Journal of Biological Physics **30:** 171–185, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

Toward Correct Protein Folding Potentials

M. CHHAJER[1] and G.M. CRIPPEN[2]
[1] *Department of Chemistry, University of North Carolina, Chapel Hill, NC 27599, U.S.A.*
[2] *College of Pharmacy, University of Michigan, Ann Arbor, MI 48109-1065, U.S.A.*

Abstract. Empirical protein folding potential functions should have a global minimum near the native conformation of globular proteins that fold stably, and they should give the correct free energy of folding. We demonstrate that otherwise very successful potentials fail to have even a local minimum anywhere near the native conformation, and a seemingly well validated method of estimating the thermodynamic stability of the native state is extremely sensitive to small perturbations in atomic coordinates. These are both indicative of fitting a great deal of irrelevant detail. Here we show how to devise a robust potential function that succeeds very well at both tasks, at least for a limited set of proteins, and this involves developing a novel representation of the denatured state. Predicted free energies of unfolding for 25 mutants of barnase are in close agreement with the experimental values, while for 17 mutants there are substantial discrepancies.

Key words: barnase, decoys, denatured state, protein folding, thermodynamic stability

1. Statistics and Decoys fail to Stabilize the Native Conformation

One way to model protein folding is to start with a detailed all-atom representation of the protein and solvent, calculate the potential energy of any configuration with a classical empirical energy function based on the customary classification of energy terms [20], and then estimate free energies by extremely lengthy molecular dynamics or Monte Carlo simulations having poorly understood convergence properties and in any case falling orders of magnitude short of simulating the experimentally known time spans for such events. In the most ambitious undertaking of this sort to date [46], a fraction of the molecular dynamics trajectories of a 36-residue protein start at an extended conformation and reach conformations resembling the native in an average sense. Here we are concerned with the alternative approach of a simplified representation of the protein, an implicit solvent model, and a much more empirical potential that includes rapid and small spatial scale entropic effects implicitly as a function of the explicit large scale conformation of the protein. The goal is to establish stability of the native state under appropriate conditions and to reproduce the experimental thermodynamics of unfolding.

One approach to devising such an empirical potential involves an extensive survey over known protein crystal structures. There have been many variations on this theme [12, 25, 26, 16, 37, 36, 19], and resulting potentials are known to be incorrect in some sense [39, 40], but they do give worse (higher) values for impressive numbers of incorrect folds compared to those for hundreds of corresponding correct native conformations. We will refer to this as the structure recognition task. While these potentials have no adjustable parameters *per se*, they do depend on the procedure and on the proteins surveyed. One might be concerned that the outcome depends on some inconsequential details rather than general underlying principles of protein folding. As an example of this type of potential, we will consider the more recent potential by Miyazawa and Jernigan [26], denoted by MJ. It was developed from a training set of 1168 proteins, and has been extensively tested against threaded decoys, which are contiguous pieces of chain taken from larger crystal structures and assigned the residue types of the native protein in question.

In optimization-based methods, native and nonnative (or decoy) structures are compared for a training set of proteins, and the parameters in the potential are explicitly adjusted to improve performance in structure recognition according to some objective. Once again, there are many variations in functional form, representation of the protein, and optimization objective [24, 39, 13, 21, 43, 44, 41, 42, 10, 27, 28, 8, 32, 3, 29]. Concerns about overfitting are aroused by such extreme examples as 80,000 adjustable parameters [29]. A more subtle concern is alternative methods of generating especially challenging sets of decoys [43, 27, 23, 45, 17, 5, 28]. As an example of this type of potential, we will consider that of Tobi and Elber [41], denoted by TE. It was developed from a training set of 572 proteins; training and testing decoys are mostly from threading but also some generated by the MONSSTER molecular dynamics program.

While MJ and TE are outwardly very successful at structure identification, we have found that adjusting potential functions so that there is even a local minimum near the native conformation is quite challenging [27]. First we checked that our coding of these potentials from the published descriptions perform satisfactorily in ungapped threading. (a) We selected a set of monomeric proteins from the original set of proteins used by the authors. For the TE potential, 55 proteins were selected; for the MJ potential, a set of 53 proteins was selected that contained 28 of their test proteins. (b) The geometric center for each side chain in each protein was determined directly from PDB atomic coordinates. These coordinates represent the position of the corresponding residues in the chain as used by the authors during their training and/or testing process. (c) Using this representation, a threading test was performed and the rank of each native conformation was determined. Decoys for each of the n proteins in the set are obtained by ungapped threading through the n-1 others. For the TE potential, 54 native conformations were ranked 1. For the MJ potential, of the 28 test proteins, Miyazawa and Jernigan report rank 1 for 27 of them whereas we got 25 of rank 1. Of the remaining 25 proteins, 23 native conformations were ranked 1 in our case while no threading test results are reported by the authors.

In order to minimize the potential functions with respect to conformation, we needed to fit the whole set of proteins to standard local geometry [27], and while

Table I. Monte Carlo minimization from the native conformation with Tobi and Elber potential[a]

Protein	Native energy	Maximum rmsd (Å)	Minimum energy[b]	No. distant trajectories[c]
1bdo	–265.3	11.8	–329.2	3
1mai	–353.1	13.1	–397.4	17
1pdo	–551.6	4.5	–581.5	0
1skz	–447.4	15.8	–508.0	85
1whi	–538.5	11.0	–576.0	9
2sns	–339.3	17.7	–344.6	28
7rsa	–306.1	12.8	–340.0	7

[a] Tobi and Elber [41].

[b] Energy of the final conformation in the MC simulation corresponding to the maximal rmsd from the native.

[c] Number of Monte Carlo trajectories out of 100 total resulting in rmsd > 6Å.

that is generally a very close fit, threading might not work as well as with the direct PDB coordinates. Since the fitting procedure produces coordinates only for the backbone heavy atoms and the C_β side chain atoms, we take the side chain center to be the point along the C_α–C_β vector at a distance from the C_α given by Hinds and Levitt [16], or just the C_α for Gly. (d) Using the fitted conformations for the natives as well as for the decoys, 34 of the natives had rank 1 for the TE potential and 44 (23 from the test set) for the MJ potential. For both the potentials, proteins which are ranked 1 in this test are also ranked 1 in the previous test.

Using seven proteins for each potential having rank 1 natives in both PDB and fitted threading tests, we conducted a Monte-Carlo energy minimization with respect to backbone torsion angles. Starting at the fitted native conformation, torsion angles were randomly perturbed by at most $\pm 20°$, and the new conformation was accepted only if the value of the potential function decreased, as in Metropolis Monte-Carlo at 0 K. The intent was to find a rather continuous downward trajectory of 50,000 steps on the energy surface, in spite of the potentials being discontinuous. For TE, the seven proteins in Table I passed threading tests against 1000 to 3340 decoys, yet several trajectories out of 100 for each protein went far from the native, extremely far in the worst case for each, except for 1pdo. While using the TE potential, no two atoms are allowed to come closer than 2.5 Å. For MJ, the seven proteins in Table II passed threading tests against 2800 to 5600 decoys, but almost all of the 30 trajectories for each went at least 6 Å from the native and much further than that in the worst case. While it is true that part of the reason for such unfolding may be the absence of hard core repulsion in MJ potential, it cannot explain the highly open state the native runs to.

Table II. Monte Carlo minimization from the native with Miyazawa and Jernigan potential[a]

Protein	Native energy (kT)	Maximum rmsd (Å)	Minimum energy[b] (kT)	No. distant trajectories[c]
1fkf	–660.7	17.1	–701.1	28
1fxd	–343.6	11.3	–427.0	22
1paz	–746.8	15.8	–858.4	28
1ycc	–579.2	16.2	–693.0	29
3b5c	–464.8	13.9	–606.0	28
5rxn	–300.9	10.9	–372.7	21
7rsa	–653.6	17.2	–680.7	29

[a] Miyazawa and Jernigan [26].

[b] Energy of the final conformation in the MC simulation corresponding to the maximal rmsd from the native.

[c] Number of Monte Carlo trajectories out of a total of 30 resulting in rmsd $>$ 6Å.

Clearly success with threading decoys is a much weaker achievement than using energy minimized decoys. In contrast, by concentrating on energy minimized decoys, we are able to construct a potential function [2] on the basis of six training proteins such that there is a local minimum within about 4 Å rmsd of the experimental native conformation for 42 proteins and within 6 Å for 71 proteins out of a total of 91 test and training proteins. Furthermore, for 89 of these 91 proteins, the native is ranked best in an ungapped threading test; the native was second best for the other two.

2. Robust Prediction of Free Energy of Unfolding

Going beyond the structure recognition problem into modeling folding thermodynamics, much less kinetics, introduces a whole new set of requirements for potential functions. Consider simply the demand that under native conditions (temperature, pH, solvent composition) the free energy of the native state (nat) should be lower than that of the denatured state (den), i.e., $\Delta G = G_{den} - G_{nat} > 0$. Success at structure recognition does not imply thermodynamic stability and vice versa [8]. Even assuming we can treat folding as a two-state process, the native and denatured states must be modeled as ensembles of conformations. Does the denatured state contain a significant amount of native folds, only a small amount of native folds, or none at all, and if the last case, then is it in the random coil state or some other state? Experimental evidence for different proteins is mixed [15], and sometimes evidence for even the same protein from two different investigators is contradictory (e.g., for barnase, see Freund et al. [11] and Takei et al. [38]).

One way to deal with disordered states is by enumerating all conformations in lattice models, either 2-D [31] or 3-D [34]. An off-lattice model [7] for the two-state denaturation of the 110 residue protein, barnase, estimates the chain entropy of the two states from lattice studies [6], and an all heavy atom potential for structure recognition [8] was modified to include solvation effects of urea so as to fit the urea denaturation at 298 K for wild-type barnase and 66 mutants. To deal with thermal denaturation, one additional parameter was required to adjust a model of density of microstates within the native and denatured states. This then reproduced the ΔG, ΔH, ΔS, T_m, and $C_p(T)$.

Another way to at least estimate ΔG of unfolding is to assess the fine details of packing, hydrogen bonding, and a number of other factors in the given crystal structure of the native protein, and use an empirically weighted sum of these factors to calculate the free energy of unfolding [14]. By this method, 1A2P.A, a high-resolution crystal structure of (wild-type) barnase, gives ΔG = 45.3 kJ/mol, in reasonable agreement with experiment (see Table III). However, if all the atomic coordinates are perturbed by adding independent random numbers uniformly distributed between 0 and 0.1 Å, the estimate changes to 39.8 kJ/mol. If the perturbations are of magnitude 0.5 Å, which is still within the resolution of the crystal structure, the estimate can change to –19.0 kJ/mol, i.e., the native is claimed to be unstable at room temperature! Clearly, the method is very sensitive to small perturbations in structure.

The motivation for this work is to obtain a potential function that is thermodynamically more robust and consistent than the more conventional approaches where the fold recognition problem and structural stability problems are treated separately. This results in two largely uncorrelated potential functions even though the two problems are highly correlated. Here, we propose a strategy to develop a potential function which is thermodynamically more consistent by treating these two problems simultaneously, as is the case in a natural process. Furthermore there is only one additional parameter needed to adjust the energy scale.

3. The Free Energy of Unfolding Via an Explicit Denatured Ensemble

We use barnase and its mutants as our model system. Barnase is 110 residues long, folds as a monomer in a two-state transition, has no disulfide bridges or large ligands, and contains all residue types but Cys and Met. The ΔG of unfolding at 298 K for barnase and its 65 mutants have been obtained by urea denaturation [33] though the crystal structures have been determined for only the wild-type and 11 of the mutants. Even though most of the mutants have been changed only at one position out of 110 possible places, there is a significant change in ΔG values, as shown in Table III. At 298 K, ΔG is 36.9 kJ/mole for the wild-type but is only 17.9 kJ/mole for the mutant where leucine has been replaced by alanine at the 14th position (L14A), thus reducing its stability by more than 50%. This is a rather surprising result even though residue 14 is in the middle of α-helix-1 and is a part of the biggest core in the folded state. None of the currently available potential

Table III. Free energy of unfolding for barnase and its mutants

protein[a]	ΔG_{exp} (kJ/mole)[b]	ΔG_{cal} (kJ/mole)
Wild[c]	36.9	36.9
I88A[c]	20.1	19.6
I88V[c]	30.0	30.5
S91A[c]	26.8	26.2
T26A[c]	29.9	30.6
Y78F[c]	32.1	32.2
L89V[c]	34.9	34.5
L14A[c]	17.9	18.3
I76A[d]	30.0	26.6
I76V[d]	32.8	32.9
I96A[d]	23.7	25.8
I96V[d]	32.6	32.5
I04V	33.5	29.5
I04A	33.8	40.0
N05A	30.2	10.3
T06G	34.3	35.2
T06A	28.4	35.9
D08A	33.3	37.0
V10T	28.6	13.1
V10A	22.4	32.6
D12A	35.3	37.1
Y13A	24.3	40.3
T16S	31.2	35.4
T16R	40.0	38.5
Y17A	29.9	36.9
N23A	27.7	28.8
Y24F	37.3	34.2
I25V	33.4	37.9
I25A	21.5	36.4
T26G	32.7	37.0
K27G	35.3	38.0
E29G	28.3	35.0
Q31S	37.2	36.5
S31A	36.6	36.6
L33Q	31.0	35.9
V36A	29.9	39.3
V36T	32.5	36.6
N41D	25.9	32.5
V45A	32.0	34.1

Table III. Continued

protein[a]	ΔG_{exp} (kJ/mole)[b]	ΔG_{cal} (kJ/mole)
V45T	27.6	23.5
I51V	32.2	30.1
I51A	17.2	28.0
D54A	24.9	37.9
D54N	29.4	27.5
I55V	34.3	28.7
I55A	30.7	29.5
N58A	28.5	28.9
N58D	38.2	25.4
V55T	31.7	24.5
K62R	35.6	30.9
N77A	30.8	21.0
N84A	30.2	26.3
V89T	23.2	–8.0
S92A	23.4	22.1
T99V	24.7	31.9
Y103F	36.1	36.8
T105V	27.9	28.0
I109V	33.5	33.2
I109A	31.6	21.3
R110A	37.2	19.4
D8A_D12A	31.0	37.3
D8A_R110A	32.2	20.9
D8A_D12A_R110A	37.6	23.1
D12A_R110A	38.5	21.6
Y13A_Y17A	20.3	45.4
T16A_Y17A	28.9	30.0

[a] XnnY represents a mutant where X in the *nn* position in the wildtype barnase has been replaced by Y.
[b] Serrano et al. [33].
[c] Training set.
[d] Test set with known experimental crystal structure.

functions can account for such dramatic changes in ΔG and at the same time place the native conformation at the global minimum of the potential function.

The crystal structures [1] of barnase and its 11 mutants for which the experimental crystal structures are available are fitted [27] to a standard geometry continuous state model [9] where each residue is represented by five interacting sites (united atom types), located at the C_β, C_α, N, C, and O atoms. All peptide bonds are kept in the trans conformation so only the 220 (ϕ, Ψ) backbone torsion angles are allowed to vary. The fitted model is within 0.5 Å rmsd in C_α coordinates [18] of the PDB structure and is used as the native structure in our calculations. Thus, the native state is a single energy, non-degenerate state in our calculations. For the remaining 54 barnase mutants, the fitted wild-type structure is used for their native states.

The potential function $E = \sum e_{ij}$ is a sum over pairwise interaction terms of the form

$$e_{ij}(d_{ij}) = \left[\frac{\cos(C_1 d_{ij})}{(C_0 + C_1 d_{ij})^{12}} + A_{ij} \cos(C_3 d_{ij}) + B_{ij} \cos(C_5 d_{ij}) \right] . \frac{(d_{max} - d_{ij})^2}{d_{max}^2} \quad (1)$$

where the first term takes into account the steric repulsion part, and the amplitude of the other two terms, A_{ij} and B_{ij}, are the adjustable parameters and control the depth and width of the potential well. This functional form allows significant flexibility in the shape of the trained energy function while keeping the number of adjustable parameters for a distance dependent interaction energy to just two per interaction. This is in contrast to the TE potential which has 13 adjustable parameters per interaction. We have taken the maximum interaction distance to be 15 Å = d_{max} beyond which $e_{ij} = 0$. The values of constants C_0, C_1, C_3 and C_5 are fixed and the same for all interactions, set at $C_0 = 0.56234133$, $C_1 = \pi/24$, $C_3 = 3\pi/24$ and $C_5 = 5\pi/24$. These values of the constants ensure that there is at most one minimum in the distance range of 0-12 Å from each of the cosine terms. Furthermore, the second and third terms in the bracket are orthogonal to each other in this range. Note that e_{ij} is a linear function of the adjustable parameters A_{ij} and B_{ij}, is continuous and differentiable for $0 \leq d_{ij} \leq 15$, $e_{ij}(0) = 1000 + A_{ij} + B_{ij}$, $e_{ij}(12) = e_{ij}(15) = 0$, and $\left.\frac{de_{ij}}{dd_{ij}}\right|_{d_{ij}=15} = 0$. We consider 24 atom types: 20 side chain atom types representing 20 different amino acids and the four main chain heavy atoms. Hence there are 300 different types of interactions and a total of 600 adjustable parameters.

Using our earlier protocol [2], initially the parameters were adjusted by quadratic programming [4] to minimize $\sum_{i,j}(A_{ij}^2 + B_{ij}^2)$ subject to $-25 \leq A_{ij}, B_{ij} \leq 50$ and

$$\Delta E = E_{non} - E_{nat} > \begin{cases} 0.3 & \rho > 0.3 \\ \rho & \rho < 0.3 \end{cases} \quad (2)$$

where ρ is a scaled measure of conformational similarity [22] between the native and a decoy, and $\rho = 0.1$ means very similar. The restricted ranges of A_{ij} and B_{ij} are used to reduce the potential surface roughness by keeping these values sufficiently small. Native sequences and structures are the first eight proteins listed in Table III. Including some of the mutants in the training set is necessary because training only with the wild type produces poor results on the mutants. Decoys were generated by threading and by random perturbation of dihedral angles followed by repeated minimization of E_{non} with respect to dihedral angles, using the current potential function. This cyclic process introduces many challenging inequalities (2), resulting in a modified potential function, readjusted decoy structures, etc. until finally no more violated inequalities can be found, and energy optimization does not move the native structure far. There is no guarantee that the native is near a global minimum of E, but it is difficult to find any lower energy conformation. Denote the resulting parameters by A^0_{ij} and B^0_{ij}.

The free energy of unfolding is calculated using

$$\Delta G_{\text{cal}} = -k_B T \left[\ln N + \ln \left(n^{-1} \sum_{j=1}^{n} \exp\left(-\frac{E_j}{k_B T} \right) \right) \right] - E_{nat} \tag{3}$$

where N is total degeneracy of the denatured state, E_j is the energy of the jth conformation in the denatured state, n is the number of conformations in the denatured state ensemble, T is the temperature, and k_B is Boltzmann's constant. Here we use the total number of self-avoiding walks of length $k = 110$ on a cubic lattice [6] estimated by $\ln N = 1.55k - 4.92136$. The energy parameters are further adjusted to give good calculated free energies of unfolding by the BFGS minimization algorithm [35] applied to

$$F = \sum_k (\Delta G_{\text{exp,k}} - \Delta G_{\text{cal,k}})^2 + w \sum_{i,j} \left[(A_{ij} - A^0_{ij})^2 + (B_{ij} + B^0_{ij})^2 \right] \tag{4}$$

where weight $w = 0.00001$ and ΔGs are in units of kJ/mole. While the first term in equation 4 tries to match the experimental and calculated values of ΔG, the second term stabilizes the adjustment process by keeping the changes in the parameters small. This objective function ensures that the energy parameters are only marginally modified and, therefore, still assign the lowest energy to the native state. A fixed set of 2000 structures is used to represent the denatured state in equation (3). Assuming that the denatured state of barnase is not a completely random coil state [15], we initially constructed the set of conformations by randomly resetting the (ϕ, Ψ) values of a random choice of 19 residues in the native conformation to values randomly selected from a database of native conformations of 313 proteins. During the ΔG matching process, if optimization with the current set of denatured states is unable to meet the matching criterion, i.e., $|\Delta G_{cal} - \Delta G_{exp}| < 1.0$ kJ/mole, then some of these 2000 conformations are replaced by new randomized conformations. After a few cycles no further substitutions were required. Even at this stage, the parameter adjustment continued on for a matter of CPU months on Sun Ultra 10 workstations, gradually building up a database of challenging nonnative conformations by local energy minimization to add to the constraints inequalities (2), while gradually minimizing the objective function (4). We denote the final potential function by PF8.

Not surprisingly, PF8 places all the eight native conformations at its apparent *global* minimum, and the estimated values of the ΔG are within ±0.7 kJ/mole of the experimental values. Application of PF8 to the other four mutants for which the experimental crystal structures are available, listed in Table III, is also quite successful. We find that with PF8, the experimental native conformations for these four mutants are at least at a local minimum of the potential function, i.e., within $\rho < 0.1$. The ΔG values for them are sufficiently close to the experimental values, within ±3.5 kJ/mole, and the effect of residue substitution on ΔG is correct, i.e., putting Ala in place of Ile reduces stability more than substituting Val in place of Ile (see Table III). While the crystal structures for various mutants are very close, the value of $\Delta\Delta G$ is significantly affected by the position of the mutated residue and those surrounding it. The ΔG values for four test set mutants with experimentally available crystal structures are well predicted by a denatured state of 2000 conformations obtained from the corresponding native. However, for other mutants without available experimental native conformations, the positions of mutated and surrounding atoms have greater error and therefore a somewhat larger denatured set was used. We applied PF8 to the remaining 54 mutants using the crystal structure of the wild-type as the reference native structure compared to the same denatured ensemble consisting of the 2000 structures used in training plus an extra 8000 randomized structures. The results for all 66 mutants are shown in Figure 1 and Table III. For the 58 test set mutants, the overall correlation coefficient is very poor ($R = 0.11$), due to a combination of good predictions and outliers. Only one mutant (V89T, not shown in Figure 1) is predicted to be thermodynamically unstable at 298 K. The twelve most conservative mutations (I→V, Y→F, T→S, K→R) have a standard deviation of 3.2 kJ/mol between experimental and calculated values. The worst results are obtained when the substitutions do not maintain the nature and size of the residues, e.g. V→T and Y→A, although some of the best predictions also involve nonconservative mutations. Possible causes for this puzzling behavior include both shortcomings of the potential, representation of the native state, and the model of the denatured state.

Even expanding the 2000 conformer denatured ensemble by another 8000 similarly randomized conformations, the denatured states of the training set and the 58 test mutants were dominated by very few structures. For all the 66 chains, only 10 different conformations were ranked one and only 36 different conformations appeared among the 10 best conformations for all chains out of the 10000 conformations. All the rank one conformations are fairly open and have pieces of α-helices and β-strands. In particular, the presence of the first α-helix near the N-terminus is very prominent. Some of these conformations are shown in Figure 2. Furthermore, for 62 of the 66 sequences, the statistical weights of the rank one conformations are greater than 0.5, thus justifying the choice of a small set of quasi-native conformations to represent the denatured state.

Applying PF8 to a set of 20 proteins of varying lengths, sequences, and crystal structures puts the native conformations of seven proteins to within 6 Å rmsd and of 16 proteins to within 8 Å rmsd of a local minimum (results not shown). The position of the local minimum near the native conformation is obtained by energy minimizing the conformation of the native with respect to torsion angles. While this

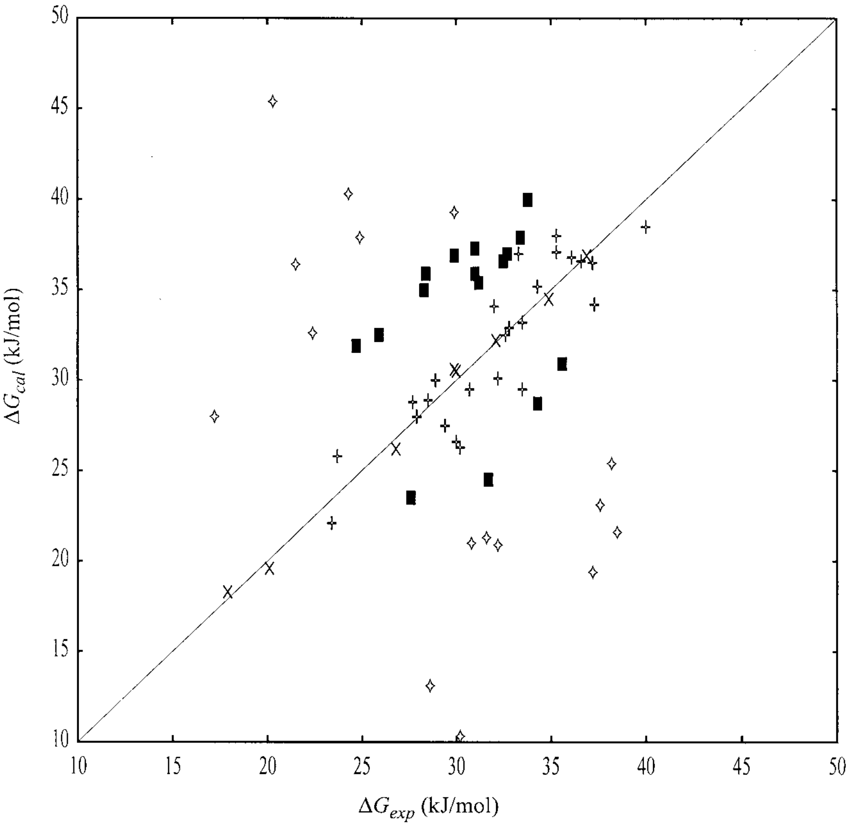

Figure 1. Correlation between experimental and calculated free energies. Correlation coefficient R = 0.11 for all 58 mutants, R = 0.59 without 17 worst outliers (open diamonds), and R = 0.89 with the 25 best performers (plus symbol). Crosses (X) represent the training set proteins.

is better performance than demonstrated for other potentials (see above), clearly the training set is quite limited in scope and cannot be expected to be as broadly applicable as our other potential of this type [2].

This work is an attempt to show that better potential functions which deal with both the fold recognition and structural stability problems can be developed, though the procedure needs to be improved. Apart from the computational time, some of the other issues are the types of interactions such as pairwise vs. many-body interactions, the nature of these interactions, and the determination of denatured state by a more direct method rather than trial and error.

4. Conclusions

Constructing empirical protein folding potentials is fraught with hazards. Success at discriminating between fixed native conformations of many different proteins

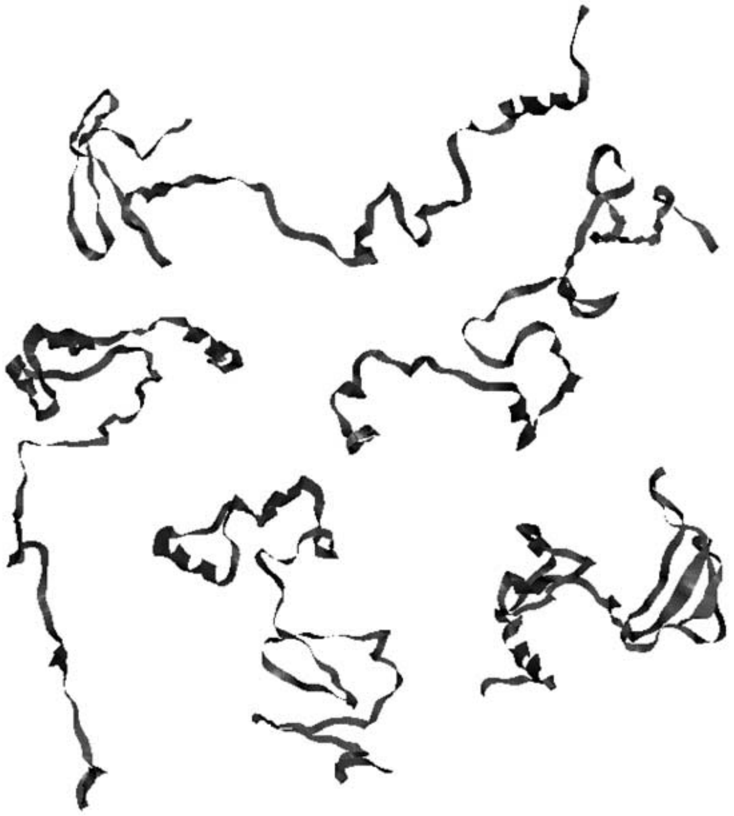

Figure 2. Some conformations in the denatured state having highest statistical weight for some barnase sequences.

and vast arrays of decoy structures does not imply there is even a local minimum anywhere near the native, much less a global one. However, even ensuring an apparent global minimum near the native does not imply that the potential can be used to calculate the free energy of unfolding. Due to the large number of conformational degrees of freedom and adjustable energy parameters, there is considerable variation remaining in the potential function even after satisfying orders of magnitude more constraints than there are adjustable parameters. In addition, free energy of unfolding calculations require some explicit or implicit model for the denatured state. Here we have developed a very computationally intensive procedure for creating a potential function and an ensemble of conformations representing the denatured state that both stabilizes the native conformation and gives good calculated free energies of unfolding for a very limited set of proteins. This suggests that the same approach may be extended to a broader range of proteins.

Acknowledgements

This work was supported by a grant from NIH (Grant #1-R01-GM59097-01). The authors would like to thank Dr Y.Z. Ohkubo for his help and allowing us to use some of his computer code. We would also like to thank the PDB contributors. Energy parameters are available from the authors upon request.

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E.: The Protein Data Bank, *Nucl. Acids Res.* **28** (2000), 235–242.
2. Chhajer, M. and Crippen, G.M.: A Protein Folding Potential that Places the Native States of a Large Number of Proteins near a Local Minimum, *BMC Struct. Biol.* **2** (2002), 4.
3. Clementi, C., Vendruscolo, M., Maritan, A. and Domany, E.: Folding Lennard-Jones Proteins by a Contact Potential, *Proteins* **37** (1999), 544–553.
4. CPLEX Version 6.6. ILOG, Inc. (2000), <http://www.ilog.com>.
5. Crippen, G.M. and Ohkubo, Y.Z.: Statistical Mechanics of Protein Folding by Exhaustive Enumeration, *Proteins* **32** (1998), 425–437.
6. Crippen, G.M.: Enumeration of Cubic Lattice Walks by Contact Class, *J. Chem. Phys.* **112** (2000), 11065–11068.
7. Crippen, G.M.: A Gaussian Statistical Mechanical Model for the Equilibrium Thermodynamics of Barnase Folding, *J. Mol. Biol.* **306** (2001), 565–573.
8. Crippen, G.M.: Constructing Smooth Potential Functions for Protein Folding, *J. Mol. Graph. Mod.* **19** (2001), 87–93.
9. Dill, K.A., Phillips, A.T. and Rosen, J.B.: Protein Structure and Energy Landscape Dependence on Sequence using a Continuous Energy Function, *J. Comput. Biol.* **4** (1997), 227–239.
10. Dombkowski, A.A. and Crippen, G.M.: Disulfide Recognition in an Optimized Threading Potential, *Protein Enging.* **13** (2000), 679–689.
11. Freund, S.M.V., Wong, K.-B. and Fersht, A.R.: Initiation Sites of Protein Folding by NMR Analysis, *Proc. Natl. Acad. Sci. USA* **93** (1996), 10600–10603.
12. Godzik. A., Kolnski, A. and Skolnick, J.: Are Proteins Ideal Mixtures of Amino Acids? Analysis of Energy Parameter Sets, *Protein Sci.* **4** (1995), 2107–2117.
13. Goldstein, R.A., Luthey-Shulten, Z. and Wolynes, P.G.: Protein Tertiary Structure Recognition using Optimized Hamiltonians with Local Interactions, *Proc. Natl. Acad. Sci. USA* **89** (1992), 9029–9033.
14. Guerois, R., Nielsen, J.E. and Serrano, L.: Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of more than 1000 Mutations, *J. Mol. Biol.* **320** (2002), 369–387. <http://fold-x.embl-heidelberg.de>
15. Hammerstrom, P. and Carlsson, U.: Is the Unfolded State the Rosetta Stone of the Protein Folding Problem?, *Biochem. Biophys. Res. Commun.* **276** (2000), 393–398.
16. Hinds, D.A. and Levitt, M.: Exploring Conformational Space with a Simple Lattice Model for Protein Structure, *J. Mol. Biol.* **243** (1994), 668–682.
17. Huang, E.S., Subbiah, S., Tsai, J. and Levitt, M.: Using a Hydrophobic Contact Potential to Evaluate Native and Near-Native Folds Generated by Molecular Dynamics Simulations, *J. Mol. Biol.* **257** (1996), 716–725.
18. Kabsch, W.: A Discussion of the Solution of the Best Rotation to relate Two Sets of Vectors, *Acta Cryst.* **A34** (1978), 827–828.
19. Kocher, J.-P.A., Rooman, M.J. and Wodak, S.J.: Factors Influencing the Ability of Knowledge-Based Potentials to Identify Native Sequence-Structure Matches, *J. Mol. Biol.* **235** (1994), 1598–1613.

20. Lazaridis, T. and Karplus, M.: Effective Energy Functions for Protein Structure Prediction, *Curr. Opin. Struct. Biol.* **10** (2000), 139–145.

21. Maiorov, V.N. and Crippen, G.M.: Contact Potential that Recognizes the Correct Folding of Globular Proteins, *J. Mol. Biol.* **227** (1992), 876–888.

22. Maiorov, V.N. and Crippen, G.M.: Size-Independent Comparison of Protein 3-Dimensional Structures, *Proteins* **22** (1995), 273–283.

23. Micheletti, C., Seno, F., Banavar, J.R. and Maritan, A.: Learning Effective Amino Acid Interactions through Interactive Stochastic Techniques, *Proteins* **42** (2001), 422–431.

24. Mirny, L.A. and Shakhnovich, E.I.: How to Derive a Protein Folding Potential? A New Approach to an Old Problem, *J. Mol. Biol.* **264** (1996), 1164–1179.

25. Miyazawa, S. and Jernigan, R.L.: Estimation of Effective Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation, *Macromolecules* **18** (1985), 534–552.

26. Miyazawa, S. and Jernigan, R.L.: Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading, *J. Mol. Biol.* **256** (1996), 623–644.

27. Ohkubo, Y.Z. and Crippen, G.M.: Potential Energy Function for Continuous State Model of Globular Proteins, *J. Comput. Biol.* **7** (2000), 363–379.

28. Park, B. and Levitt, M.: Energy Functions that Discriminate X-ray and Near Native Folds from Well-Constructed Decoys, *J. Mol. Biol.* **258** (1996), 367–392.

29. Park, B.H., Huang, E.S. and Levitt, M.: Factors Affecting the Ability of Energy Functions to Discriminate Correct from Incorrect Folds, *J. Mol. Biol.* **266** (1997), 831–846.

30. Park, K., Vendruscolo, M. and Domany, E.: Towards an Energy Function for the Contact Map Representation of Proteins, *Proteins* **40** (2000), 237–248.

31. Reith, D., Huber, T., Muller-Plathe, F. and Torda, A.E.: Free Energy Approximations in Simple Lattice Proteins, *J. Chem. Phys.* **114** (2001), 4998–5005.

32. Samudrala, R. and Moult, J.: An All-Atom Distance Dependent Conditional Probability Discriminatory Function for Protein Structure Prediction, *J. Mol. Biol.* **275** (1998), 895–916.

33. Serrano, L., Kellis, J.T., Cann, P., Matouschek, A. and Fersht A.R.: The Folding of an Enzyme II. Substructure of Barnase and the Contribution of Different Interactions to Protein Stability, *J. Mol. Biol.* **224** (1992), 783–804.

34. Shakhnovich, E.I.: Proteins with Selected Sequences Fold into Unique Native Conformation, *Phys. Rev. Lett.* **72** (1994), 3907–3910.

35. Shanno, D.F. and Phua, K.H.: Remark on Algorithm 500: Minimization of Unconstrained Multi-Variate Function [E4], *ACM Trans. Math. Software* **6** (1980), 618–622.

36. Sippl, M.J.: Boltzmann's Principle, Knowledge-Based Mean Fields and Protein Folding. An Approach to the Computational Determination of Protein Structures, *J. Comput-Aided Mol. Design* **7** (1993), 473–501.

37. Skolnick, J., Jaroszewski, L., Kolinski, A. and Godzik A.: Derivation and Testing of Pair Potentials for Protein Folding. When is the Quasichemical Approximation Correct?, *Protein Sci.* **6** (1997), 676–688.

38. Takei, J., Chu, R.-A. and Bai, Y.: Absence of Stable Intermediates on the Folding Pathway of Barnase, *Proc. Natl. Acad. Sci. USA* **97** (2000), 10796–10801.

39. Thomas, P.D. and Dill, K.A.: An Interactive Method for Extracting Energy-like Quantities from Protein Structure, *Proc. Natl. Acad. Sci. USA* **93** (1993), 11628–11633.

40. Thomas, P.D. and Dill, K.A.: Statistical Potentials Extracted from Protein Structures: How Accurate are They?, *J. Mol. Biol.* **257** (1996), 457–469.

41. Tobi, D. and Elber, R.: Distance-Dependent, Pair Potential for Protein Folding: Results from Linear Optimization, *Proteins* **41** (2000), 40–46.

42. Tobi, D., Shafran, G., Linial, N. and Elber, R.: On the Design and Analysis of Protein Folding Potentials, *Proteins* **40** (2000), 71–85.

43. Vendruscolo, M. and Domany, E.: Pairwise Contact Potentials are Unsuitable for Protein Folding, *J. Chem. Phys.* **109** (1998), 11101–11108.

44. Vendruscolo, M., Najmanovich, R. and Domany, E.: Can a Pairwise Contact Potential Stabilize Native Protein Folds against Decoys Obtained by Threading?, *Proteins* **38** (2000), 134–148.

45. Wang, Y., Zhang, H., Li, W. and Scott, R.A.: Discriminating Compact Nonnative Structures from the Native Structure of Globular Proteins, *Proc. Natl. Acad. Sci. USA* **92** (1995), 709–713.

46. Zagrovic, B., Snow, C.D., Shirts, M.R. and Pande, V.S.: Simulation of Folding of a Small Alpha-Helical Protein in Atomistic Detail using Worldwide-Distributed Computing, *J. Mol. Biol.* **323** (2002), 927–937.