# Optimizing energy potentials for success in protein tertiary structure prediction

Ting-Lan Chiu[1] and Richard A Goldstein[1,2]

**Background**: Success in solving the protein structure prediction problem relies on the choice of an accurate potential energy function. For a single protein sequence, it has been shown that the potential energy function can be optimized for predictive success by maximizing the energy gap between the correct structure and the ensemble of random structures relative to the distribution of the energies of these random structures (the Z-score). Different methods have been described for implementing this procedure for an ensemble of database proteins. Here, we demonstrate a new approach.

**Results**: For a single protein sequence, the probability of success (i.e. the probability that the folded state is the lowest energy state) is derived. We then maximize the average probability of success for a set of proteins to obtain the optimal potential energy function. This results in maximum attention being focused on the proteins whose structures are difficult but not impossible to predict.

**Conclusions**: Using a lattice model of proteins, we show that the optimal interaction potentials obtained by our method are both more accurate and more likely to produce successful predictions than those obtained by other averaging procedures.

Addresses: [1]Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055, USA. [2]Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055, USA.

Correspondence: Richard A Goldstein
E-mail: richardg@umich.edu

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

Received: 17 November 1997
Revisions requested: 06 January 1998
Revisions received: 25 February 1998
Accepted: 12 March 1998

Published: 07 May 1998
http://biomednet.com/elecref/1359027800300223

Folding & Design 07 May 1998, 3:223–228



## Introduction

Research into the specific function and mechanism of a protein generally starts with knowledge of its three-dimensional structure. The only way of determining the structures of most proteins is through laborious and time-consuming experimental methods such as X-ray crystallography or multi-dimensional NMR. In contrast, determining the sequences of proteins has become relatively easy. The development of a general method for the prediction of protein tertiary structure based on the protein sequence remains, unfortunately, one of the great unsolved problems in computational biophysics.

Predicting the three-dimensional conformation of a correctly folded protein can be divided into two distinct steps: the construction of a fitness function to evaluate the various conformations; and the search through various possible conformations for the 'best' prediction most likely to represent the native state. Neither part of this problem has proven particularly tractable. The choice of an appropriate fitness function has been a matter of intense debate. If we assume that the correct conformation of the folded protein represents the structure of minimum free energy, the most natural cost function is the value of this free energy. Unfortunately, the parameters describing the free energy, especially interactions between the protein and the solvent, are still the subject of much uncertainty. In addition, an accurate energy function would require a complete representation of all the atoms in the protein, and some representation of the solvent degrees of freedom, thus greatly increasing the conformational space to be searched. For this reason, there has been increasing interest in developing appropriate potential functions that work for simplified 'reduced representations' of the protein conformations.

Although it is possible to develop energy functions based on empirical measurements or calculations for small organic molecules, it has become common to look at the statistical properties of the database of proteins of known structures to ascertain the values of the various energetic parameters. Many of the potentials currently being developed can be considered variations of 'potentials of mean force' [1–3], derived from a statistical analysis of the protein database based on the quasi-chemical approximation of Miyazawa and Jernigan [4]. In the development of these potentials, the distribution of interactions in folded proteins is assumed to represent an uncorrelated thermodynamic weighting of the interaction energy. Although these potentials have achieved some degree of success, they suffer from a number of problems. Progress has been made in justifying this approach from theoretical principles [4–6], but there is no *a priori* reason to believe that the interactions in proteins in the respective ground states of an ensemble of biological proteins would obey Boltzmann statistics. 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 in order to avoid interactions with the solvent, they will tend to be clustered near each other, resulting in 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 [7] and the principle of minimal frustration [8] imply that correlations between interactions may arise in order to facilitate the folding process).

An alternative approach has been based on deriving energy functions that are optimized to predict protein conformations, generally by ensuring that the energy of the correctly folded state is as low as possible compared with the lowest energy incorrect states [9–12]. Implicit in this approach is that it is necessary not only to stabilize the correct structures, but also to destabilize incorrect ones. Such a principle has also been used to design amino acid sequences that would fold into a given native state [13,14]. There have been a number of variations on this principle. Goldstein, Luthey-Schulten and Wolynes (GLW; [10,11,15,16]) approached this problem using both techniques drawn from spin-glass theory and from Bayesian statistics. According to their work, the important quantity was the difference in energy level of the correct native state compared with the average energy level of the random conformations ($\Delta$), divided by the standard deviation of the energy levels of the random conformations ($\Gamma$); this is similar to the Z-score in the sequence-alignment literature [17]. The best energy potential would be the potential that maximized this Z-score for the proteins in the database. For an ensemble of database proteins, GLW chose to maximize $\langle\Delta\rangle/\langle\Gamma\rangle$, where the averages are over the database proteins, in order to enable a closed-form solution for the optimal energy function. In contrast, Mirny and Shakhnovich (MS; [12]) maximized the harmonic average of individual Z-scores to obtain the optimal potential for a set of proteins, motivated by the desire for proteins with low Z-scores to dominate the averaging procedure.

In this paper, we present a new approach towards finding the optimal potential for a set of proteins. It is based on the principle that it is always best to optimize the quantity that you are most interested in maximizing. If we are interested in developing an energy potential that is as successful as possible at predicting protein structures, we should optimize the success rate. This averaging procedure allows us to concentrate on the proteins with intermediate Z-scores rather than the ones with extremely low or high Z-scores, thus neglecting the proteins whose predictions are either highly unlikely or overly easy.

Given these different approaches towards the problem, certain questions emerge. Which approach gives the most accurate energy potentials? Which potentials are more successful at predicting protein structures? Answering either of these questions is difficult. Because we do not know the true potentials, we cannot evaluate which derived one is most accurate. Although there have been canonical sets of protein structures developed to test structure prediction algorithms, the results are necessarily anecdotal and are complicated by differences in implementation of the various techniques.

An approach towards answering questions such as these was pioneered by Thomas and Dill [18] using lattice models. The basic idea was to imagine a reality in which proteins are described by self-avoiding random walks on lattices and 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.

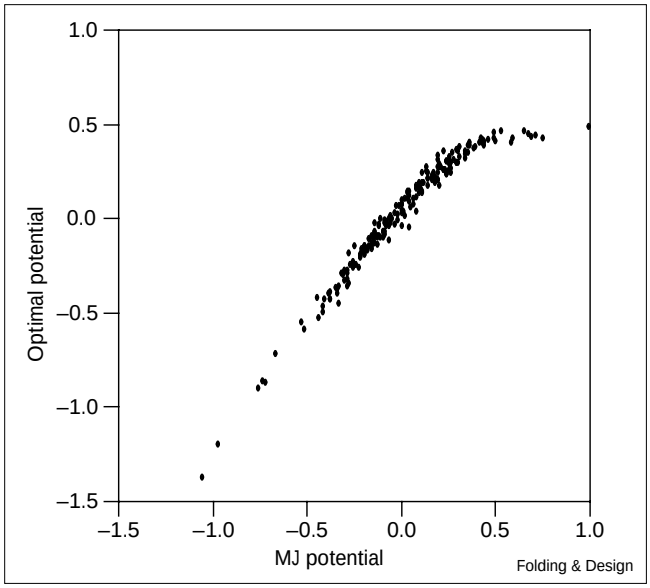
In general, we find that the approach of optimizing the probability of success generates potentials that are more accurate than generated either by optimizing $\langle\Delta\rangle/\langle\Gamma\rangle$ as in the GLW method or the harmonic mean of the Z-scores as in the more recent MS approach. We also demonstrate that our method is significantly more likely to be successful at predicting the structures of proteins not in the database.

## Results

As mentioned above, there have been a number of methods proposed for optimizing interaction potentials based on maximizing Z-scores over a training database. The most fundamental difference is the nature of the averaging over the various proteins. In the GLW approach, $\langle\Delta\rangle/\langle\Gamma\rangle$ is optimized. In the MS approach, the harmonic mean of the Z-scores, $\langle 1/Z\rangle^{-1}$ is optimized. We tried the most obvious approach, optimizing the simple mean of the Z-scores ($Z_{avg}$). In our (CG) approach, we optimize the average probability of a successful prediction, calculated as described in the Materials and methods section.

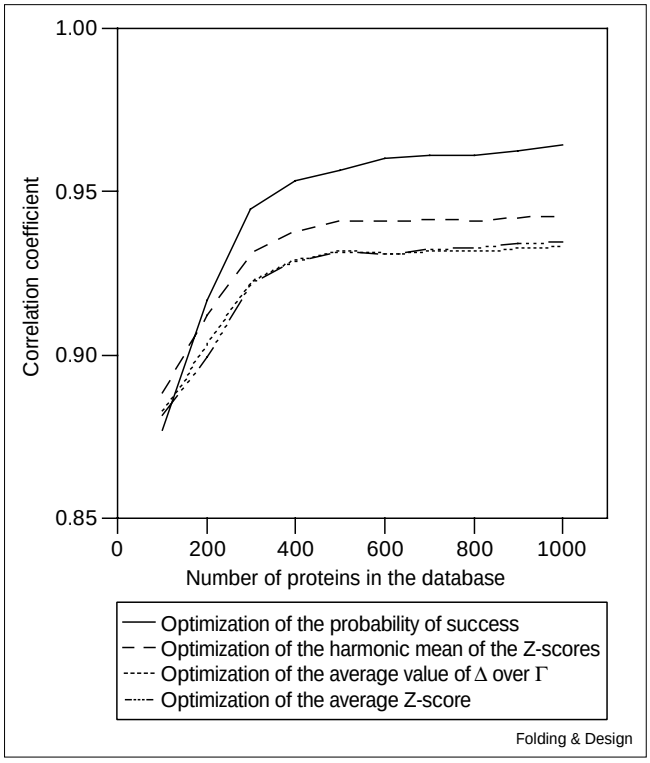
In order to compare these various methods, a database was constructed consisting of 1000 27-residue proteins made up of random amino acid sequences. These proteins were assumed to fold into a state confined to a $3\times3\times3$ three-dimensional cubic lattice, in which the distances between adjacent residues are all of unit length. It is possible to enumerate all 103,346 possible self-avoiding walks on the lattice. A contact exists if two residues are on adjacent sites but are not adjacent in sequence. It was assumed that the true energy function for these lattice proteins was the one developed by Miyazawa and Jernigan (MJ; [4]), which implicitly includes the effect of interactions between the protein and the solvent. Using this interaction potential, we

**Figure 1**




Optimal potential derived by our approach compared with the 'correct' MJ [4] potential.

**Figure 2**




Correlation coefficients as a function of the number of proteins in the database for various approaches. Optimization of the probability of success, CG (this work); optimization of the harmonic mean of the Z-scores, MS [12]; optimization of the average value of Δ over the average value of Γ, GLW [10,11,15,16]; and optimization of the average Z-score ($Z_{avg}$).

were able to calculate the energies of all possible compact conformations for every sequence. The conformation with the lowest energy is the correct native structure corresponding to that particular sequence. The 1000 different sequences corresponded to 992 unique native folded states.

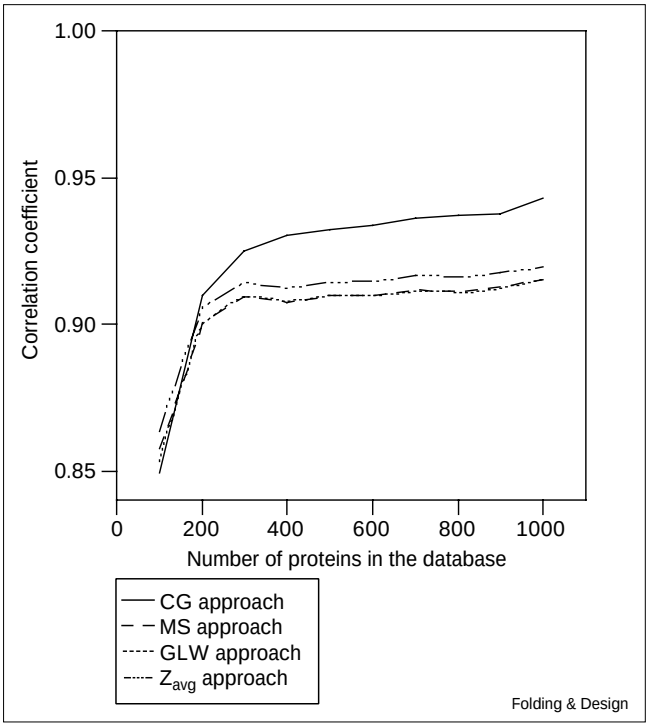
Because the interactions are assumed to be symmetric, the energy function is specified by the 210 contact potentials representing all possible pairs of amino acids. As only relative energy levels are relevant, and all possible structures have the same total number of contacts, adding or multiplying a constant to the derived potentials will not change the result. It is therefore necessary to eliminate two degrees of freedom to obtain a unique set of optimal potentials. We did this by setting two interactions equal to their corresponding MJ potentials and optimizing the other 208 interaction potentials. For all the different approaches, the optimization was performed using the sequential quadratic algorithm of the NAG software package (Numerical Algorithms Group Ltd, Oxford, UK).

The relative values of the potentials optimized for overall success versus the 'correct' MJ potentials are shown in Figure 1. We measured the accuracy of the derived potentials by calculating the correlation coefficients between these potentials and the MJ potentials. These coefficients as a function of database size for the various optimization methods are shown in Figure 2.

As mentioned above, the synthetic database was constructed of random sequences and their associated native states, assumed to be the conformations of lowest energy using the MJ potential. This implicitly assumes that all sequences are possible and represent viable, foldable proteins. In contrast, theoretical models and lattice simulations have suggested that only proteins with an adequate value of the Z-score would be able to fold [10,19–21]. In order to investigate how this constraint on the dataset would affect the results of the optimization, lattice proteins were grouped according to their Z-scores and a database was constructed of sequences with Z-scores > 5.0. ~0.87% of random protein sequences fulfilled this criteria. The accuracy of the various optimization methods for calculating the true energy potential as a function of the database size for this second database is shown in Figure 3.

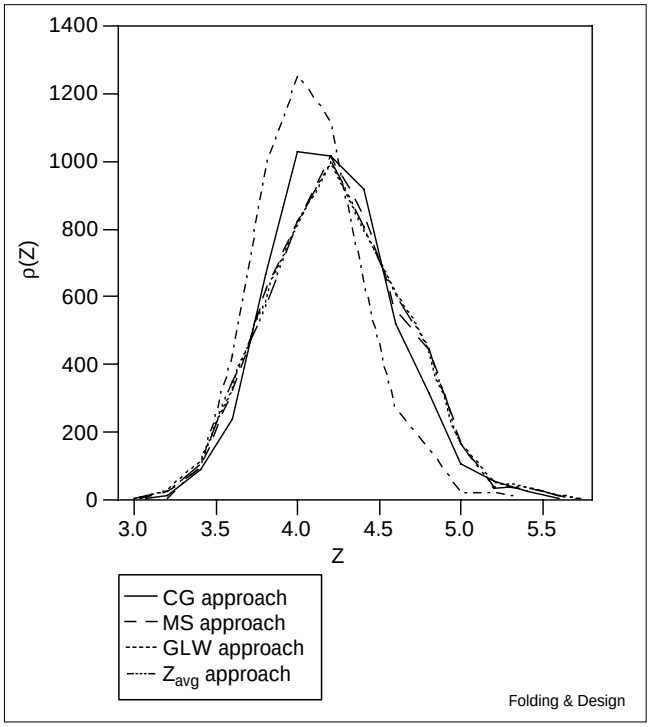
The purpose of these optimization procedures is to maximize our ability to predict the native conformations of proteins of unknown structure. In order to compare the various optimization methods, we generated a second indepzendent database of 1000 test proteins with their 992 different corresponding native states, and calculated the fraction of these proteins that would have their correct structure selected using the various optimized

**Figure 3**




Correlation coefficients as a function of the number of proteins in the database of proteins which all have Z-scores > 5.0.

**Figure 4**




The distribution of proteins in the test set with a given value of Z-score for various approaches.

potentials. All the potentials did reasonably well at selecting the correct structure from among the 992 possibilities; the CG method described above yielded the highest success rate of 98.9%, compared with success rates of 98.3% for the MS method, 97.9% for the GLW method, and 97.7% for the $Z_{avg}$ method. The relative rates fell significantly when the potentials were used to predict the correct structure from among all 103,346 possible compact lattice structures. The CG method was successful with 61.9% of the test set proteins, significantly higher than the success rates achieved with the MS potential (54.2%), $Z_{avg}$ potential (50.5%) and GLW potential (49.8%).

The distribution of Z-scores calculated for these test proteins with the different energy functions is plotted in Figure 4. All optimization methods yield significantly higher Z-scores than the true energy function. In contrast to the other optimization schemes, the CG method works to maximally increase the Z-score of proteins for which the structures are possible yet difficult to predict — those with Z-scores of ~4.0.

## Discussion

In this paper, we propose a novel approach towards finding the optimal potential based on a set of database proteins. Specifically, we developed an expression to directly quantify the probability of success in predicting protein structures. All the various optimization methods discussed in this paper are equivalent for a single database protein, optimizing the Z-score for that protein. The difference between these procedures is in how this method is generalized to a larger heterogeneous database that contains some proteins that may be easier to predict than others. Previous methods have been developed in order to allow closed-form solutions (GLW) or to focus on the proteins that are the hardest to predict (MS). In contrast, we maximized the average probability of success and aimed to predict as many protein structures as possible, providing more accurate energy parameters that are more likely to generate accurate predictions.

For the hardest protein structures to predict, which had low Z-scores and for which the probability of a successful prediction is small, increasing the Z-scores will not greatly increase the probability of a successful prediction. Similarly, in the cases where the Z-score is high and confident predictions can be made, further optimization is not warranted. Our method concentrates directly on proteins whose structure prediction is challenging, but not impossible. It is exactly this ability to focus on the proteins at the border between predictability and

non-predictability that allows us to maximally increase the overall rate of success.

As shown in Figure 1, the optimization approach tends to be inaccurate for the repulsive interactions with large positive contact energies. The biggest discrepancy between these two potentials is the tendency to underestimate strongly destabilizing interactions, also observed in optimization procedures that optimize the harmonic mean of the Z-scores [12]. This is likely to be a result of the use of a Gaussian approximation for the random states, in that the Z-score is actually decreased by having bad states overly high in energy as a result of the increase in the standard deviation of the energies of the random states, $\Gamma$.

There are then two sources of errors in the optimized potentials — the systematic bias towards underestimating destabilizing interactions and the more random scatter of the potentials around this bias. This scatter may be a result of the size of the database of known structures or may involve more complicated effects such as correlations between the various contact potentials. In order to separate the consequences of these two effects on the prediction accuracy, we represented the systematic bias by modeling the relationship between true and optimized potentials (Figure 1) with a cubic spine. We then calculated the accuracy of our method with the systematic bias removed and only the scatter remaining. Similarly, we calculated the effect of removing the scatter and computed the accuracy of our method if the contact potentials were assumed to lie exactly on the cubic-spline fit. The ability of the potential to distinguish the correct structure from all possible compact structures increased from 61.9% to 73.6% when the systematic bias was removed. Similarly, the success rate increased to 71.0% when the deviations around the systematic bias were eliminated. This suggests that we should have a significant increase in prediction accuracy through the use of a theoretical model that fits the energy distributions better than the Gaussian distribution. Similarly, the use of a larger dataset and a model that better includes the correlations between the contact potentials might similarly increase the prediction accuracy.

## Materials and methods

We are generally interested in predicting the native structure of protein m by identifying the native conformation $C_{NS}^m$ from a large set of N possible conformations $\{C_k\}$. We do this by calculating $\{E_k^m\}$, the energy of protein sequence m for every possible conformation k using some unspecified energy function $H$. We then choose the lowest energy state as our predicted structure. We choose correctly only if the lowest energy state, computed using $H$, is indeed the correct state, or alternatively, if every incorrect state is higher in energy. The question is, what is the best energy function to use for this application?

Let us assume that $\rho_m(E_r)$, the distribution of energies of protein m in the random conformations computed using $H$, is a Gaussian centered at $\bar{E}_r^m$ with standard deviation $\Gamma_m$, whereas the correct native-state structure has energy $E_{NS}^m$. The probability that any individual random structure with energy $E_r$ drawn from $\rho_m(E_r)$ has an energy larger than that of the correct structure is given by:

$$\begin{aligned} P(E_r > E_{NS}^m) &= \int_{E_{NS}^m}^{\infty} \rho_m(E_r)\, dE_r \\ &= \frac{1}{\sqrt{2\pi}\,\Gamma_m} \int_{E_{NS}^m}^{\infty} \exp\left( \frac{(E_r - E_r^m)^2}{2\Gamma_m^2} \right) dE_r \\ &= 0.5 + 0.5\ \mathrm{erf}\left( \frac{Z_m}{\sqrt{2}} \right) \end{aligned} \quad (1)$$

where $Z_m$ (the Z-score for protein m) = $(\bar{E}_r^m - E_{NS}^m)/\Gamma_m$.

In order for us to be successful in correctly predicting the structure from among the N incorrect alternatives, all the other structures must have an energy > $E_{NS}^m$. $P(S_m)$, the probability of 'success' for sequence m, is given by:

$$P(S_m) = \left[ 0.5 + 0.5\ \mathrm{erf}\left( \frac{Z_m}{\sqrt{2}} \right) \right]^N \quad (2)$$

For a single sequence, the probability of success is a monotonically increasing function of the Z-score, so the optimal energy function is the one that maximizes this quantity. For an ensemble of proteins, we are interested in generating the largest possible number of correct predictions. This corresponds to optimizing the average probability of success. So, in contrast to maximizing either $\langle\Delta\rangle/\langle\Gamma\rangle$, as done by Wolynes and coworkers [10,11,15,16], or the harmonic mean of Z, as done by Mirny and Shakhnovich [12], we maximize the average probability that the energy function would yield a correct prediction:

$$\left\langle P(S_m) \right\rangle = \left\langle 0.5 + 0.5\ \mathrm{erf}\left( \frac{Z_m}{\sqrt{2}} \right) \right\rangle \quad (3)$$

## Acknowledgements

We would like to thank Kurt Hillig and James Raines for computational assistance, and Sridhar Govindarajan and Michael Thompson for helpful discussions. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, NIH Grant LM0577, and NSF equipment grant BIR9512955.

## References

1. 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.
2. Godzik, A., Kolinski, A. & Skolnick, J. (1992). A topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.* **227**, 227-238.
3. Sippl, M.J. & Weitckus, S. (1992). Detection of native-like models for amino-acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **13**, 258-271.
4. Miyazawa, S. & Jernigan, R.L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534-552.
5. Berg, O.G. & von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins. *J. Mol. Biol.* **193**, 723-750.
6. Finkelstein, A.V., Gutin, A.M. & Badretdinov, A.Y. (1995). Boltzmann-like statistics of protein architectures. *Subcell. Biochem.* **24**, 1-26.
7. Gō, N. (1983). Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183-210.
8. Bryngelson, J.D. & Wolynes, P.G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA* **84**, 7524-7528.

9. Crippen, G.M. (1991). Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry* 30, 4232-4237.
10. Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1992). Optimal protein folding codes from spin glass theory. *Proc. Natl Acad. Sci. USA* 89, 4918-4922.
11. Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1992). Protein tertiary structure recognition using optimized hamiltonians with local interactions. *Proc. Natl Acad. Sci. USA* 89, 9029-9033.
12. Mirny, L.A. & Shakhnovich, E.I. (1996). How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* 264, 1164-1179.
13. Shakhnovich, E.I. & Gutin, A.M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA* 90, 7195-7199.
14. Shakhnovich, E.I. (1994). Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* 72, 3907-3910.
15. Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1994). A Bayesian approach to sequence alignment algorithms for protein structure recognition. In *Proceedings of the 27th Annual Hawaii International Conference on System Sciences* (Hunter, L., ed.), pp. 306-315. IEEE Computer Society Press, Los Alamitos, USA.
16. Goldstein, R.A., Luthey-Schulten, Z.A. & Wolynes, P.G. (1996). The statistical mechanical basis of sequence alignment algorithms for protein structure recognition. In *New Developments in Theoretical Studies of Proteins* (Elber, R., ed.), pp. 359-388. World Scientific, Singapore.
17. Doolittle, R.F. (1981). Similar amino acid sequences: chance or common ancestry. *Science* 214, 149-159.
18. Thomas, P.D. & Dill, K.A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* 257, 457-469.
19. Fukugita, M., Lancaster, D. & Mitchard, M.G. (1993). Kinematics and thermodynamics of a folding heteropolymer. *Proc. Natl Acad. Sci. USA* 90, 6365-6368.
20. Šali, A., Shakhnovich, E.I. & Karplus, M.J. (1994). Kinetics of protein folding: a lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* 235, 1614–1636.
21. Chan, H.S. & Dill, K.A. (1994). Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.* 100, 9238-9257.


**Because *Folding & Design* operates a 'Continuous Publication System' for Research Papers, this paper has been published on the internet before being printed. The paper can be accessed from http://biomednet.com/cbiology/fad – for further information, see the explanation on the contents pages.**