Determination of optimal Chebyshev-expanded hydrophobic discrimination function for globular proteins

by B. Fain
Y. Xia
M. Levitt

We describe the development of a scoring function designed to model the hydrophobic effect in protein folding. An optimization technique is used to determine the best functional form of the hydrophobic potential. The scoring function is expanded using the Chebyshev polynomials, for which the coefficients are determined by minimizing the Z-score of native structures in the ensembles of alternate conformations. (The Z-score is the score relative to the mean, measured in units of standard deviation.) The derived effective potential is tested on decoy sets conventionally used in such studies. The function is able to discriminate very well between correct and incorrect folds, despite the fact that it simply counts the number of neighbors of each amino acid. Our results show that the techniques of Z-score optimization and Chebyshev expansion work, and work well. Our results also confirm that hydrophobic effect is one of the principal driving forces in protein folding.

Introduction

A potential function which distinguishes native and native-like conformations from non-native structures is essential to protein structure prediction [1–10].

In this work we present three major ideas, each of which is used to design a potential function. First, we define a very simple effective potential for protein

0018-8646/01/$5.00 © 2001 IBM

structure prediction. This potential is designed to mirror the hydrophobic interaction and simply counts the number of neighbors of each amino acid. Second, we introduce a novel method of representing the shape of a potential which makes no assumptions about the functional form of the potential. Third, we utilize a recently developed procedure [11, 12] to train our potential by adjusting the parameters of the function to minimize the Z-score of sets of native structures with respect to alternate conformations. Finally, we evaluate our derived effective energy by applying it to decoys from the Decoys'R'Us [13] database.

We initially chose to model the hydrophobic interaction as a test for our procedures of Z-score optimization and Chebyshev polynomial expansion. We were surprised to discover that the effective potential came to possess significant discriminating power. The investigation of this power, which would illuminate the role of the hydrophobic effect in proteins, is important and interesting. Because we are already testing several new concepts in this paper, we relegate the investigation of the hydrophobic effect to another work.

Hydrophobic interaction

Every attempt to predict the three-dimensional structure of a protein's native state demands knowledge of the interaction potential between amino acids [2, 5, 14]. Since the classic work of Kauzmann [15] it has generally been believed [16] that one of the most important forces involved in folding is the hydrophobic effect. During the folding process, residues with charged and polar side chains remain exposed to the solvent, and those with hydrophobic side chains segregate into the interior of a globular protein [17, 18].

Series of experiments [19–22] as well as theoretical studies [16, 23–27] show evidence that the nonspecific interaction and placement of hydrophobic residues is a more critical determinant of protein structure than local sequence-dependent interactions. These results suggest that hydrophobic forces must be included in any theoretical expression of the conformation energy of a polypeptide. In addition, it may be possible to achieve significant discrimination between the ground state and alternate states of a polypeptide using only information about the arrangement of hydrophobic residues. Huang et al. [5] have shown that hydrophobic potentials alone can distinguish 99% of the correct folds for a suitably chosen set of alternatives, namely threaded decoys. Similar results have been reported by Cassari and Sippl [28].

In this paper we derive a hydrophobic potential which has significant power to discriminate among a challenging set of alternate conformations from the Decoys'R'Us (*http://dd.stanford.edu*) database [13, 29]. The novel aspect of our work is that in deriving an optimal hydrophobic potential, the only assumptions we make are that the hydrophobic interaction is residue-specific and that its strength depends on the accessible surface area of each amino acid. We let the optimization procedure dictate the very form of the interaction. The discrimination power of this rather simple potential compares favorably with potentials of much greater complexity.

Optimization strategies

There are currently two main approaches to extracting coarse-grained potentials between pairs of amino acids. The first approach, pioneered by Miyazawa and Jernigan [26], is based on the quasi-chemical approximation. It derives conformational energies by comparing the distributions of amino acids occurring in native structures of proteins to those of the random compact conformations. This approach has been used by many researchers [14, 30–33], and it was well reviewed by Sippl [34] as well as by Wodak and Roman [35].

The main flaw of potentials of mean force is that the quasi-chemical approximation may not be valid [36]. Recently Thomas and Dill [37] tested the method on exactly solvable lattice models. They showed that although the extracted and exact potentials do have common elements (which accounts for the current popularity of potentials of mean force), the two indeed do not correlate very well.

An alternate strategy was originally suggested by Maiorov and Crippen [38], and has since been the subject of considerable activity [11, 38–42]. The basic idea is to parameterize a suitably chosen Hamiltonian and then to adjust the parameters in such a way that a collection of native states assumes either the lowest or one of the lowest energies compared with an ensemble of incorrectly folded alternate structures. We use a variation of the second method to optimize our potential.

Theory

Simplified representation of the protein

It is currently believed that all-atom potentials are required to properly model the dynamics of protein folding [43]. The high level of detail combined with the (relatively) slow speed of today's computers limits the time scale over which we can follow the folding process to the order of one μs. Levitt [44] developed a now common solution to the burden of computational complexity: Avoid a detailed description of the amino acid by representing the side chain with a point approximating the centroid of the side chain. In our model the virtual side chain is a point

3.0 Å from the C^α along the C^α–C^β vector, where C^α and C^β refer to the alpha and beta carbons of the polypeptide chain. We position the Gly centroid at the C^α atom.

The model can be used [3, 5, 45] with either a fixed centroid distance or a sequence-dependent distance. Simplicity motivated our choice of a fixed distance, although performance does improve somewhat with a sequence-dependent centroid location.

Hydrophobic potential

When a hydrophobic residue is buried in the interior of the protein, it will necessarily have many neighboring residues. Viswanadhan [46] has shown that the average number of neighbors within 10 Å of a given residue correlates well with its hydrophobicity. For this reason, we assume that the energy contribution from each residue will depend on the number of residues within a 10-Å shell surrounding it. Explicitly,

$$E_i = E_a(n), \tag{1}$$

where a denotes a specific amino acid type, and n is the number of neighbors. We want to represent $E_a(n)$ as a linear combination of suitably chosen appropriate basis functions. Our choice of a basis differs from Crippen's [40]; we decided to represent our potential as a linear sum of Chebyshev polynomials (see Appendix A). Since the Chebyshev representation is most naturally applied to functions defined on the interval $[-1, 1]$, we transform the functional dependence of E on n as follows:

$$E_i = E_a\left(\frac{n-10}{n+10}\right). \tag{2}$$

The transformation $n \rightarrow (n - 10)/(n + 10)$ maps the possible number of near neighbors $[0, \infty]$ to an interval $[-1, 1]$. We chose 10 as the crossover point because 10 is roughly the number of neighbors with which an amino acid becomes buried. It is not necessary to choose this parameter exactly, but getting it in the right ballpark helps the Chebyshev expansion converge more rapidly.

The final functional form for the hydrophobic energy of a protein length N becomes

$$E_{\text{burial}} = \sum_{i=1}^{N} E_i = \sum_{i=1}^{N} \sum_{k} C_{a,k} T_k \left(\frac{n-10}{n+10}\right), \tag{3}$$

where a indexes the amino acid type, k is the order of the Chebyshev polynomial T_k, and n is the number of neighbors within a 10-Å radius of the amino acid i.

Because we want the resolution of our potential to be of order 0.1 (recall that the number of neighbors is an integer quantity), we choose to retain terms no higher than $O6$ in the Chebyshev expansion. (See the section on methods for a fuller explanation.) In addition, the 0th term is omitted because it is a sum of constants and contributes equally to any conformation of the same protein. We then have $20 \times 6 = 120$ coefficients $C_{a,k}$, which completely determine the potential.

Our representation of the potential has several highly desirable properties. First, it assumes nothing of the form of the interaction other than the fact that the energy depends on the amino acid type and its degree of burial. Optimization of the discriminating power of the potential will determine the very functional form of the interaction. Second, choosing the Chebyshev expansion allows us to represent the interaction with great accuracy using a very small number of parameters. This reduction in the number of parameters is paramount for any optimization scheme.

Z-score optimization

Our optimization scheme follows the general outline of the method of Mirny and Shakhnovich [11]. We choose a training set and construct alternate structures for each chosen protein. We then optimize the parameters of our Hamiltonian to minimize the average Z-score of the native structures relative to their corresponding alternates. (The details of the computation are described in the section on methods, and also in [12].) We decided to optimize the average Z-score and not the harmonic mean as in Reference [11] because during minimization the individual Z-scores may have different signs, thus rendering the harmonic mean useless. As soon as all of our Z-scores became negative, the minimizer switched to the harmonic mean. The final results for the two methods did not differ appreciably.

An alternate strategy would have been to insist that each of the native structures assumed the lowest energy with respect to the decoys [9, 47]. We prefer to minimize Z-score for several reasons. First, it is quite probable that despite our best efforts, the native conformations of some of the proteins in our training set are not the ground state of that particular sequence. Our optimization procedure allows non-native states to occasionally occupy states of lower energy than native. A related motive is to allow "near-native" states to be lower than the native conformation. At this stage of protein structure prediction development, we consider picking, say, a 1-Å structure from a set of decoys a success. A second advantage of using Z-score minimization is speed. The computational complexity of the optimization depends only on the number of parameters in the Hamiltonian and on the number of training sequences; it does not depend on the number of structures in each ensemble of the training set (see Appendix B).

Table 1 Proteins in the initial training set.

1a1x	1a32	1aep	1aho	1aie	1ail	1ako	1aly	1amx	1arb
1bb9	1bd8	1bea	1bfg	1bgf	1ble	1bm8	1bv1	1c25	1cby
1cex	1cfr	1chd	1cyw	1dun	1fna	1gpr	1gvp	1hoe	1hyp
1i1b	1ifc	1kid	1koe	1kte	1lcl	1lfb	1lki	1lxa	1mjc
1msc	1nkd	1noa	1pbv	1pdo	1pht	1pne	1ptf	1r69	1rcb
1rpl	1sfp	1tfe	1tig	1tlk	1tud	1tul	1utg	1vcc	1vie
1wer	1whi	1who	1xat	2end	2igd	2pii	2pth	2rn2	2tgi
3pte									

Results

Construction of training sets

We trained our function on a set of 71 sequences, listed in **Table 1**.

The structures were selected using three criteria. First, to ensure variety in our training set, the sequence identity of the structures was less than 35%. Second, to reinforce structural variation, each protein was chosen from a different SCOP (protein classification database) [48] family. Finally, to avoid structures of low quality, we kept only proteins with a SPASI (structure quality database) [49] score of ≤ 0.25.

Each member of our training set consists of the native structure and a set of one thousand alternate conformations ("decoys"). For each sequence we perturbed the native structure (starting the Monte Carlo trajectory with the native structure allows us to generate near-native decoys easily) with a simulated annealing routine [50, 51] and produced 1000 alternate conformations. We designed the decoy-generation procedure to produce decoys for which the root mean square deviation (RMSD) from native ranged from 0 to the radius of gyration (RG) of the native structure. In addition, the simulated annealing was designed to produce structures with RG similar to that of the native conformation, thus ensuring compactness.

We wanted to force the function to differentiate between native structure and alternate structures that have a relatively low (1–6) α-carbon (CA) RMSD from the native. Typically training sets are generated by threading onto diverse alternate structures [29]. It thus becomes exponentially difficult to produce decoys with near-native RMSDs [52]; as a result, the decoy set is "not challenging." We felt that just as threading decoys are considered insufficient for testing a potential, they should be equally inadequate for training one.

Training

The training procedure attempts to find the lowest average Z-score for all of the ensembles corresponding to the training proteins (listed in Table 1). The reader should consult Equation (B6) and surrounding paragraphs. The parameter space consists of $20 \times 6 = 120 c_k$. Because *a priori* we have no notion what the Z-score surface looks like, and also because 120 variables is a relatively large number, we chose to minimize with a simplex version of the simulated annealing procedure [50]. The temperature is decreased linearly from $\tau = 100$ (an arbitrary "large" value) to $\tau = 10^{-3}$. The annealing is restarted several times. At the end of the run, the result is refined with a downhill simplex method [51], which is equivalent to setting $\tau = 0$.

We cannot be certain that each of the proteins in our training set represents a true minimum of energy. Several things could go wrong: It is possible that co-factors unlisted in the PDB file were present either *in vitro* or *in vivo*; the shape of the molecule might have been significantly distorted by crystallization; or the molecule might also actually be a dimer or a multi-mer. To safeguard against these errors, we ensured self-consistency by removing from the training set all proteins that did not achieve a Z-score lower than -0.5. This value is high enough to allow small disulfide-rich proteins to be included in the training set. Three proteins were removed from the initial set; our final training set consists of 68 proteins.

The final average Z-score for the training set is -3.48. **Figure 1** is a plot of the energy function for three representative amino acids, showing the burial preferences for hydrophobic valine, hydrophilic arginine, and intermediate glycine. Val prefers many neighbors, while Arg would rather have only a few neighbors. (Recall that our reference state consists only of compact structures; thus, even a hydrophilic residue must have neighbors.) Finally, the burial preferences of Gly are between those of Arg and Val. When examining Figure 1, the reader should note that energy values for fewer than two neighbors and 30 or more neighbors are arbitrary. All residues have at least two neighbors, and, because of their excluded volume, no residues have more than 50 neighbors.

Discrimination power

Recently, the evaluation of functions has been made convenient by a database of decoy sets [13] (*http://dd.stanford.edu*). We tested our potential on two families

of decoy sets. The first, also known as the Park–Levitt set [29], was produced by perturbing the loop degrees of freedom of the native structure and then selecting protein-like conformations. This set has been used in several comparisons of potentials [29, 32, 53]. The second family of decoys was produced by Kesar and Levitt by using minimization with a complex potential which contains a significant pairwise component. We selected the K–L family because pairwise potentials are easily deceived by this set, possibly because each member of the set is a local minimum of a pairwise potential. The performance of our potential is summarized in **Tables 2** and **3**. For comparison we also display the Z-scores of two other potentials. Shell is a Myazawa–Jernigan [26] pairwise contact potential, and DB is a complex potential reported by Simons et al. [53]. We chose the Shell potential because, although slightly dated, its basic features are the same as those of most modern potentials [32]. The DB potential was chosen because, considering its performance in the latest CASP, it is one of the best discriminating functions available. Tables 2 and 3 show the Z-scores for the three potentials, as well as the rank of the native and the first near-native conformations in each decoy set. (We define "near-native" structures to be closer than 3.5 Å RMSD from the native fold.) We do not know the performance of the scoring function of Simons et al. on the Kesar–Levitt set because the scoring function was not available at the time of this writing.

The performance of our burial function is surprisingly strong. Our potential counts only the number of neighbors around each residue. The Shell potential does the same, but each neighboring amino acid brings a different contribution (in other words, it is pairwise residue specific). The DB potential is very complex, containing pairwise components of different burial classes, a secondary structure packing term, and other contributions. However, we manage to achieve Z-scores which are very close to those of our colleagues. Even more significantly, we achieve excellent discrimination of near-native decoys, which is crucial for structure prediction.

The Kesar–Levitt set unveiled further surprises. Shell does predictably badly on the set because the decoys were minimized with a potential which has a pairwise component. The performance of our function, however, remains similar to its performance on the P–L decoy set. We were unable to evaluate the discrimination of near-natives because the decoy sets in the K–L family do not have near-native structures.

Two proteins presented us with some difficulty: 2cro and 4pti. The latter has three disulfide bonds which stabilize it and probably reduce the protein's tendency to be hydrophobically stable. The reason for the stubbornness of 2cro is unclear. We can take partial

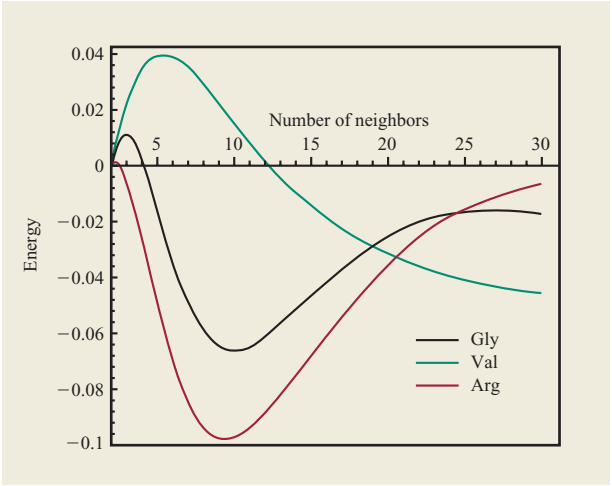

Figure 1

Burial preferences for Val, Arg, and Gly. Regions containing fewer than two neighbors and more than 30 neighbors do not contribute to discrimination.

Table 2 Performance of scoring functions on the Park–Levitt decoy set. The set contains approximately 700 decoys of each protein. The three elements separated by colons in the right-hand columns are *native rank* : *near-native rank* : *Z-score*.

Protein	*Burial*	*Shell (M–J)*	*Simons et al.*
1ctf	1 : 1 : −2.9	13 : 1 : −2.3	1 : 1 : −3.5
1r69	2 : 1 : −2.4	4 : 2 : −2.4	1 : 7 : −2.0
1sn3	3 : 1 : −2.1	14 : 1 : −2.1	1 : 1 : −2.2
2cro	60 : 1 : −1.3	19 : 1 : −1.8	371 : 6 : −1.2
3icb	10 : 1 : −1.6	3 : 1 : −2.5	1 : 2 : −2.8
4pti	62 : 9 : −1.7	5 : 2 : −2.3	21 : 23 : −2.3
4rxn	6 : 2 : −2.2	2 : 1 : −2.8	21 : 3 : −2.8
ave	−2.0	−2.3	−2.4

Table 3 Performance of scoring functions on the Kesar–Levitt decoy set. The set contains 500 decoys of each protein. Most of the proteins in this collection do not have near-native decoys. The three elements separated by colons in the right-hand columns are *native rank* : *near-native rank* : *Z-score.*

Protein	*Burial*	*Shell (M–J)*
1bba	4 : 10 : −2.5	500 : 9 : +2.9
1ctf	1 : n/a : −3.5	16 : n/a : −1.9
1dtk	34 : n/a : −0.9	26 : 1 : −1.2
1fc2	9 : n/a : −2.1	321 : n/a : +0.4
1igd	1 : n/a : −2.8	10 : n/a : −2.0
2cro	250 : n/a : −0.4	46 : 2 : −1.3
2ovo	7 : n/a : −2.2	246 : n/a : +0.6
4pti	32 : n/a : −1.4	248 : 1 : +0.7
smd3	1 : n/a : −5.0	1 : n/a : −4.1
ave	−2.3	−0.7

solace in the fact that other potentials have difficulty with this protein as well.

Conclusion

Our initial aim was to test two procedures, Z-score optimization and the Chebyshev expansion of the potential, in a simulated predictive environment. We also wanted to see how well the functional form in Equation (2) captured the burial propensities of the various amino acids. This we have accomplished. In addition, our optimized potential, despite its simplicity, performed with considerable strength and consistency. This positive result is likely brought about by two factors. First, the hydrophobic effect plays a significant role in protein folding. In addition, our optimization procedure, the Chebyshev approximation, or the form of the potential capture the hydrophobic effect well.

The promising results of this work suggest clear avenues for further investigation. We are currently working on other components of the energy function, and we are trying to generate better training sets. The significant role of the hydrophobic effect also deserves further inquiry.

Appendix A: Chebyshev expansion

An excellent exposition of the Chebyshev expansion can be found in Chapter 5 of Reference [54]. We briefly restate some of the more useful properties of the approximation. The Chebyshev polynomial of degree n is defined by

$$T_n(x) = \cos(n \arccos x). \tag{A1}$$

Explicitly, the polynomials are the following:

$$\begin{aligned}
&T_0(x) = 1;\\
&T_1(x) = x;\\
&T_2(x) = 2x^2 - 1;\\
&T_3(x) = 4x^3 - 3x;\\
&T_4(x) = 8x^4 - 8x^2 + 1;\\
&\quad \ldots\\
&T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) n \geq 1.
\end{aligned} \tag{A2}$$

Although the T_n are defined only on the interval $[-1, 1]$, a simple change of variable allows the expansion to be used to represent a function between two arbitrary limits, $[a, b]$:

$$y = \frac{x - \frac{1}{2}(b + a)}{\frac{1}{2}(b - a)}. \tag{A3}$$

There are two main reasons for picking the Chebyshev expansion. The first is that the error is spread out smoothly over the approximated interval. In fact, the Chebyshev approximating polynomial is very nearly the same as the minimax polynomial. The second is that the Chebyshev approximation achieves a very low error for relatively few terms, thus leaving us fewer parameters to deal with.

Appendix B: Z-score optimization

Our optimization scheme minimizes the average Z-score of the training set native structures relative to their respective decoys. For each ensemble the Z-score is defined as

$$Z \equiv \frac{V(0) - \langle V(i) \rangle}{\sigma}, \tag{B1}$$

where the $\langle\ \rangle$ and σ are, respectively, the ensemble average and standard deviation, 0 denotes the native conformation, and $0 \leq i \leq N$ runs over all of the conformations in the ensemble.

Because our potential function is a linear combination of terms,

$$V(i) = \sum_{k=0}^{k_{\max}} c_k V_k(i), \tag{B2}$$

where the c_k are the parameters we are optimizing, we can achieve significant simplification. We substitute Equation (B2) into Equation (B1). When the order of summation on i and k is interchanged, the numerator of Equation (B1) becomes

$$\sum_{k=0}^{k_{\max}} c_k \left[V_k(0) - \frac{1}{N} \sum_{i=0}^{N} V_k(i) \right]. \tag{B3}$$

Next we consider the denominator, which is the square root of the variance. The variance of a linear sum can be decomposed as follows:

$$\operatorname{var}\left(\sum_{k=0}^{k_{\max}} c_k V_k \right) = \sum_{m=0}^{k_{\max}} \sum_{n=0}^{k_{\max}} c_m c_n \operatorname{cov}(V_m, V_n), \tag{B4}$$

in which the ensemble covariance matrix is defined as

$$\operatorname{cov}(V_m, V_n) \equiv \langle (x_m - \mu_m)(x_n - \mu_n) \rangle, \tag{B5}$$

where μ is the ensemble mean. Putting Equations (B3) and (B4) together, we get

$$Z = \frac{\sum_{k=0}^{k_{\max}} c_k \left[V_k(0) - \frac{1}{N} \sum_{i=0}^{N} V_k(i) \right]}{\sqrt{\sum_{m=0}^{k_{\max}} \sum_{n=0}^{k_{\max}} c_m c_n \operatorname{cov}(V_m, V_n)}}. \tag{B6}$$

The value of Equation (B6) is that one can precalculate the actual basis functions V_k and the covariance matrix for each ensemble. Consequent adjustment of the parameters

c_k requires us to simply perform matrix and vector multiplication.

Acknowledgments

Many thanks to the present and former members of the Levitt group for help and discussions. We wish to thank Patrice Koehl for providing excellent routines for constructing structures from dihedral angles. B. F. wishes to thank the A. P. Sloan Foundation and the U.S. Department of Energy for financial support. Y. X. is a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported in part by Grant No. DE-FG03-95ER62135 to M. L. from the U.S. Department of Energy.

References

1. A. Bauer and A. Beyer, *Proteins: Struct. Funct. Genet.* **18,** 254 (1994).
2. J. Bowie, R. Luthy, and D. Eisenberg, *Science* **253,** 164 (1991).
3. S. Bryant and C. Lawrence, *Proteins: Struct. Funct. Genet.* **16,** 92 (1993).
4. R. Goldstein, Z. Luthey-Schulten, and P. Wolynes, *Proc. Natl. Acad. Sci. USA* **89,** 9029 (1992).
5. E. Huang, S. Subbiah, and M. Levitt, *J. Mol. Biol.* **252,** 709 (1996).
6. T. Jones, W. Taylor, and J. Thornton, *Nature (Lond.)* **358,** 86 (1992).
7. M. Levitt, *J. Mol. Biol.* **170,** 723 (1983).
8. C. Ouzounis, C. Sander, M. Sharf, and R. Schneider, *J. Mol. Biol.* **232,** 805 (1993).
9. F. Seno, A. Maritan, and J. Banavar, *Proteins: Struct. Funct. Genet.* **30,** 224 (1998).
10. R. Srinivasan and G. Rose, *Proteins: Struct. Funct. Genet.* **22,** 81 (1995).
11. L. Mirny and E. Shakhnovich, *J. Mol. Biol.* **264,** 1164 (1996).
12. Y. Xia and M. Levitt, *J. Chem. Phys.* **113,** 9318 (2000).
13. R. Samudrala and M. Levitt, *Prot. Sci.* **9,** 1399 (1999).
14. M. Sippl, *J. Mol. Biol.* **213,** 859 (1990).
15. W. Kauzmann, *Advan. Prot. Chem.* **14,** 1 (1959).
16. K. Dill, *Biochemistry* **29,** 7133 (1990).
17. M. Perutz, J. Kendrew, and H. Watson, *J. Mol. Biol.* **13,** 669 (1965).
18. G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus, *Science* **229,** 834 (1985).
19. S. Kamtekar, J. Shiffer, H. Xiong, J. Babik, and M. Hecht, *Science* **262,** 1680 (1993).
20. W. Lim and R. Sauer, *Nature (Lond.)* **339,** 31 (1989).
21. D. Sali, M. Bycroft, and A. Fersht, *J. Mol. Biol.* **220,** 779 (1991).
22. D. Shortle and A. Meeker, *Biochemistry* **29,** 8033 (1990).
23. P. Argos, *J. Mol. Biol.* **197,** 331 (1987).
24. B. Cohen, S. Presnell, and F. Cohen, *Prot. Sci.* **2,** 2134 (1993).
25. A. Lesk and C. Chothia, *J. Mol. Biol.* **136,** 225 (1980).
26. S. Miyazawa and R. Jernigan, *Macromolecules* **18,** 534 (1985).
27. R. Sweet and D. Eisenberg, *J. Mol. Biol.* **171,** 479 (1983).
28. G. Cassari and M. Sippl, *J. Mol. Biol.* **224,** 725 (1992).
29. B. Park and M. Levitt, *J. Mol. Biol.* **258,** 367 (1996).
30. M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. Sippl, *J. Mol. Biol.* **216,** 167 (1990).
31. S. Miyazawa and R. Jernigan, *J. Mol. Biol.* **256,** 623 (1996).
32. R. Samudrala and J. Moult, *J. Mol. Biol.* **275,** 895 (1998).
33. S. Tanaka and H. Scheraga, *Macromolecules* **9,** 945 (1976).
34. F. Sippl, *Curr. Opin. Struct. Biol.* **5,** 229 (1995).
35. S. Wodak and M. Roman, *Curr. Opin. Struct. Biol.* **3,** 247 (1993).
36. A. Ben-Naim, *J. Chem. Phys.* **107,** 3698 (1997).
37. P. Thomas and K. Dill, *J. Mol. Biol.* **257,** 457 (1996).
38. V. Maiorov and G. Crippen, *J. Mol. Biol.* **227,** 876 (1992).
39. T. Chiu and R. Goldstein, *Folding & Design* **3,** 223 (1998).
40. G. Crippen, *J. Mol. Biol.* **260,** 467 (1996).
41. G. Crippen, *Folding & Design* **1,** S58 (1997).
42. R. Goldstein, Z. Luthey-Schulten, and P. Wolynes, *Proc. Natl. Acad. Sci. USA* **89,** 4918 (1992).
43. W. van Gunsteren, *Computer Simulation of Biomolecular Systems*, ESCOM Science Publishers, B.V. Leiden, Netherlands, 1989.
44. M. Levitt, *J. Mol. Biol.* **104,** 59 (1976).
45. K. Simons, C. Kooperberg, E. Huang, and D. Baker, *J. Mol. Biol.* **268,** 209 (1997).
46. V. Viswanadhan, *Int. J. Biol. Macromol.* **9,** 39 (1987).
47. J. Mourik, C. Clementi, A. Maritan, F. Seno, and J. Banavar, *J. Chem. Phys.* **110,** 10123 (1999).
48. A. Murzin, S. Brenner, and T. Hubbard, *J. Mol. Biol.* **247,** 536 (1995).
49. S. Brenner, P. Koehl, and M. Levitt, *Nucl. Acids Res.* **28,** 254 (2000).
50. N. Metropolis, M. Rosenbluth, A. Teller, and E. Teller, *J. Chem. Phys.* **21** (1953).
51. J. Nelder and R. Mead, *Comput. J.* **7,** 308 (1965).
52. B. Reva, A. Finkelstein, and J. Skolnik, *Folding & Design* **3,** 141 (1998).
53. K. Simons, I. Ruczinski, C. Kooperberg, B. Fox, C. Bystroff, and D. Baker, *Proteins: Struct. Funct. Genet.* **34,** 82 (1999).
54. W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge University Press, New York, 1992.

Received December 18, 2000; accepted for publication January 31, 2001

Boris Fain *Department of Structural Biology, Stanford University, Stanford, California 94305 (bfain@stanford.edu).* Dr. Fain is a Research Fellow in the Levitt Laboratory, Department of Structural Biology, Stanford University. He received his B.A. degree in mathematics from the University of California at Berkeley in 1990, and his Ph.D. in physics from UCLA in 1997. He came to Stanford as a Sloan/DOE Fellow in 1997 to work on protein folding.

Yu Xia *Department of Chemistry and Department of Structural Biology, Stanford University, California 94305 (yuxia@csb.stanford.edu).* Mr. Xia is a Ph.D. student in the Levitt Laboratory working on protein structure prediction. He received his B.S. degree in chemistry with a minor in computer science from Beijing University in 1995. He is a recipient of the Howard Hughes Predoctoral Fellowship.

Michael Levitt *Department of Structural Biology, Stanford University, Stanford, California 94305 (michael.levitt@stanford.edu).* Dr. Levitt has worked on computational biology for more than thirty years, focusing on physical simulation, structure prediction, and large-scale sequence-structure comparisons. He received his B.Sc. degree in physics from King's College, London, in 1967 and his Ph.D. from Cambridge University in 1972.