PROTEINS: Structure, Function, and Genetics 41:157–163 (2000)

How to Generate Improved Potentials for Protein Tertiary Structure Prediction: A Lattice Model Study

Ting-Lan Chiu[1] **and Richard A. Goldstein**[1,2]*
[1]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*
[2]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **Success in the protein structure prediction problem relies heavily on the choice of an appropriate potential function. One approach toward extracting these potentials from a database of known protein structures is to maximize the *Z*-score of the database proteins, which represents the ability of the potential to discriminate correct from random conformations. These optimization methods model the entire distribution of alternative structures, reducing their ability to concentrate on the lowest energy structures most competitive with the native state and resulting in an unfortunate tendency to underestimate the repulsive interactions. This leads to reduced accuracy and predictive ability. Using a lattice model, we demonstrate how we can weight the distribution to suppress the contributions of the high-energy conformations to the *Z*-score calculation. The result is a potential that is more accurate and more likely to yield correct predictions than other *Z*-score optimization methods as well as potentials of mean force. Proteins 2000;41:157–163.** © **2000 Wiley-Liss, Inc.**

Key words: contact potential; lattice proteins; fold recognition; protein folding; *Z*-score

INTRODUCTION

Methods to predict the tertiary structure of a target protein generally have three parts: there must be a way of defining a space of possible structures, a search strategy for exploring this space, and a potential energy function that is used to evaluate how well the protein sequence "fits" into any of the possible structures. The choice of a potential energy function is critical to the success of such procedures. In the absence of methods for generating sufficiently accurate potentials using ab initio quantum mechanical calculations or parameters drawn from small molecules, most of these potential functions are derived from the database of known protein structures using one of two approaches. The first approach is to calculate so-called "potentials of mean force" (POMF)[1–3] based on the quasi-chemical approximation of Miyazawa and Jernigan.[4] An alternative method to deriving potential functions is through optimization, generally by ensuring that the energy of the correctly folded structure of a set of training proteins is as low as possible compared with those of the incorrect structures.[5–12] The ability of the potentials to predict the structure of novel proteins is measured by evaluation the accuracy of the predictions made on an independent set of test proteins. Implicit in this approach is that it is necessary not only to stabilize the correct structures, but also to destabilize incorrect ones.

Goldstein, Luthy-Shulten, and Wolynes (GLW) approached this problem using both techniques drawn from spin-glass theory and from Bayesian statistics.[7,8,13,14] According to their work, the important quantity was the difference in the energy of the correct native state compared to the average energy of the random conformations (Δ), divided by the standard deviation of the random-conformation energies (Γ), similar to what has been called a *Z*-score in the sequence-alignment literature.[15] The best energy potential would be the potential that maximized this *Z*-score averaged over the proteins in the training database. A number of different implementations of this approach have been described, including maximization of $\langle\Delta_m\rangle/\langle\Gamma_m\rangle$ as first proposed by GLW[7,8,13,14] and maximization of $\langle 1/Z_m\rangle^{-1}$, as suggested by Mirny and Shakhnovich (MS).[10] We (CG) developed an expression to directly quantify the probability of success in predicting protein structures as a function of the *Z*-score.[16] We then maximized the average probability of success over the ensemble of training-set proteins.

Although performance in true blind tests with real proteins is the most definitive proof of a method, the small number of proteins in these tests make the results necessarily anecdotal. In addition, the tests with true proteins measures the success of the overall prediction package including the search over protein structures and the parametrization of the energy function, making it difficult to compare interaction derivation methods directly. This is especially problematic as we do not know the true interaction energies that are being approximated. One approach to this problem using lattice models was introduced by Thomas and Dill.[17] They created a model of proteins as self-avoiding walks on a cubic lattice, specified an energy

Grant sponsor: College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies; Grant sponsor: National Institutes of Health; Grant number: LM0577; Grant sponsor: National Science Foundation; Grant number: BIR9512955.

Ting-Lan Chiu's present address is Laboratory of Computational Genomics, Donald Danforth Plant Science Center, 893 N. Warson Road, St. Louis, MO 63141.

*Correspondence to: Richard A. Goldstein, Department of Chemistry, Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@umich.edu

Received 5 November 1999; Accepted 11 May 2000

function, and then created a lattice-protein database by finding the ground state of a set of amino acid sequences. The different methods for deriving interaction parameters were then applied to this database, and the accuracy of various methods compared. Using this approach, we can implement the same search strategy, the same parametrization of the energy function, and create a model that fulfills the approximations and assumptions of all of the various models. In this way, we can directly compare the different methods for deriving potentials, independently of these other factors.

In our previous article, we used this lattice-model method to demonstrate that our average probability of success method generates the most accurate potentials of the various *Z*-score based optimization approaches, and that the potentials calculated using the CG method is significantly more likely to be successful at predicting the structures of nondatabase proteins.[16] Even so, in further work we found that the POMF approach generates slightly more accurate potentials with a marginally higher success rate at predicting the correct structure for an independent test set. One factor was the tendency of the CG method, like other *Z*-based optimization schemes, to underestimate strongly repulsive interactions,[16] as noted by a number of investigators.[10,18] In addition, the *Z*-score is a function of the distribution of energies of the entire ensemble of alternative structures. It is not clear how relevant the very high energy alternative structures are to the prediction method. In this paper, we again turn to lattice models to show how this underestimation of the repulsive interactions is an inherent aspect of the Gaussian approximation used to model the distribution of energies of the random conformations. We improve our previous approach of maximizing the average probability of success by incorporating a refined *Z*-score calculation procedure that concentrates on the low-energy alternative structures. We find in tests with lattice proteins that this new approach generates potentials that are more accurate than those generated by previous *Z*-score based optimization methods as well as the POMF approach, even when the most questionable assumptions of the POMF method are satisfied by the model. We further show that, under these conditions, our method is significantly more likely to be successful at predicting the structures of an independent set of test proteins.

METHODS

Z-Score Based Optimization Methods

In all energy-based prediction schemes, we calculate the energy of each target sequence m in each of an ensemble of possible conformations $\{\mathscr{C}_k\}$ consisting of a set of random conformations $\{\mathscr{C}_r\}$ as well as the native-state (NS) conformation $\mathscr{C}_{\mathrm{NS}}$. Of the set of resulting energies $\{E_k^m\}$, the conformation of lowest energy is the predicted structure. The prediction is a success if the lowest energy conformation is the true native state, that is, all of the random conformations are higher in energy.

Z-score based optimization methods are based on maximizing the ability of the potential to discriminate between the correct and various incorrect structures. Similarly to the method used in sequence-alignment statistics, we calculate the *Z*-score by comparing the difference between the native state energy and the average of the distribution of energies of the other possible conformations with the width of this distribution. Mathematically, the *Z*-score for sequence m, Z_m, is equal to $Z_m = \Delta_m/\Gamma_m$ where $\Delta_m = \bar{E}_r^m - E_{\mathrm{NS}}^m$, with E_{NS}^m is the energy of sequence m in its native state, $\bar{E}_r^m$ is the average of the energies of the random conformations, and Γ_m is the width of the distribution of energies of these random conformations. If the *Z*-score is large, it means that the energy of the native state is well-separated from the energies of the random conformations, so predicting the correct native state can be done with confidence. The smaller the *Z*-score, the more likely it is that some random conformation will have an energy lower than that of the native state when calculated with the model potential, resulting in an incorrect prediction. We can use this approach to develop a potential from a training set of proteins with their associated structures by finding the potential that maximizes the *Z*-score of the proteins in the training set.

Maximizing the Probability of Success

In contrast to earlier methods that maximized $\langle\Delta_m\rangle/\langle\Gamma_m\rangle$ or $\langle 1/Z_m\rangle^{-1}$ where the averages are over the set of training proteins, we developed a potential to maximize the average probability of making a correct prediction. Let us assume that $\rho_m(E_r)$, the distribution of energies of protein m in the random conformations, is a Gaussian centered at $\bar{E}_r^m$ with standard deviation Γ_m. For us to be successful in correctly predicting the structure from among the N incorrect alternatives, *all* of the other structures have to have energy greater than E_{NS}^m. We assume that the energies of the incorrect alternatives are randomly and independently chosen from $\rho_m(E_r)$. $P(\mathscr{S}_m)$, the probability of "success" for sequence m, is given by[16]

$$P(\mathscr{S}_m) = \left(0.5 + 0.5\ \mathrm{erf}\left(\frac{Z_m}{\sqrt{2}}\right)\right)^N \qquad (1)$$

For an ensemble of proteins, we were interested in generating the largest possible number of correct predictions. As the total number of correct predictions is equal to the number of attempts times the average probability of success, we maximized $\langle P(\mathscr{S}_m)\rangle = \langle(0.5 + 0.5\ \mathrm{erf}(Z_m/\sqrt{2}))^N\rangle$, the probability that the energy function would yield a correct prediction averaged over the proteins in the training set.

New CG Method

As mentioned in the introduction, all *Z*-based schemes (including ours) tend to underestimate repulsive interactions. We can see why this happens by considering under what circumstances Z_m is increased. Maximization of the *Z*-score occurs when Δ_m is maximized and Γ_m is reduced. Γ_m is decreased when the energy of the highest-energy conformations are reduced. This results in a tendency of the optimization procedure to underestimate the repulsive interactions found more often in these high-energy states.

More generally, these Z-based schemes model the entire ensemble of random states. In fact, the challenge is to distinguish between the native state and other low-lying states.

We can reduce the systematic bias to a considerable amount and concentrate on the low-lying states by suppressing the high-energy states in calculating the Z-scores. We do it by weighting the contribution of random conformation r with energy E_r by $\exp(-\alpha E_r)$, where α is a positive real number, and then model the resulting data by Gaussian $\times \exp(-\alpha E)$. The parameters Δ_m and Γ_m can thus be extracted for calculating the Z-score, as follows. Following the weighting, the average of the resulting data, $\bar{e}_r^m$, is given by

$$\bar{e}_r^m = \frac{\sum_{r=1}^{N} E_r \exp(-\alpha E_r)}{\sum_{r=1}^{N} \exp(-\alpha E_r)} \tag{2}$$

with $\bar{e}_r^{2^m}$ similarly given by

$$\bar{e}_r^{2^m} = \frac{\sum_{r=1}^{N} E_r^2 \exp(-\alpha E_r)}{\sum_{r=1}^{N} \exp(-\alpha E_r)} \tag{3}$$

Modeling the resulting data by $\rho_m(E_r) \times \exp(-\alpha E_r)$, $\mathrm{Est}(\bar{e}_r^m)$ and $\mathrm{Est}(\bar{e}_r^{2^m})$, the expected values of $\bar{e}_r^m$ and $\bar{e}_r^{2^m}$, are given by

$$\mathrm{Est}(\bar{e}_r^m) = \frac{\int_{-\infty}^{\infty} E_r \exp(-\alpha E_r)\exp\left(-\frac{(E_r - \bar{E}_r^m)^2}{2\Gamma_m^2}\right) dE_r}{\int_{-\infty}^{\infty} \exp(-\alpha E_r)\exp\left(-\frac{(E_r - \bar{E}_r^m)^2}{2\Gamma_m^2}\right) dE_r}$$

$$= \bar{E}_r^m - \alpha\Gamma_m^2 \tag{4}$$

and

$$\mathrm{Est}(\bar{e}_r^{2^m}) = \frac{\int_{-\infty}^{\infty} E_r^2 \exp(-\alpha E_r)\exp\left(-\frac{(E_r - \bar{E}_r^m)^2}{2\Gamma_m^2}\right) dE_r}{\int_{-\infty}^{\infty} \exp(-\alpha E_r)\exp\left(-\frac{(E_r - \bar{E}_r^m)^2}{2\Gamma_m^2}\right) dE_r}$$

$$= \Gamma_m^2 + (\bar{E}_r^m - \alpha\Gamma_m^2)^2. \tag{5}$$

or $\Gamma_m^2 = \mathrm{Est}(\bar{e}_r^{2^m}) - (\mathrm{Est}(\bar{e}_r^m))^2$ and $\Delta_m = \mathrm{Est}(\bar{e}_r^m) + \alpha\Gamma_m^2 - E_{\mathrm{NS}}^m$. Equating calculated and estimated values of $\bar{e}_r^m$ and $\bar{e}_r^{2^m}$ leads to

$$\Gamma_m^2 = \bar{e}_r^{2^m} - (\bar{e}_r^m)^2 \tag{6}$$

and

$$\Delta_m = \bar{e}_r^m + \alpha\Gamma_m^2 - E_{\mathrm{NS}}^m. \tag{7}$$

We can obtain Δ_m and Γ_m using Eqs. (6) and (7), calculate the value of $Z_m = \Delta_m/\Gamma_m$, and compute $P(\mathcal{S}_m)$ using Eq. (1). We then can adjust the potential to maximize $\langle P(\mathcal{S}_m)\rangle$. Note that the new CG method is equivalent to the older CG method when $\alpha = 0$.

Potentials of Mean Force

The most common approach toward extracting interaction potentials from a training dataset is the potential of mean force method. According to this approach, the contact potential between residue types a and b separated by k peptide units along the chain is given by

$$\Delta E_k^{ab}(s) = -kT \ln\left[\frac{f_k^{ab}(s)}{f_k(s)}\right] \tag{8}$$

where s is the spatial distance between residues a and b; $f_k^{ab}(s)$ and $f_k(s)$ denote the relative frequencies of ab contacts and all amino acid pair contacts observed in the database for given values of k and s, respectively.[3,19–21] $f_k^{ab}(s)$ is equal to n_s^{ab}, the number of ab pairs at a distance s, divided by the total number of ab pairs, n_T^{ab}; $f_k(s)$ is equal to n_s the number of pairs at a distance s, divided by the total number of pairs, n_T. To account for small datasets, the value of $f^{ab}(s)/f(s)$ is estimated as

$$\frac{f^{ab}(s)}{f(s)} = \left(\frac{1}{1 + n_T^{ab}\sigma}\right)\frac{n_s}{n_T} + \left(\frac{n_T^{ab}\sigma}{1 + n_T^{ab}\sigma}\right)\frac{n_s^{ab}}{n_T^{ab}} \tag{9}$$

with $\sigma = \frac{1}{50}$.[19] The value of σ reflects the relative importance given to the pre-data estimate of $f^{ab}(s)/f(s) = n_s/n_T$ and the value calculated based on the data alone.

Tests With Lattice Proteins

Given these different approaches to derive potential energy functions, we wished to identify the approach that generates the most accurate energy functions and are the most successful at predicting protein structures. To address this issue, we applied the method pioneered by Dill using lattice models.[17] The basic idea is to imagine a reality where proteins are described by self-avoiding random walks on lattices, and where the energy function is specified in advance. We can generate a synthetic database of random sequences and their corresponding native states. We then determine the accuracy with which scientists living in this lattice world could reconstruct the true energy function by applying one method or another to the synthetic database. We can also see how successful these scientists would be in predicting the structure of other lattice proteins based on their approximate energy functions. Although obviously simplified, the use of lattice models provides us with a method to evaluate the various optimization methods without the complications and limitations inherent when dealing with a restricted set of real proteins and where the interactions are known with limited accuracy. Although success with lattice models does not guarantee similar performance for predicting the structure of biological proteins, the hope is that insights obtained with the lattice models can be applied to this more complicated problem.

Our lattice model consists of a chain of 27 monomers, confined to a $3 \times 3 \times 3$ three-dimensional maximally-compact cubic lattice, with each monomer located at one lattice point. There are 103,346 possible conformations represented by the set of self-avoiding walks unrelated by rotations or reflections. We assume that the energies of the various conformations are given by a simple contact energy of the form:

$$E = \sum_{\langle ij \rangle} \gamma_{\mathrm{MJ}}(\mathcal{A}_i, \mathcal{A}_j) u(1 - r_{ij}) \quad (10)$$

where $\gamma_{\mathrm{MJ}}(\mathcal{A}_i, \mathcal{A}_j)$ is the contact energy when amino acid $\mathcal{A}_i$ at position i in the sequence is in contact with amino acid $\mathcal{A}_j$ at position j, and $u(1 - r_{ij})$ is a step-function equal to 1 if nonsequential amino acid locations i and j are in contact, that is, on adjacent lattice points. It is assumed that the true energy function for these lattice proteins was the one developed by Miyazawa and Jernigan (MJ).[4] The conformation of lowest energy is assumed to be the native state. All of the other 103,345 conformations make up the set of random conformations.

We assembled three independent training sets of 750, 1,000, 2,500, and 5,000 27-residue proteins, each protein with a random sequence. For each training protein m, we calculated the energy of all 103,346 conformations, considered the structure of lowest energy to be the native state, and calculated Δ_m, Γ_m, and $Z_m = \Delta_m/\Gamma_m$. We then used each of the training sets to estimate the interaction parameters $\gamma(\mathcal{A}_i, \mathcal{A}_j)$ using a variety of different methods including maximizing the value of $\langle \Delta_m \rangle / \langle \Gamma_m \rangle$ as first proposed by GSW, maximizing the value of $\langle 1/Z_m \rangle^{-1}$, as suggested by MS, as well as the simplest procedure, maximization of $\langle Z_m \rangle$. We also maximized $\langle P(\mathcal{S}_m) \rangle$, the probability of success, calculated using both the CG approach and the new CG approach described previously. The estimated potentials were assumed to be symmetric (i.e., $\gamma(\mathcal{A}_i, \mathcal{A}_j) = \gamma(\mathcal{A}_j, \mathcal{A}_i)$), resulting in 210 parameters to be estimated. As the Z-scores are unchanged by additive or multiplicative factors applied to $\gamma(\mathcal{A}_i, \mathcal{A}_j)$, we set two interactions equal to their corresponding MJ potentials and optimized the other 208 interaction potentials using the quasi-Newton algorithm of the NAG software package (Numerical Algorithms Group Ltd., Oxford, UK).

We also estimated the potentials using the POMF approach. To apply this method in a fair manner to the $3 \times 3 \times 3$ three-dimensional cubic lattice model, two modifications are necessary. Firstly, because the contact potentials in the model only depend upon the types of amino acids, the k dependence in POMF method should be averaged out when applied to the lattice model. Secondly, in the lattice model, pairs of noncovalently connected amino acids are either interacting if they are in contact (on adjacent lattice sites) or not interacting. For this reason, there are two possible values of s, corresponding to $s = 1$ (contact, C) or $s > 1$ (noncontact, N). We calculate the energy for these two situations, and subtract the noncontact potential from the contact potential to obtain the potential for a given pair of amino acids in contact relative to the same pair not in contact.

For the $3 \times 3 \times 3$ three-dimensional cubic lattice model, there are always 28 contacts formed out of 156 pairs of residues that can possibly be in contact. $n_{s=1}/n_T$ is therefore equal to 28/156, whereas $n_{s=0}/n_T = 128/156$. kT in Eq. (8) is a multiplicative constant that does not affect the accuracy of the potential or the ability to discriminate between various protein conformations.

Once the various estimates of $\gamma(\mathcal{A}_i, \mathcal{A}_j)$ are made, we use two different measures to evaluate the success of the estimation. The first is to ask how well do the estimated contact potentials agree with the true $\gamma_{\mathrm{MJ}}(\mathcal{A}_i, \mathcal{A}_j)$ potential used to construct the database? Because we are not interested in constants of proportionality, we calculate the correlation coefficient between the true parameters and the estimated parameters. More important is how well these potentials work at predicting the structure of proteins not in the training set. For each of the training sets, we prepared a separate 1,000-protein test set of random sequences, and find the true native state of these sequences in the same manner as for the training set. We then calculate the energy of all 103,346 conformations using the various estimates of the contact parameters, and take the conformation of lowest energy with these estimates as the predicted structure. We then compute the success rate, that is, the fraction of the proteins in the test set whose structure is correctly predicted.

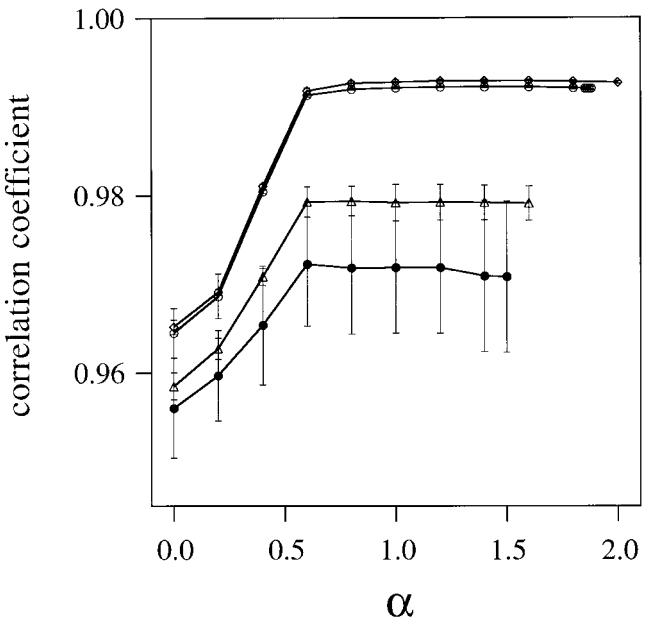

Fig. 1. Correlation coefficients as a function of the α value for various database sizes: datasets of 750 sequences (●), 1,000 sequences (△), 2,500 sequences (○), and 5,000 sequences (◇). Error bars represent the variation of results among three independent sets of equal size.

RESULTS

In contrast to the various other methods, the New CG approach has one more adjustable parameter, α. Figure 1 shows the correlation coefficient of the potentials derived using the new CG method compared with the true MJ potentials as a function of α for different sizes of the training database. As can be seen, the inclusion of the parameter results in an increased ability to extract more accurate potential functions compared with the original CG method ($\alpha = 0$). This translates directly into an improved ability to predict the structure of the proteins in the test set, as shown in Figure 2. The increase in the accuracy of the predictions is considerable: from an average of 67% prediction accuracy to an average of approximately 84% accuracy for the largest training sets.

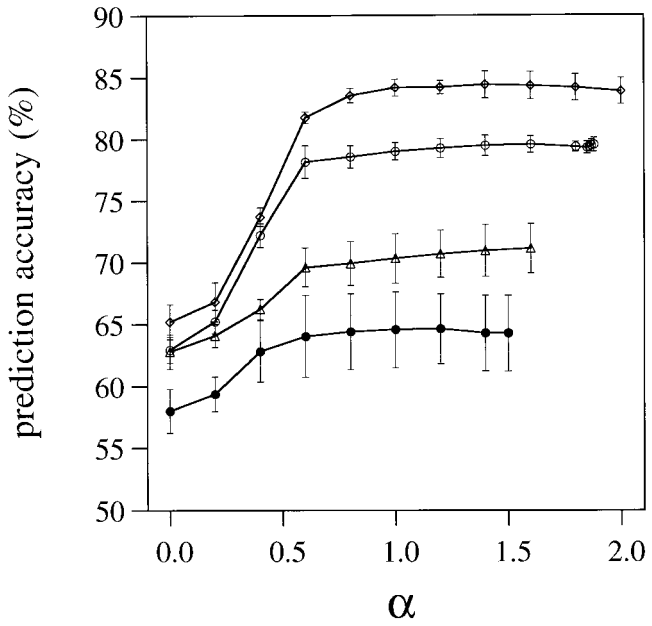

Fig. 2. Prediction accuracy as a function of the α value for various database sizes: datasets of 750 sequences (●), 1,000 sequences (△), 2,500 sequences (○), and 5,000 sequences (◇). Error bars represent the variation of results among three independent sets of equal size.

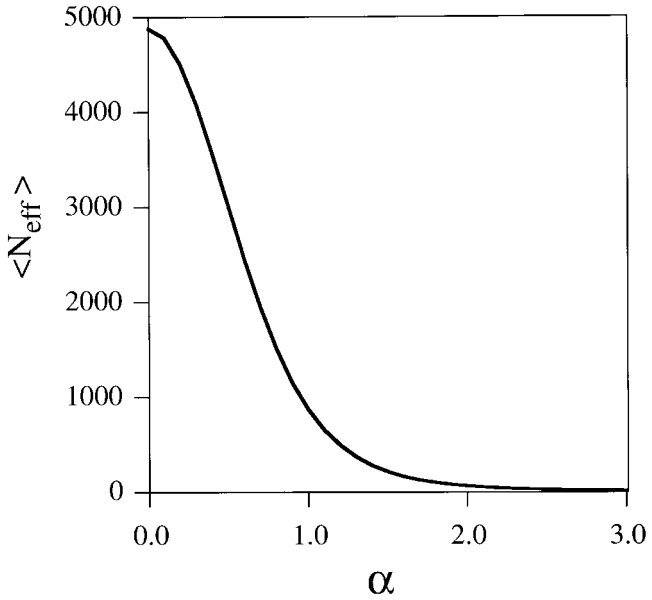

Fig. 3. Average effective number of states, defined in Eq. (11), as a function of α for one of the 5,000-sequence datasets. For $\alpha = 0$, $\langle N_{\text{eff}} \rangle$ equals the total number of structures. As some of the sequences in the dataset fold into the same structure, this total is equal to 4,878.

There is a tendency for reduced performance with large values of α, especially for the smaller training sets. As α increases, fewer random conformations contribute to the calculation of Z_m. This is shown in Figure 3, which shows the average effective number of states $\langle N_{\text{eff}} \rangle$ as a function of α for one of the datasets of size 5,000, with $\langle N_{\text{eff}} \rangle$ defined as

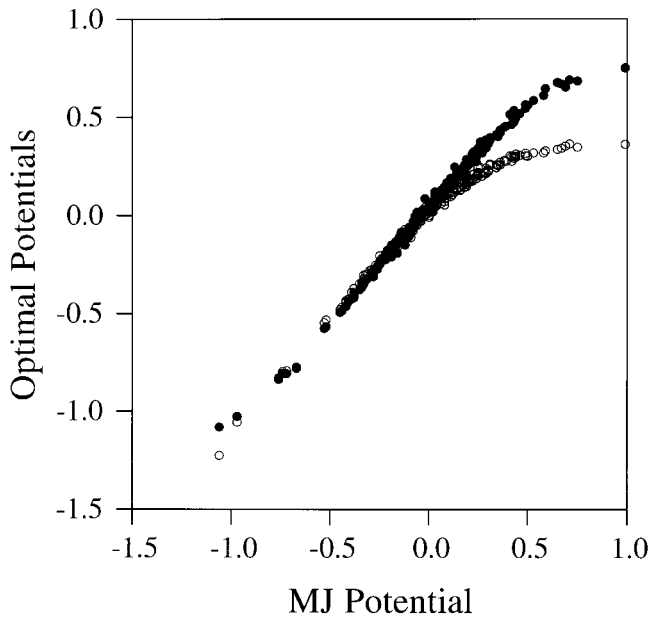

Fig. 4. Optimal potentials derived by the CG method (○) and new CG method at $\alpha = 1.4$ (●) compared with the "correct" MJ potential.

$$\langle N_{\text{eff}} \rangle = \left\langle \left(\sum_{r=1}^{N} \left(\frac{\exp(-\alpha E_r)}{\sum_{r=1}^{N} \exp(-\alpha E_r)} \right)^2 \right)^{-1} \right\rangle \qquad (11)$$

For larger values of α there is a succeptability to overtraining, as the optimization procedure is increasingly able to optimize for these few particular random conformations. This is less of a problem for the larger training sets.

As mentioned in the introduction, the earlier CG method tended to underestimate the value of the destabilizing interactions, as shown in Figure 4 (open circles).[16] This consistent underestimation was also observed in other *Z*-score optimization methods.[10] The new CG method, in contrast, tends to do a better job estimating these potentials (Fig. 4, filled circles). It is this increased accuracy that results in the improved performance in structure prediction.

Table I details a comparison between the different estimation strategies for a 5,000-protein training set of proteins, as measured by average correlation coefficients with the true MJ potential as well as average prediction accuracy on a 1,000-protein test set. Especially as measured by prediction accuracy, the new CG method with $\alpha = 1.4$ has a significantly higher success rate than any of the other *Z*-score based optimization schemes as well as the Potential of Mean Force.

DISCUSSION

The first implementation of this approach maximized $\langle \Delta \rangle / \langle \Gamma \rangle$, where the averages are over the database proteins, to enable a closed-form solution for the optimal energy function. In contrast, Mirny and Shakhnovich (MS) maximized the harmonic average of individual *Z*-scores to obtain the optimal potential for a set of proteins, motivated

TABLE I. Comparison of Various Methods[†]

Method	Correlation coefficient	Prediction accuracy
GLW	0.938	53.1%
Z-avg	0.939	53.8%
MS	0.946	56.0%
CG	0.965	63.9%
POMF	0.987	68.7%
New CG	0.993	82.9%

[†]Average correlation coefficients between the potentials derived by various methods and the "true" MJ potential. These methods include maximization of $\langle\Delta_m\rangle/\langle\Gamma_m\rangle$ (GLW), $\langle Z_m\rangle$ (Z-avg), $\langle 1/Z_m\rangle^{-1}$ (MS), maximizing the probability of a successful prediction using the previous CG approach (CG) and the New CG approach described in this paper (α = 1.4) (new CG), as well as calculating the potential using the potential of mean force approach (POMF). Also shown is the average percentage of correct predictions made on the test database using the variously derived potentials. All results are based on 5,000-sequence training sets and 1,000-sequence test sets.

by the desire for proteins with low Z-scores to dominate the averaging procedure.[10] This averaging procedure allowed us to concentrate on the proteins with intermediate Z-scores rather than the ones with extremely low or high Z-scores, thus giving less weight to proteins whose predictions are either highly unlikely or overly easy.

In the development of these potentials, the distribution of interactions in folded proteins is assumed to represent an uncorrelated Boltzmann weighting of the interaction energy. Although these potentials have achieved some degree of success, they suffer from a number of problems. The theoretical justification for this approach is difficult and nonintuitive.[4,22,23] In addition, the potentials of mean force generally assume statistical independence of the various interactions. This is quite problematic, both for trivial reasons (if hydrophobic residues are buried away from the surface to avoid interactions with the solvent, they will tend to be clustered near each other, causing a greater chance that they will be in contact even in the absence of interactions between them) and deeper reasons (the consistency principle of Gō and the principle of minimal frustration imply that correlations between interactions may arise in order to facilitate the folding process.[24,25] The growing use of more complicated potential energy functions and information derived from alternative sources (such as experimental results) will make assumptions of statistical independence increasingly invalid.

It has been proposed that the underestimation of the repulsive interactions in optimized potentials results from the random error[10] or because of overoptimization of the potentials.[18] We propose the reason for this systematic bias is that, as we maximize the Z-score (and thus reduce Γ), we overstabilize the high-energy random structures, resulting in the underestimation of the repulsive potentials. This is supported by the observation that we can reduce the systematic bias to a considerable amount by suppressing the high-energy state contribution to the Z-score calculation. Using this approach, we improve our previous CG approach by incorporating a refined Z-score implementation procedure into the calculation of the average probability of success. As can be seen in Table I, this

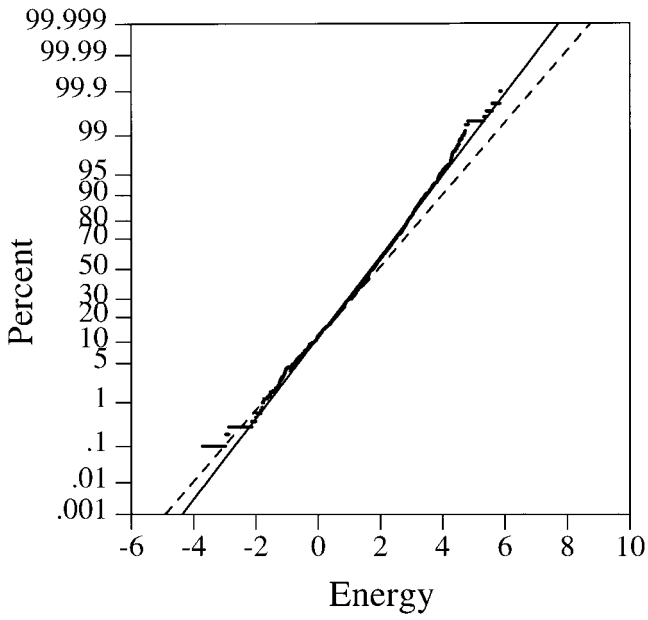

Fig. 5. The normal probability plot showing the cumulative distribution of the energies of a typical set of random structures (· · ·), compared with the models generated using the CG (——) and New CG methods (– – –).

new CG method reproduces the "real" potential with the correlation coefficient as high as 0.993. This improved method generates potentials that are more accurate than those generated by other potential derivation methods including previous Z-score based optimization methods as well as the potential-of-mean-force approach.

There is another advantage to focus on the lower-energy part of the Gaussian distribution. As the precision of our calculation of the average probability of success is especially dependent on the lower energy tail, this indicates that there might be an advantage to more accurately modeling this part of the distribution. Figure 5 shows a normal probability plot showing the cumulative of the distribution of the energies of a typical set of random structures, compared with the models generated using the CG and new CG methods. A normal probability plot will be a straight line plot, to within sampling error, if Gaussian is a good representation of $\rho_m(E_r)$. As can be seen, the new CG method captures the distribution of the low-energy states better than the previous Gaussian representation.

We can understand the existence of the optimal value of α by considering how systematic bias and random bias vary as α is modified. The new CG is completely identical to CG when α is equal to zero. Even though the systematic bias is reduced as α is increased, random bias increases because of the fall of the effective number of states contributing to the calculation of the averages, resulting in reduced accuracy as well as a strong dependence of the potentials on the training set, as shown in Figures 2 and 3. As α becomes larger, the Z-score calculation becomes dominated by the few alternative low-energy conformations, similar in spirit to the optimization procedure of Crippen.[5,6,9,11] The danger is that the small number of alternative conformations can induce sampling errors,

especially for small datasets. Incorporation of α as an additional parameter allows us to adjust the number of alternative states being considered, interpolating between giving too much emphasis to thermodynamically-irrelevant high-energy conformations or to the fewest low-energy states.

ACKNOWLEDGMENTS

We would like to thank Todd Raeker and Michael Kitson for computational assistance.

REFERENCES

1. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
2. Godzik A, Kolinski A, Skolnick J. A topology fingerprint approach to the inverse protein folding problem. J Mol Biol 1992;227:227–238.
3. Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. J Mol Biol 1992;224:725–732.
4. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
5. Crippen GM. Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 1991;30: 4232–4237.
6. Maiorov VN, Crippen GM. Contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
7. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein folding codes from spin glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
8. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
9. Crippen GM, Maiorov VN. In: Merz KM, Grand SML, editors. The protein folding problem and tertiary structure prediction. Boston: Birkhauser; 1994. p 231–277.
10. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
11. Crippen GM. Easily searched protein folding potentials. J Mol Biol 1996;260:467–475.
12. Seno F, Maritan A, Banavar JR. Interaction potentials for protein folding. Proteins 1998;30:244–248.
13. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. A Bayesian approach to sequence alignment algorithms for protein structure recognition. In: Hunter L, editor. Proc. 27th Annual Hawaii International Conference on System Sciences. Los Alamitos: IEEE Computer Society Press; 1994. p 306–315.
14. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. The statistical mechanical basis of sequence alignment algorithms for protein structure recognition. In: Elber R, editor. New developments in theoretical studies of proteins. Singapore: World Scientific; 1996. p 359–388.
15. Doolittle RF. Similar amino acid sequences: chance or common ancestry? Science 1981;214:149–159.
16. Chiu TL, Goldstein RA. Optimizing energy potentials for success in protein tertiary structure prediction. Folding Design 1998;3:223–228.
17. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
18. Zhang L, Skolnick J. How do potentials derived from structural databases relate to "true" potentials? Protein Sci 1998;7:112–122.
19. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. J Mol Biol 1990;213:859–883.
20. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models. J Mol Biol 1990;216:167–180.
21 Sippl MJ. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. Curr Opin Struct Biol 1995;5:229–235.
22. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. J Mol Biol 1987;193:723–750.
23. Finkelstein AV, Gutin AM, Badretdinov AY. Boltzmann-like statistics of protein architectures. Subcell Biochem 1995;24:1–26.
24. Gō N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
25. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84: 7524–7528.