compute Z-scores using a set of alternative conformations obtained by threading a sequence through a representative set of protein structures. To find a potential, one needs to generate a set of alternative conformations, and then, use this set to compute individual Z-scores. This procedure is computationally inexpensive since the set of conformations obtained by threading does not depend on the potentials used. However, when dynamic sampling techniques (Monte Carlo, molecular dynamics, growth procedures etc.), are used for *ab initio* folding, the set of alternative conformations is not known in advance and, more importantly, depends on the potential applied. In this case one has to make some assumptions about the ensemble of alternative conformations that will enable one to compute average energy $\langle E\rangle_{\text{conf}}$ and variance of energy $\sigma_{\text{conf}}(E)$ over the ensemble of alternative conformations. In this study we show how to optimize pairwise potential for *ab initio* folding of a simple model and for threading.

Derivation of parameters for pairwise potential

Pairwise potential

We consider pairwise contact potential, i.e. the energy of a conformation is a sum of the energies of pairwise contacts between monomers, which are not nearest neighbors in sequence:

$$E(\boldsymbol{\xi}, \boldsymbol{\Delta}) = \sum_{1\leqslant i<j\leqslant N} U(\xi_i, \xi_j)\Delta_{ij} \tag{6}$$

where $\Delta_{ij} = 1$ if monomers i and j are in contact and $\Delta_{ij} = 0$ otherwise. Various definitions of contacts can be used (Kocher *et al*., 1994). ξ_i defines the type of amino acid residue in position i. Potential is given by **U** matrix, where $U(\xi, \eta)$ is energy of a contact between amino acids of types ξ and η.

Optimization of potential

The optimization of $F(\mathbf{U}) = -\langle Z\rangle_{\text{harm}}$ is performed by the Metropolis Monte Carlo procedure in space of potentials, i.e. at each step a cell $U(\xi, \eta)$ of the matrix **U** is chosen randomly and a small random number $r \in[-0.1, 0.1]$ is added to $U(\xi, \eta)$. This change is accepted if it increases $F(\mathbf{U})$ and rejected with probability:

$$1 - \exp(-\frac{\delta F}{T_{\text{opt}}})$$

if it decreases $F(\mathbf{U})$. T_{opt} is the temperature of optimization. Optimization of potentials starts from a completely random potential and stops when the target function changes less than on $\epsilon = 0.01$ for the last 20,000 steps.

Computation of Z-score

For a given sequence ξ, potential **U** and generated set of alternative conformations $\boldsymbol{\Delta}^{(k)}$, $k = 1, \ldots K$ one can compute Z of the native conformation $\boldsymbol{\Delta}^N$:

$$Z(\boldsymbol{\xi}, \boldsymbol{\Delta}^N) = \frac{E(\boldsymbol{\xi}, \boldsymbol{\Delta}^N) - \langle E(\boldsymbol{\xi}, \boldsymbol{\Delta}^{(k)})\rangle_k}{\sigma(E(\boldsymbol{\xi}, \boldsymbol{\Delta}^{(k)}))_k} \tag{7}$$

where index k denotes averaging over alternative conformations.

Instead of computing the energy of a sequence in each alternative conformation whenever we need Z, we compute the average quantities for the set of conformations and use these averages to compute Z-scores for any sequence and any potential. These average quantities are computed only once, which saves a significant amount of computer time.

The energy of an individual conformation for a pairwise Hamiltonian is given by equation (6). Hence, the average energy for the set of conformation is:

$$\langle E(\boldsymbol{\xi}, \boldsymbol{\Delta}^{(k)})\rangle_k = \sum_{1\leqslant i<j\leqslant N} U(\xi_i, \xi_j)\langle\Delta_{ij}\rangle_k \tag{8}$$

where $\langle\Delta_{ij}\rangle$ is the average density of contacts between residues number i and j in the set of alternative conformations:

$$\langle\Delta_{ij}\rangle_k = \frac{1}{K}\sum_{k=1}^{K}\Delta_{ij}^{(k)} \tag{9}$$

Note that one can compute the matrix of average density of a contact $\langle \Delta_{ij} \rangle$ only once for a set of conformations and later use this matrix to compute the average energy for a sequence ξ and any potential **U** (see equation (8)).

Similarly for $\sigma(E)$,

$$\sigma^2(E(\xi, \Delta^{(k)}))_k = \langle E^2 \rangle_k - \langle E \rangle_k^2 = \sum_{1 \leqslant i < j \leqslant N} \sum_{1 \leqslant l < m \leqslant N} U(\xi_i, \xi_j) U(\xi_l, \xi_m) T_{ij,lm} \tag{10}$$

where $T_{i,j,l,m}$ is a contact correlator,

$$T_{ij,lm} = \langle \Delta_{ij}^{(k)} \Delta_{lm}^{(k)} \rangle_k - \langle \Delta_{ij}^{(k)} \rangle_k \langle \Delta_{lm}^{(k)} \rangle_k \tag{11}$$

which depends only on the set conformations and can be computed in advance for a given set. Once $\langle \Delta_{ij} \rangle$ and $T_{ij,lm}$ are computed, one can easily compute the value of Z-score for a given sequence ξ, conformation Δ^N and potential **U**:

$$Z(\xi, \Delta^N, \mathbf{U}) = \frac{\sum_{1 \leqslant i < j \leqslant N} U(\xi_i, \xi_j)(\Delta_{ij}^N - \langle \Delta_{ij} \rangle_k)}{\sqrt{\sum_{1 \leqslant i < j \leqslant N} \sum_{1 \leqslant l < m \leqslant N} U(\xi_i, \xi_j) U(\xi_l, \xi_m) T_{ij,lm}}} \tag{12}$$

Alternative conformations

For each protein in the dataset, we build an ensemble of alternative conformations which contains conformations with the same compactness as the native one, i.e. the same number of residue-residue contacts. In fact, instead of generating a huge number of conformations, we assume that: (1) the contacts in the alternative conformations are distributed independently and uniformly and that (2) the number of contacts is the same as in the native conformation. These assumptions allow one to compute the average density of contacts $\langle \Delta_{ij} \rangle$ and correlator $T_{ij,lm}$ as:

$$\langle \Delta_{ij} \rangle = \frac{n}{n_{\text{total}}} \tag{13}$$

and

$$T_{ij,lm} = \begin{cases} \dfrac{1}{n_{\text{total}}^2} & \text{if } ij \neq lm \\ \dfrac{1}{n_{\text{total}}} - \dfrac{1}{n_{\text{total}}^2} & \text{if } ij = lm \end{cases} \tag{14}$$

where n is the number of contacts in the native conformation, and n_{total} is the total number of topologically possible contacts.

Following this, the value of Z score can be computed for each protein in the dataset and a given potential **U** using equation (12). Lattice model simulations show that sequences with low values of Z are able to fold fast to their native conformations (Abkevich *et al*., 1994; Gutin *et al*., 1995).

To test our derivation method and compare it with other techniques, we first turn to a simple lattice model which allows us to test a potential by performing *ab initio* folding of protein starting from a random conformation. Next, we apply our method to derive the parameters from the dataset of well-resolved protein structures.

Sequence design

The aim of our sequence design is to find a sequence (for a given potential) that delivers a low Z-score to a given conformation. The procedure starts from a random sequence with a given amino acid composition. Although different sequences have different amino acid composition, the composition in the dataset corresponds to those of the native proteins (Creighton, 1993).

At each step we choose two residues at random and attempt to permute them. A change of Z-score (δZ) associated with this permutation is computed. If this permutation decreases the value of Z-score ($\delta Z < 0$), then this permutation is accepted, otherwise ($\delta Z > 0$) the permutation is rejected with probability:

$$1 - \exp\left(-\frac{\delta Z}{T_{sel}}\right)$$

The procedure stops when either no changes in sequence have occurred in the last 1000 $\dot{N}$ steps or if a preset value of Z-score is reached (Z_{target}). Using this procedure and setting different values Z_{target}, we are able to generate sequences that provide the required value of Z-score for a given conformation.

Acknowledgements

We are grateful to Alexander Gutin and Victor Abkevich for many fruitful discussions. L.M. is grateful to Michael Schwarz for interesting discussions. We thank Richard Goldstein for explaining to us their averaging procedure. This work was funded by the Packard Foundation.

The parameter set is available from our anonymous ftp-site paradox.harvard.edu; file Euv.dat in the directory/pub/leonid.

References

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Free energy landscape for protein folding kinetics, intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062.

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995). Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **252**, 460–471.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164–170.

Bryant, S. H. & Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins: Struct. Funct. Genet.* **16**, 92–112.

Bryngelson, J. D. (1994). When is a potential accurate enough for structure prediction? *J. Chem. Phys.* **100**, 6038–6045.

Covell, D. G. (1994). Low resolution models of polypeptide chain collapse. *J. Mol. Biol.* **25**, 1032–1043.

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*, W. H. Freeman and Co, New York.

DeWitte, R. & Shakhnovich, E. (1994). Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Protein Sci.* **3**, 1570–1581.

Elofson, A. Le Grand, S. & Eisenberg, D. (1995). Local moves: an efficient algorithm for simulation of protein folding. *Proteins: Struct. Funct. Genet.* **23**, 73–82.

Finkelstein, A. V. & Reva, B. A. (1991). Search for the most stable folds of protein chains. *Nature,* **351**, 497–499.

Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Ya. (1993). Why are the same protein folds used to perform different functions? *FEBS Letters,* **325**, 23–28.

Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Ya. (1995). Why do protein architectures have Boltzmann-like statistics? *Proteins: Struct. Funct. Genet.* **23**, 142–149.

Godzik, A., Kolinski, A. & Skolnick, J. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* **4**, 2101–2117.

Goldstein, R., Luthey-Schulten, Z. A. & Wolynes, P. (1992). Optimal protein-folding codes from spin-glass theory. *Proc. Natl Acad. Sci. USA,* **89**, 4918–4922.

Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995). Evolution-like selection of fast-folding model proteins. *Proc. Natl Acad. Sci. USA,* **92**, 1282–1286.

Hao, M. & Scheraga, H. (1996a). How optimization of potential function affects protein folding. *Proc. Natl Acad. Sci. USA,* **93**, 4984–4989.

Hao, M.-H. & Scheraga, H. A. (1996b). Optimizing potential function for protein folding. *J. Phys. Chem.* **100**, 14540–14548.

Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668–682.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86–89.

Kocher, J. P., Rooman, M. J. & Wodak, S. J. (1994). Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* **235**, 1598–1613.

Kolaskar, A. S. & Prashanth, D. (1979). Empirical torsional potential functions from protein structure data. *Int. J. Pept. Protein Res.* **14**, 88–98.

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

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

Lattman, E. E. (1995). Protein structure prediction. *Proteins: Struct. Funct. Genet.* **23**, 295–462.

Levitt, M. (1976). A simplified representation of protein conformation for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59–107.

Lemer, C., Rooman, M. & Wodak, S. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 321–332.

Maiorov, V. N. & Crippen, G. M. (1992). Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* **227**, 876–888.

Mirny, L. & Domany, E. (1996). Protein fold recognition and dynamics in the space of contact maps. *Proteins: Struct. Funct. Genet.* In the press.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* **18**, 534–552.

Miyazawa, S. & Jernigan, R. (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.

Nishikawa, K. & Matsuo, Y. (1993). Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.* **6**, 811–820.

Pande, V., Grosberg, A. & Tanaka, T. (1995). How accurate must potentials be for successful modeling of protein folding? *J. Chem. Phys.* **103**, 9482–9491.

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

Rooman, M., Kocher, J. A. & Wodak, S. (1992). Extracting information on folding from aminoacid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry,* **32**, 10226–10238.

Sali, A. Shakhnovich, E. I. & Karplus, M. (1994a). Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235**, 1614–1636.

Shakhnovich, E. I. & Gutin, A. M. (1991). Influence of point mutations on protein structure. Probability of a neutral mutation. *J. Theoret. Biol.* **149**, 537–546.

Shakhnovich, E. I., Farztdinov, G. M., Gutin, A. M. & Karplus, M. (1991). Protein folding bottle-necks: a lattice Monte Carlo simulation. *Phys. Rev. Letters,* **67**, 1665–1667.

Shakhnovich, E. I. & Gutin, A. M. (1993a). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA,* **90**, 7195–7198.

Shakhnovich, E. I. & Gutin, A. M. (1993b). A novel approach to design of stable proteins. *Protein Eng.* **6**, 793–800.

Shakhnovich, E. I. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Letters* **72**, 3907–3910.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883.

Skolnick, J. & Kolinski, A. (1990). Simulations of the folding of a globular protein. *Science,* **250**, 1121–1125.

Socci, N. D. & Onuchic, J. N. (1994). Folding kinetics of protein-like heteropolymers. *J. Chem. Phys.* **101**, 1519–1528.

Tanaka, S. & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules,* **9**, 945–950.

Thomas, P. & Dill, K. (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**, 1531–1548.

Vasques, M., Nemethy, G. & Scheraga, H. (1994).

Conformational energy calculations on polypeptides and proteins. *Chem. Rev.* **94**, 2183–2239.
Wilson, C. & Doniach, S. (1989). Computer model to dynamically simulate protein folding: studies with crambin. *Proteins: Struct. Funct. Genet.* **6**, 193–209.
Wodak, S. & Rooman, M. (1993). Generating and testing protein folds. *Curr. Opin. Struct. Biol.* **3**, 247–259.
Yue, K., Fiedig, K., Thomas, P. Chan, H. S., Shakhnovich, E. I. & K. A. (1995). A test of lattice protein folding algorithms. *Proc. Natl Acad. Sci. USA,* **92**, 325–329.

Edited by F. E. Cohen

(Received 15 July 1996; accepted 18 September 1996)