Protein Engineering vol.11 no.9 pp.749–752, 1998

Optimizing potentials for the inverse protein folding problem

Ting-Lan Chiu[1] and Richard A.Goldstein[1,2,3]

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

[3]To whom correspondence should be addressed

Inverse protein folding, which seeks to identify sequences that fold into a given structure, has been approached by threading candidate sequences onto the structure and scoring them with database-derived potentials. The sequences with the lowest energies are predicted to fold into that structure. It has been argued that the limited success of this type of approach is not due to the discrepancy between the scoring potential and the true potential but is rather due to the fact that sequences choose their lowest-energy structure rather than structures choosing the lowest-energy sequences. Here we develop a non-physical potential scheme optimized for the inverse folding problem. We maximize the average probability of success for a set of lattice proteins to obtain the optimal potential energy function, and show that the potential obtained by our method is more likely to produce successful predictions than the true potential.

Keywords: contact potential/lattice proteins/sequence recognition/structure recognition/protein folding

Introduction

There has been much effort expended in developing methods to predict the structure of a protein based on its amino acid sequence. While there is still not a general solution to this problem, progress has been made in the more limited problem of fold recognition, identifying whether a given sequence folds into a previously observed structure. One complication is that it is not necessarily true that the structure of this sequence has been previously observed.

Eisenberg and co-workers developed a different approach to this problem, called the 'inverse folding problem' (Bowie *et al*., 1991). Can we find the sequences in the database that fold into a particular protein structure? A related question involves designing a sequence to fold into a given structure (Shakhnovich *et al*., 1991; Yue and Dill, 1992; Šali *et al*., 1994a; Godzik, 1995). The general approach to this problem has been to thread each sequence in the database onto the given structure, and to use a database-derived potential to score the sequence. The sequences with the lowest energies calculated with this potential are predicted to fold into the target structure. Generally potentials developed for structure prediction are also used for inverse folding, the rationale being that fold recognition and sequence recognition both deal with the compatibility of a sequence with a structure, and therefore what works for one should also work for the other.

Sippl and Crippen have argued that the limited success of this type of approach is not due to the discrepancy between the database-derived potential and the true potential but is rather due to the fact that nature uses the true potential to select the best structure for a sequence rather than the other way around (Sippl, 1993; Crippen and Maiorov, 1995; Crippen, 1996). An illustration of this point is given by Figure 1, which shows the energy of two sequences, 1 and 2, in two different structures, *A* and *B*. It is generally assumed that proteins obey the 'thermodynamic hypothesis', that the native state represents the state of lowest free energy (Anfinsen, 1973; Govindarajan and Goldstein, 1998). In this case, sequence 1 would fold into structure *A* while sequence 2 would fold into structure *B*. There is no similar hypothesis that states that the *sequence* with the lowest energy in a given structure would necessarily fold into that structure. In this example, sequence 2 would fold into structure *B* although sequence 1 has a lower energy in this structure. Because of this reason, even a perfectly accurate energy function may not be adequate for inverse folding. As an example, Crippen used a two-dimensional square lattice model involving two residue types to show that even if the true energy function was known in advance, sequence identification using the true energy can yield only limited success (Crippen, 1996).

While the thermodynamic hypothesis gives the structure prediction problem a conceptual basis absent in the inverse folding problem, there is a rough symmetry from a machine-learning point of view. There have been a number of approaches introduced to develop optimal energy functions, that is, functions that are optimized for protein structure prediction rather than for physical-chemical accuracy (Crippen, 1991; Maiorov and Crippen, 1992; Goldstein *et al*., 1992a,b; Mirny and Shakhnovich, 1996; Chiu and Goldstein, 1998). It may be possible to optimize a potential for accuracy in inverse folding that may actually work better than the true energy function.

According to the approach developed by Wolynes and co-workers, the optimal energy function for structure prediction is the one that maximized the energy difference between the energy of the target protein in its correct structure versus the average energy of the same protein in random structures, normalized by the distribution of energies of the random structures—a quantity that they called 'R' (Goldstein *et al*., 1992a,b). For a set of training proteins, they derived the energy function that maximized the average energy difference divided by the average distribution of energies of the random structures, a formulation that allowed for a closed-form solution. Later work improved on this approach by including the presence of correlations in the energy landscape and the effect of the energy potential on the conformations sampled at a finite temperature (Hao and Scheraga, 1996; Koretke *et al*., 1996).

In their formulation of the inverse folding problem, Eisenberg and co-workers measured the significance of a match by calculating the gap between the energy of the correct sequence in the given conformation compared with the distribution of the energies of random sequences in the same conformation, a quantity which they designated as the 'Z-score' (Bowie

et al., 1991), using the standard notation used to evaluate sequence–sequence alignments (Doolittle, 1981). In later work investigating the properties of protein sequences and their ability to fold, Shakhnovich used the term 'Z-score' to characterize the distribution of energies of a given sequence in a variety of conformations, in the same sense as Wolynes' 'R'. Shakhnovich then developed a variation on the R/Z-score optimization procedure for structure prediction by maximizing the harmonic average of the R/Z-score over a training set of proteins, a procedure that gives greater weight to the proteins that are hardest to predict compared with the earlier Wolynes formulation (Mirny and Shakhnovich, 1996). More recently, Chiu and Goldstein developed an approach that optimizes the average success rate (Chiu and Goldstein, 1998). This averaging procedure concentrates on the protein conformations with intermediate R/Z-scores rather than the ones with extremely low or high R/Z-scores, thus neglecting the protein conformations whose predictions are either highly unlikely or overly easy.

Here we describe the development of a non-physical potential optimized for the inverse folding problem, based on our modification of the R/Z-score optimization criterion described previously (Chiu and Goldstein, 1998). Following the approach of Thomas, Dill and Crippen, we generated a synthetic database of random sequences and their corresponding native states where the true energy function was specified in advance (Crippen, 1996; Thomas and Dill, 1996). We derived the optimal energy function for inverse folding using our method, and compared true and optimal energy functions in their performance in identifying sequences corresponding to these structures. In general, we find that our approach of optimizing the average probability of success generates potentials that are more likely to be successful at predicting the sequences of non-database proteins than the true potential.

The use of multiple notation has led to some confusion. In this paper, we use the term 'Z-score', as it was the first term used explicitly for the inverse folding problem, as well as the general notation used in statistics for quantities of this type.

Materials and methods

The method for generating an optimized energy potential for inverse folding is similar to our previously described method of optimizing an energy potential for fold recognition (Chiu and Goldstein, 1998). We are generally interested in predicting the sequence that folds into native state conformation m by identifying the correct sequence S^m_{NS} from a large set of N possible sequences $\{S_k\}$. We do this by calculating $\{E^m_k\}$, the energy of every sequence k in protein conformation m using some specified energy function $\mathcal{H}$. We then choose the sequence with the lowest energy as the sequence likely to fold into that structure. We choose correctly only if the lowest energy sequence, computed using $\mathcal{H}$, is indeed the correct sequence, or alternatively, if every incorrect sequence 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 the random sequences in conformation m, is a Gaussian centered at $\bar{E}^m_r$ with standard deviation Γ_m, while the correct sequence has energy E^m_{NS}. The probability that any individual random sequence with energy E_r drawn from $\rho_m(E_r)$ has an energy larger than that of the correct sequence is given by

$$P(E_r > E^m_{\mathrm{NS}}) = \int_{E^m_{\mathrm{NS}}}^{\infty} \rho^m_v(E_r)\, dE_r \qquad (1)$$

$$= \frac{1}{\sqrt{2\pi}\Gamma_m} \int_{E^m_{\mathrm{NS}}}^{\infty} \exp\left(\frac{(E_r - \bar{E}^m_r)^2}{2\Gamma^2_m}\right) dE_r$$

$$= 0.5 + 0.5\ \mathrm{erf}\left(\frac{Z_m}{\sqrt{2}}\right)$$

where Z_m, the Z-score for conformation m, is

$$Z_m = (\bar{E}^m_r - E^m_{\mathrm{NS}})/\Gamma_m\,.$$

In order for us to be successful in correctly predicting the sequence from among the N incorrect alternatives, *all* of the other sequences have to have energy greater than E^m_{NS}. $P(S_m)$, the probability of 'success' for conformation m, is given by

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

For a single conformation the probability of success is a monotonically increasing function of 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. We maximize $\langle P(S_m)\rangle = \left\langle \left(0.5 + 0.5\ \mathrm{erf}\left(\frac{Zm}{\sqrt{2}}\right)\right)\right\rangle$, the average probability that the energy function would yield a correctly identified sequence.

Results

A database was constructed consisting of 1000 27-residue proteins made up of random amino acid sequences. These proteins were assumed to fold into one of the 103 346 possible self-avoiding walks on a 3×3×3 three-dimensional cubic lattice, where the distances between adjacent residues are all of unit length. A contact exists if two residues are on adjacent sites but 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), which implicitly includes the effect of interactions between the protein and the solvent (Miyazawa and Jernigan, 1985). Using this interaction potential, we 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 for that particular sequence. The 1000 different sequences corresponded to 992 unique native states. To simplify the problem, we discarded eight sequences so that there is only one correct sequence for each conformation in the database.

As 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

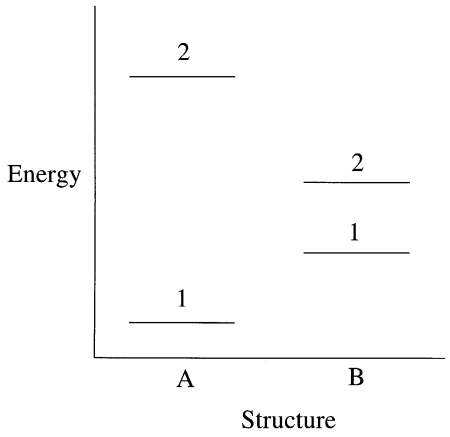

Fig. 1. Energy diagram for two proteins. Based on the 'thermodynamic hypothesis', sequence 1 folds into structure *A*, while sequence 2 folds into structure *B*. An inverse folding approach based on an accurate energy function would incorrectly pick sequence 1 to fold into structure *B*, since it has lower energy than sequence 2 in this structure.

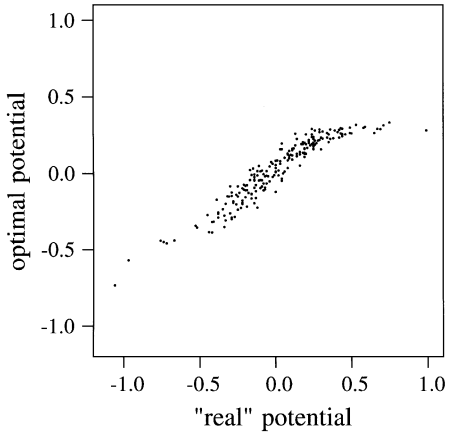

Fig. 2. Optimal potential derived by our approach compared with the 'correct' MJ potential.

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. A trial potential energy function was chosen at random. For each structure, the Z-score was calculated by determining the average and standard deviation of the energies of the random and correct sequences when folded into that structure, and the probability of a correct assignment was calculated using Equation 2. This probability, averaged over all 992 observed structures, was optimized using a sequential quadratic algorithm (E04UCF) of the NAG software package (Numerical Algorithms Group Ltd, Oxford, UK), similar to the SOL/NPSOL subroutine described by Gill and Murray (1974). The values of the optimized potentials compared with the original MJ potentials are shown in Figure 2.

We then computed what fraction of these proteins would have their correct sequence selected using both the optimized potential and MJ potential. Our potential yields the higher success rate of 70.6%, compared with a success rate of 50.2% for the 'true' MJ potential. We also calculated the Z-score for each protein in the database with both energy functions, and calculated the probability of a successful prediction using Equation 2 with $N = 992$. The fraction of these training set proteins with a probability of success greater than any cut-off is plotted in Figure 3.

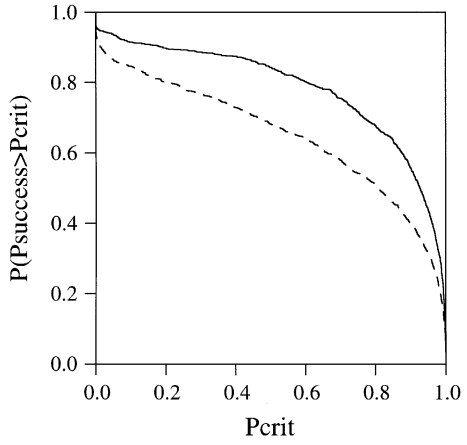

Fig. 3. The fraction of proteins in the training set with probability of success higher than cut-off P_{crit} calculated using the optimal potential (—) and MJ potential (– – –).

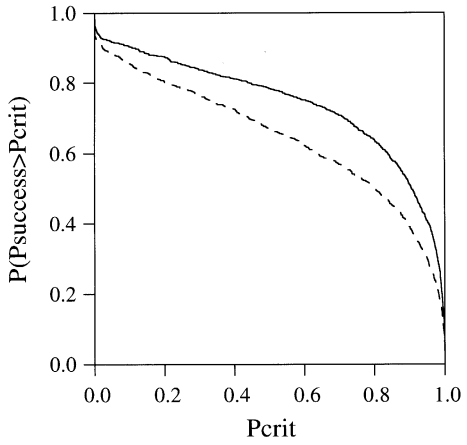

Fig. 4. The fraction of proteins in the test set with probability of success higher than cut-off P_{crit} calculated using the optimal potential (—) and MJ potential (– – –).

The purpose of our optimization procedure is to maximize our ability to predict the correct sequences of proteins of given structures. In order to compare optimal potential with MJ potential, we generated a second independent database of 1000 test proteins with their 991 different corresponding native states. Our potential yields the higher success rate of 64.0%, compared with a success rate of 48.7% for MJ potential. The fraction of the test set proteins with a probability of successful identification greater than any cut-off is shown in Figure 4. As can be seen, optimizing the average probability of success for the original training set of proteins increases the probability that proteins in the test set will have their sequences correctly predicted, compared with the use of the true energy function.

Not all amino acid sequences represent possible proteins. In order to be viable, proteins must be able to fold rapidly enough and be sufficiently stable to avoid irreversible processes such as aggregation and proteolysis. The set of lattice proteins described above, representing random sequences, is somewhat unrealistic in that the majority of them are unlikely to correspond to foldable proteins. Theoretical and computational work has suggested that the ability of a protein to successfully fold into its native state is dependent on the ratio between the energy gap separating the native state and the average of the non-native states and the standard deviation in the energies of these non-native conformations, corresponding to the original

Wolynes R-score (Goldstein *et al*., 1992a,b; Šali *et al*., 1994a,b). In order to construct a more biologically representative test of the various methods, we constructed a second data set of proteins consisting of proteins with R-scores greater than 5.0, a condition satisfied by only 0.3% of the random protein sequences. Using the same methods described above, we optimized our potentials for this second training database, and tested the potential on an equivalent test set, also consisting of proteins with R-scores greater than 5.0. These proteins have more stable native states, and are correspondingly easier to identify with the inverse folding procedure. Prediction accuracy using the exact MJ potential increased from 48.7 to 81.1%, while the optimized potential identified sequences corresponding to a given fold increased from 64.0 to 91.9%. Even in the case of proteins selected based on their ability to fold, the optimized potentials do better than the true potential.

Discussion

As shown in Figure 2, the potential generated by the optimization approach is largely correlated with the correct potential. As described in previous work, the tendency for the optimal potential to underestimate strongly destabilizing interactions is possibly the result of the use of a Gaussian approximation for the energies of the random sequences, and the resulting importance of thermodynamically-irrelevant high-energy structures (Chiu and Goldstein, 1998). Conversely, the differences between the optimal and true strongly attractive energy potentials is not observed in potentials optimized for fold recognition (Chiu and Goldstein, 1998), and thus is likely to represent a consequence of differences in the problems being addressed. For fold recognition, the correct energy potential would be the most effective; this is not true for the inverse folding problem.

As shown in Figures 3 and 4, the true potential had limited success in solving the inverse folding problem, indicating that the ability of increasingly accurate energy functions to solve this problem is limited. The true potential only tells us why one structure is preferred for a particular sequence, rather than indicating what sequences would most likely fold into a specific structure. We can achieve greater accuracies in the inverse folding problem by developing an optimized scheme so that the correct sequence would have the lowest potential relative to the other sequences.

In this paper, we find the optimal potential for inverse folding problem based on a set of database proteins. Specifically, we developed an expression to directly quantify the probability of success in predicting protein sequences. Our method is generalized to a larger heterogeneous database, containing some proteins that may be easier to predict than others. We maximized the average probability of success aiming to predict as many proteins as possible, providing more accurate energy parameters that are more likely to generate accurate predictions.

It is unlikely that we will soon have a general and comprehensive method for either protein structure prediction or inverse folding. Rather, we can concentrate on extending the number of proteins whose structure can be predicted. For the hardest proteins to predict with low Z-scores, where the probability of a successful prediction is small, increasing the Z-scores will not greatly increase the probability of a successful prediction. Conversely, in those 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 predictions 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.

Acknowledgements

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

Anfinsen,C. (1973) *Science*, **181**, 223–230.
Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) *Science*, **253**, 164–170.
Chiu,T.L. and Goldstein,R.A. (1998) *Folding Design*, **3**, 223–228.
Crippen,G.M. (1991) *Biochemistry*, **30**, 4232–4237.
Crippen,G.M. (1996) *Proteins*, **26**, 167–171.
Crippen,G.M. and Maiorov,V.N. (1995) In Bohr,H. and Brunak,S. (eds) *Protein Folds. A Distance-based Approach*. CRC Press, New York, pp. 189–201.
Doolittle,R.F. (1981) *Science*, **214**, 149–159.
Gill,P.E. and Murray,W. (eds) (1974) *Numerical Methods for Constrained Optimization*. Academic Press, London.
Godzik,A. (1995) *Protein Engng*, **8**, 409–416.
Goldstein,R.A., Luthey-Schulten,Z.A. and Wolynes,P.G. (1992a) *Proc. Natl Acad. Sci. USA*, **89**, 4918–4922.
Goldstein,R.A., Luthey-Schulten,Z.A. and Wolynes,P.G. (1992b) *Proc. Natl Acad. Sci. USA*, **89**, 9029–9033.
Govandarajan,S. and Goldstein,R.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5545–5549.
Hao,M.H. and Scheraga,H.A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 4984–4989.
Koretke,K.K., Luthey-Schulten,Z. and Wolynes,P.G. (1996) *Protein Sci.*, **5**, 1043–1059.
Maiorov,V.N. and Crippen,G.M. (1992) *J. Mol. Biol.*, **227**, 876–888.
Mirny,L.A. and Shakhnovich,E.I. (1996) *J. Mol. Biol.*, **264**, 1164–1179.
Miyazawa,S. and Jernigan,R.L. (1985) *Macromol.*, **18**, 534–552.
Shakhnovich,E.I., Farztdinov,G., Gutin,A.M. and Karplus,M. (1991) *Phys. Rev. Lett.*, **67**, 1665–1668.
Sippl,M.J. (1993) *J. Comp.-Aided Mol. Design*, **7**, 473–501.
Thomas,P.D. and Dill,K.A. (1996) *J. Mol. Biol.*, **257**, 457–469.
Šali,A., Shakhnovich,E.I. and Karplus,M.J. (1994a) *J. Mol. Biol.*, **235**, 1614–1636.
Šali,A., Shakhnovich,E.I. and Karplus,M.J. (1994b) *Nature*, **369**, 248–251.
Yue,K. and Dill,K.A. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 4163–4167.

Received December 24, 1997; revised March 11, 1998; accepted May 1, 1998