ELSEVIER

Constructing smooth potential functions for protein folding

Gordon M. Crippen

College of Pharmacy, University of Michigan, Ann Arbor, MI, USA

A protein folding potential function ideally has several properties: it favors the native conformations for a number of protein sequences over a variety of nonnative folds; it can guide the search over conformations for the native state; it reflects changes in stability of the native fold due to changes in sequence; and it is relatively insensitive to small changes in conformation. While these are not mutually incompatible goals, attaining one property does not ensure that the others are satisfied. Examples are given of simple potentials having one property but lacking others. A new functional form of a folding potential is described where interactions between all nonhydrogen atoms are used to estimate interresidue interactions and implicit solvation. Its parameters can be adjusted to satisfy the above properties at least for barnase and a few other proteins. © 2001 by Elsevier Science Inc.

Keywords: protein folding; barnase; urea denaturation; free energy of protein folding

INTRODUCTION

Arbitrary Potential Functions

Consider the classical ab initio protein folding problem: given an amino acid sequence, calculate whether it has a well-defined native conformation at some reasonable temperature range and solvent composition; if so, predict what that conformation is. Numerous protein folding potentials or scoring functions have been devised to help solve this problem, or at least to treat a relaxation of the problem where the native conformation exists and must simply be selected from a set of decoys. We view this arbitrary scoring function, *E*, to be an energy-like function of conformation, *c*, and sequence, *s*, so that lower values of *E(c,s)* are more favorable, but there is not necessarily any physically significant zero value for the scale or physical unit.

Certainly it has been argued that basic physics should be our guide to constructing *E*, that it should include as many realistic features as possible, and that even its functional form should reflect the way we classify the different contributions to the overall Hamiltonian.[1] This is why standard all-atom potential functions have Lennard-Jones and Coulomb terms. The alternative viewpoint is that *E* ought to depend on the coordinates of the idealized point atoms and should be invariant under rigid translation and rotation of the whole system of molecules under consideration, but otherwise almost any functional form is worth considering. It has been shown that any such function can be approximated to any desired precision by a polynomial in the squared interatomic distances. [2] However, just as some series approximations converge more rapidly than others, this result says nothing about how many terms are required or how high the degrees of the terms must be. As an amusing example, consider an assortment of 36 minimal-energy conformations of the tetrapeptide N-acetyl-Ser-Val-Gly-Ser-N′-methylamide, according to the MMFF force field. MMFF is a detailed, realistic potential function with explicit hydrogen atoms. [3] It has the usual assortment of bond stretching, bond-angle bending, and nonbonded terms. Since the 36 conformations are local energy minima, they don't involve any very unfavorable interactions, but they do span a range of 20 kcal/mol. Forward stepwise linear regression[4] is able to find an empirical function involving a constant and 23 terms that are only linear in the squared distances, yet the standard deviation of the fit is a mere 0.15 kcal/mol.

$$E \approx \left\{ \begin{array}{l} \quad 5.344 + 0.020d^2_{7,2} + 0.110d^2_{10,1} - 1.191d^2_{11,8} \\ - 0.145d^2_{12,3} + 0.007d^2_{13,12} + 0.021d^2_{15,4} + 0.120d^2_{14,19} \\ + 0.121d^2_{16,9} - 0.046d_{21,2} - 0.417d^2_{21,17} + 0.049d^2_{21,18} \\ + 25.684d^2_{23,22} + 0.086d^2_{24,6} + 0.030d^2_{25,7} - 0.038d^2_{26,24} \\ + 0.019d^2_{27,20} + 0.692d^2_{28,25} + 0.046d^2_{29,13} + 0.072d^2_{32,1} \\ - 0.025d^2_{32,4} - 0.042d^2_{32,5} + 2.516d^2_{32,31} - 0.251d^2_{33,30} \end{array} \right. \tag{1}$$

Atoms are labelled as in Figure 1. Most atoms are involved in some statistically significant interaction, particularly hydrogens, which are more sensitive to small changes in dihedral angles. There are both local and long-range interactions, such that drawing a connecting line for each term would make a very messy figure. Clearly this is an outrageous exercise in arbitrary fitting because the real potential function, MMFF, is certainly not linear in some selection of squared distances.

The situation becomes even more outrageous if we assume

Corresponding author: G. M. Crippen, College of Pharmacy, University of Michigan, Ann Arbor, MI, USA

E-mail address: gcrippen@umich.edu (G. M. Crippen)

Journal of Molecular Graphics and Modelling 19, 87–93, 2001
© 2001 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00132-7

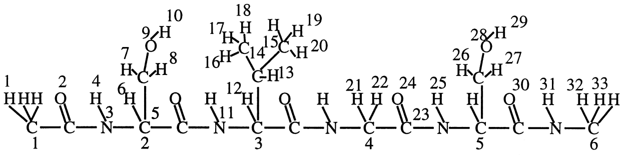

Figure 1. Critical interatomic distances used in the fit to the MMFF energies. Above are atom numbers used to designate all-atom pairs in Equation 1; below are the C^{α} atom numbers in Equation 2.

the potential function should use only C^{α} positions (six, including the capping methyl carbons) as a simulation of a potential function that uses a simplified representation of a protein. The best linear function uses a constant term and six terms that are linear in squared distances, but the standard deviation is 5.4 kcal/mol. The best quadratic function involves 14 terms, yielding a standard deviation of 4.3 kcal/mol. Only by resorting to a cubic function with 35 terms can the standard deviation of the fit be reduced to 0.18 kcal/mol.

$$E \approx \begin{cases} 0.001 + 0.198d_{3,1}^2d_{2,1}^2 - 0.039d_{3,1}^4 + 0.060d_{4,2}^2d_{4,1}^2 + 0.086d_{4,3}^2d_{3,2}^4 \\ - 0.070d_{4,3}^4d_{3,2}^2 - 0.0001d_{5,1}^2d_{3,1}^2 - 0.049d_{5,1}^2d_{4,2}^2 + 0.066d_{5,1}^2d_{4,3}^2 \\ - 0.003d_{5,1}^2d_{4,1}^2 - 0.104d_{5,2}^2d_{2,1}^2 + 0.153d_{5,3}^2d_{4,3}^2 - 0.030d_{5,3}^4 \\ + 0.093d_{5,3}^2d_{3,2}^2 + 0.215d_{5,4}^2d_{5,3}^2 - 0.034d_{5,4}^2d_{5,3}^2d_{3,2}^2 + 0.215d_{5,4}^2d_{5,3}^2 \\ + 0.001d_{5,4}^4d_{3,2}^2 + 0.028d_{6,1}^2d_{2,1}^2 - 0.006d_{6,1}^2d_{3,1}^2 - 0.034d_{6,2}^2d_{4,3}^2 \\ + 0.009d_{6,2}^2d_{5,1}^2 - 0.003d_{6,3}^2d_{6,2}^2 + 0.009d_{6,3}^2d_{3,1}^2 - 0.014d_{6,3}^2d_{4,1}^2 \\ + 0.003d_{6,3}^2d_{4,2}^2 - 0.006d_{6,4}^2d_{5,1}^2 - 0.003d_{6,4}^2d_{6,2}^2 + 0.012d_{6,4}^2d_{6,3}^2 \\ + 0.084d_{6,5}^2d_{3,2}^2 - 0.042d_{6,5}^2d_{3,2}^2d_{2,1}^2 - 0.059d_{6,5}^2d_{4,1}^2 + 0.110d_{6,5}^2d_{5,2}^2 \\ + 0.084d_{6,5}^2d_{3,2}^2 + 0.020d_{6,5}^4d_{2,1}^2 \end{cases} \tag{2}$$

Here the numbering of the C^{α} atoms is given in the lower part of Figure 1. All the squared distances are used, no term is linear in squared distances, but only a few are cubic. Even though MMFF is really a function of all atomic positions, the energies of this broad sampling of conformations can be accurately fit to a function of only C^{α}—C^{α} squared distances.

Pessimistically speaking, one can draw two conclusions about this exercise. Fitting some functions of conformational energy, such as the given energies themselves for several conformations, can be done without regard to the functional form of the true energy. Success at fitting says more about perseverance and the set of observations to be fitted than it does about the appropriateness of the functional form. Secondly, the level of detail in representing the protein has nothing to do with success in fitting. The most one can hope for is that all-atom potentials might have simpler terms than united-atom or united-residue potentials require.

Devious Potential Functions

Let *c(s)* be the native conformation for a given sequence, *s*. Figure 2 illustrates a few unsatisfactory potential functions that nonetheless have many of the desired features. The first impulse is to adjust the parameters of the potential function *E* such that it has a local minimum at the native conformation. Unfortunately, the local behavior of a function of many conformational variables and adjustable parameters has little relation to its global behavior. Such functions tend to have many local minima because they are usually formulated as sums of interactions between many particles, and the relative positions of these particles are complicated functions of the basic conformational variables, even though the individual terms may be unimodal functions of the distances between particles. Thus in curve (a) the native conformation is precisely a local minimum of *E*, but the global minimum corresponds to a very different structure. Some would view this as quite unsatisfactory, while others would happily note that the native still has a respectable Z-score because most of the potential function's surface is substantially higher than *E(c(s),s)*. In fact, the naive approach to achieving even this result is to adjust the parameters so that the gradient of *E* with respect to *c* at the native conformation is zero. Unless the Hessian of *E* is also constrained to be positive definite, the result is usually curve (b), where trivially the native conformation is at the global minimal value of *E*, but so is every other conformation. The potential energy surface is extremely smooth in the sense that small changes in conformation always result in zero change in *E*, but there is also zero correlation between *E* and conformational resemblance to the native. Consequently, *E* is of utterly no use in locating the native conformation. Curve (c), on the other hand, is useless due to insufficient sampling of nonnative conformations. Certainly

$$E(c,s) - E(c(s),\, s) > M \tag{3}$$

for some positive margin *M* for every nonnative structure *c* that has been examined. Furthermore, the potential is smooth in the sense that for every two conformations *A* and *B*, it satisfies a Lipschitz condition

$$\left|E(A,\, s) - E(B,\, s)\right| < L\|A - B\| \tag{4}$$

where $L > 0$ is the Lipschitz constant, and $\|A - B\| \geq 0$ is some measure of conformational similarity, such as the C^{α} atom RMSD or ρ.[5] In this schematic one-dimensional conformation space example, *L* is just the slope of the line. The problem with curve (c) is that although there is a good positive correlation between the value of *E* and the conformational dissimilarity with the native for all the sampled nonnative structures, none have been sampled to the left of the native. The global minimum may lie far to the left, and a more extensive search for nonnatives will find many violations of Equation 3.

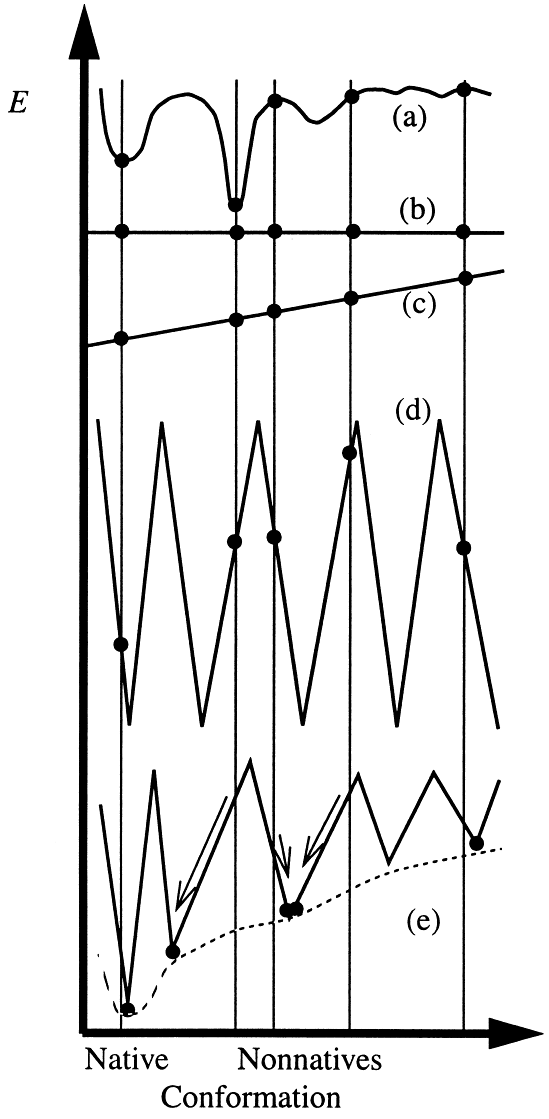

Figure 2. Unsatisfactory folding potentials, E, *for a given sequence as a function of conformation. (a) Local versus global minima; (b) trivial uniform global minimum; (c) native far from global minimum; (d) trivial rough energy surface; (e) rough surface but smooth and desirable progression of local minima.*

The most insidious case in Figure 2 is curve (d), where E has many local minima that may even be of comparable depth, but most of the potential surface has substantially higher values. E is not at all smooth, corresponding to a very large value of L in Equation 4. Adjusting the parameters may amount to nothing more than shifting the rough landscape so that the native conformation $c(s)$ happens to be located near a rare, low value of E, so that for many randomly sampled nonnative conformations c, the probability is high that inequality (3) will be satisfied. This effect is even more pronounced for high-dimensional conformation spaces. While it may be true that $E(c(s),s)$ is substantially lower (more favorable) than any other $E(c,s)$ even for many different sequences s, there is no noticeable correlation between E and similarity to the native, so that E will not help simulated annealing, molecular dynamics, or any sort of global search algorithm find the native structure. Local minimization from arbitrary starting conformations is very helpful for producing violations of inequality (3), leading to revised energy parameters where at least the lower bound envelope of the surface exhibits the desired correlation between locally minimized values of E and similarity to the near-native locally minimal E structure[6,7] as in curve (e).

Rough Potential Functions

Roughness of E with respect to conformation is a serious problem because a high (unfavorable) value has little correlation with the depth of a nearby local minimum that must be located by an expensively accurate optimization of conformation. One cause is the urge to incorporate a realistic steric repulsion term, often varying as $1/r_{ij}^{12}$, where r_{ij} is the distance between atoms i and j. Then every energetically favorable, tightly packed globular protein conformation is very close to a singularity where $E \rightarrow \infty$. Therefore in Equation 4, $L = \infty$. Even when the steric repulsion at short interatomic distances is differentiable and bounded,[8] E is customarily calculated as a sum of many terms, which may vary in magnitude with conformational change, and there may be both positive and negative terms: $E = \sum_i t_i$. Suppose E is a contact/no contact function, such as that of Miyazawa and Jernigan,[9] where a small change in conformation may change some term t_i from a nonzero value to zero, or vice versa. Then a useful estimate of roughness is

$$R = \frac{\max_i |t_i|}{|\sum_j t_j|} \tag{5}$$

so that R is large if E consists of many large magnitude terms that nearly cancel each other in the sum. Of course, the large R case is even more troublesome in finite precision computer arithmetic due to the propagation of roundoff error.

One way to make a discontinuous contact potential smoother is to carefully choose the interparticle distance cutoff to be at a local minimum of the radial distribution curve. That way, a small change in coordinates switches on or off a small number of terms compared with the number of contact terms that are on. For example, if the interacting particles are hard spheres as closely packed as in a liquid, a good choice of cutoff is the distance between the first and second packing shells. Unfortunately, the heavy atoms in a protein are so loosely packed that the radial distribution function is featureless except for a maximum corresponding to covalently bonded atom pairs and a second maximum for atom pairs linked by two covalent bonds (Figure 3). Consequently, any all-or-none distance cutoff choice affects relatively many terms when the conformation is even slightly perturbed. For example, a potential similar to that of Miyazawa and Jernigan[9] evaluated on the crystal structure of barnase (PDB entry 1A2P.A) can change by as much as 7% when all the residues are translated by random vectors having components in the range of 0 to 0.2 Å.

Roughness can be ameliorated by changing from all-or-none terms based on sharp cutoffs to sigmoidal switching function terms.[10,11] Then technically E is no longer discontinuous and may have continuous first, second, or even higher derivatives with respect to conformation. However, a condition number like R may still be large if there is a lot of cancellation of terms, and a large term can greatly change its magnitude due to some

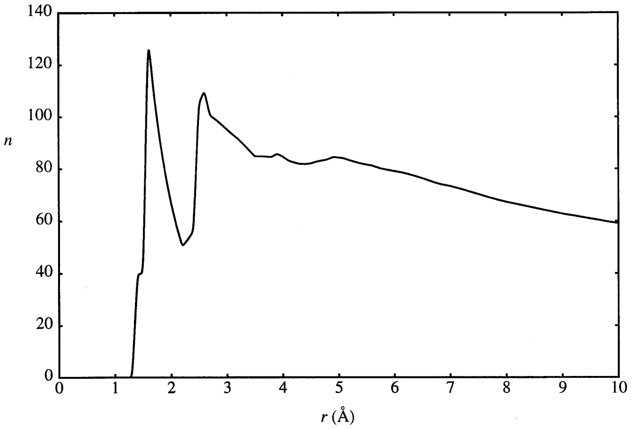

Figure 3. Cumulative radial density of heavy atoms in the crystal structure of barnase (1A2P.A). r = distance between atoms; n = (number of atom pairs closer than distance r)/r^3.

small change in conformation. An alternative approach featured in this study is to use more and smaller terms in the summation and to define them in a way that most small changes in conformation that change some terms from nonzero values to zero simultaneously change other terms in the opposite way.

A protein-folding potential may be used in other ways beyond simply favoring the native conformation over nonnative structures. Even in very artificial test situations where we are given the correct functional form of the "true" free energy, *E(c,s)*, the complete set of all possible conformations, *c*, all folding sequences, *s*, and their corresponding native conformations *c(s)*, *E* is not necessarily uniquely determined by Equation 3.[12] For example, consider barnase (bacillus amyloliquefaciens ribonuclease; EC: 3.1.27). The native conformations and the thermodynamics and kinetics of folding of wild-type barnase and many of its mutants have been studied by several groups. The objective in this study is to construct a potential function that not only favors the native conformation over some assortment of nonnatives for several different protein sequences, but to further adjust it so that it also can account for changes in the ΔG of folding of barnase upon mutation of the sequence. Furthermore, the potential should be smooth in the sense discussed above so that the results should not depend on fine details of the native or denatured conformations.

METHODS

Functional Form

The potential starts with the coordinates of all nonhydrogen atoms in a protein, as in a well-resolved crystal structure. The atoms are grouped into 22 different types of subsets: the backbone nitrogen (Nbb), the backbone carbonyl (Cbb), and the 20 different sidechains plus the C^α (Gly, Ala, etc.). Thus the four heavy atoms of a glycyl residue (labelled as N, CA, C, and O in a PDB file) are viewed as three subsets, namely Nbb = {N}, Cbb = {C, O}, and Gly = {CA}. For an alanyl residue, the sidechain subset Ala = {CA, CB}, and a large sidechain subset like Phe has 8 atoms. From a survey of 23 protein crystal structures (PDB entries 135L, 189L, 1AGX, 1AK2, 1APA, 1ARN, 1BTL, 1BW4, 1DRO, 1DSB.A, 1EAL, 1JUD, 1LBA,

Table 1. Maximum number of contacts between subsets and maximum total contacts for each subset

	Gly	Ala	Val	Leu	Ile	Cys	Met	Phe	Pro	Tyr	His	Trp	Ser	Thr	Lys	Arg	Asp	Asn	Glu	Gln	Nbb	Cbb
Gly	1	2	4	5	5	3	5	8	4	9	7	11	3	4	6	8	5	5	6	6	1	2
Ala	2	4	8	10	10	6	10	16	8	18	14	22	6	8	12	16	10	10	12	12	2	4
Val	4	8	16	20	20	12	20	30	16	36	27	44	12	16	24	30	20	20	24	24	4	8
Leu	5	10	20	25	25	15	24	38	20	44	33	52	15	20	29	40	25	25	29	29	5	10
Ile	5	10	20	25	24	15	24	37	20	42	35	48	15	19	30	37	23	24	30	30	5	10
Cys	3	6	12	15	15	9	15	20	12	27	21	32	9	12	16	24	15	15	18	17	3	6
Met	5	10	20	24	24	15	24	36	20	44	34	51	15	18	24	36	25	22	28	28	5	10
Phe	8	16	30	38	37	20	36	59	32	63	51	59	24	28	47	61	40	39	41	44	8	16
Pro	4	8	16	20	20	12	20	32	16	35	28	43	12	16	24	32	20	20	24	24	4	8
Tyr	9	18	36	44	42	27	44	63	35	73	60	72	27	35	51	67	36	44	47	52	9	18
His	7	14	27	33	35	21	34	51	28	60	48	72	21	28	37	53	35	35	39	36	7	14
Trp	11	22	44	52	48	32	51	59	43	72	72	10	27	44	64	77	43	48	50	64	11	22
Ser	3	6	12	15	15	9	15	24	12	27	21	27	9	12	18	24	15	15	18	18	3	6
Thr	4	8	16	20	19	12	18	28	16	35	28	44	12	16	24	30	20	20	24	24	4	8
Lys	6	12	24	29	30	16	24	47	24	51	37	64	18	24	33	37	30	30	36	36	6	12
Arg	8	16	30	40	37	24	36	61	32	67	53	77	24	30	37	48	40	38	43	47	8	16
Asp	5	10	20	25	23	15	25	40	20	36	35	43	15	20	30	40	25	25	29	30	5	10
Asn	5	10	20	25	24	15	22	39	20	44	35	48	15	20	30	38	25	25	30	30	5	10
Glu	6	12	24	29	30	18	28	41	24	47	39	50	18	24	36	43	29	30	35	35	6	12
Gln	6	12	24	29	30	17	28	44	24	52	36	64	18	24	36	47	30	30	35	35	6	12
Nbb	1	2	4	5	5	3	5	8	4	9	7	11	3	4	6	8	5	5	6	6	1	2
Cbb	2	4	8	10	10	6	10	16	8	18	14	22	6	8	12	16	10	10	12	12	2	4
total	75	144	244	314	312	200	315	482	251	505	451	569	208	280	333	448	319	309	313	373	75	145

Table 2. Pairwise subset interaction parameters (cal/mol × 10^{-2})

	Gly	Ala	Val	Leu	Ile	Cys	Met	Phe	Pro	Tyr	His	Trp	Ser	Thr	Lys	Arg	Asp	Asn	Glu	Gln	Nbb	Cbb
Gly	−23	15	−54	8	15	−23	69	−26	−1	−7	−21	52	64	−14	119	−6	−59	20	−103	186	27	−3
Ala		−57	274	24	18	21	−11	−32	−39	−66	−103	−10	−84	78	72	−42	−3	−223	2	163	7	0
Val			32	−57	−27	−25	−14	110	−30	70	1283	470	197	322	60	−185	−241	−2	−164	200	−19	−33
Leu				0	30	−34	−40	−16	193	−11	93	−1	−36	41	−5	11	−143	179	24	243	18	0
Ile					86	−36	−50	−6	22	21	51	56	−164	50	−49	−14	−91	−94	−24	−95	0	−4
Cys						−25	−14	−54	38	6	−8	−3	−72	−86	42	−6	2	−1	37	29	−45	−45
Met							54	−18	16	−31	23	232	0	−107	79	7	−30	34	159	32	−39	−28
Phe								−56	−504	12	−616	−1	258	50	19	23	2	−107	−14	−37	−8	19
Pro									11	−24	41	−6	29	−146	16	8	−53	152	26	3	35	28
Tyr										−83	−8	−23	−13	86	−4	−53	444	55	23	8	−4	−4
His											90	7	−97	−173	20	12	14	168	75	−1	14	9
Trp												11	−11	245	−25	−15	316	−66	−33	−51	6	2
Ser													−41	32	30	20	50	104	−24	19	−11	8
Thr														52	38	17	−31	−21	88	−136	−15	−26
Lys															61	−11	85	−6	−149	−4	60	25
Arg																129	−1	−124	−11	42	11	9
Asp																	54	−67	72	−42	−6	28
Asn																		−90	299	−20	−2	4
Glu																			−142	−811	−4	0
Gln																				−1337	21	−7
Nbb																					21	−27
Cbb																						3

1LCL, 1LKI, 1MML, 1NFA, 1PAZ, 1RSY, 1SFE, 1WBC, 2UCE, 8I1B), the maximum numbers of heavy atom contacts between subsets and total for a subset were noted (Table 1). As in the analysis of the relation between solvent exposure and atom–atom contacts by Colonna-Cesari and Sander,[13] a contact is defined as two heavy atoms in different subsets that are less than 6.4 Å apart, because then it is unlikely that a water molecule could lie between them. For a pair of small subsets like Ala and Cbb, the maximum observed number of contacts is just the product of the number of atoms in the two subsets, but due to the short contact cutoff distance, the maximum is smaller than that for large subsets. In order to assess overall solvent accessibility, the survey also noted the maximum total contacts between a given type of subset and the rest of the protein (last row in Table 1).

For a given sequence and all-atom conformation of a protein, the packing is summarized by a list of all pairwise subset interactions. Each interaction (i, j) notes the sequence numbers (ri ,r_j), the types of the two subsets (t_i ,t_j), the relative strength of the interaction (v_{ij}), and its relative contributions to the burial of the two subsets (b_i ,b_j). The relative strength $0 < v_{ij} \leq 1$ is just the number of atom–atom contacts for the interaction divided by the maximum observed number for those types. The relative burial, b_i, of subset i due to interaction *(i, j)* is the number of atom–atom contacts in the interaction divided by the maximum total number of contacts observed for a subset of type t_i. Thus for any one interaction, $0 < b_i \leq 1$, and summing over all interactions involving that subset, $\Sigma b_i \leq 1$. Although the interaction list is calculated from the true sequence and the coordinates of all the appropriate atoms, it is summarized in terms of v_{ij} and (b_i, b_j) to use it for energy calculations when threading. In threading, part of a long protein chain is used as a conformational template for a smaller, different amino acid sequence. Thus the interaction list is edited by deleting interactions involving some of the residues and by changing the types of sidechain subsets to those of the new sequence. It is assumed that the relative strengths of the remaining interactions are unchanged by these type changes and the relative burial contributions also remain the same, although deleting residues of the original structure will lower the total burial of some of the remaining subsets.

There are 22 × 23/2 = 253 (unordered) types of subset pairs. The energy is taken to be a sum over the interactions

$$E = \sum p_{t_i,t_j} v_{ij} \tag{6}$$

where each term depends only on the relative strength and the corresponding parameter, p. E is constructed to be linear in the 253 parameters so that a set of inequalities (Equation 3) derived from comparing the native conformation with nonnatives found by threading can be used to determine their values.

Parameter Adjustment

The term threading in this study refers to producing nonnative conformations of a given (single polypeptide chain) sequence of length n_s by taking the structure of another template protein having chain length $n_t \geq n_s$ and using all $n_t - n_s + 1$ contiguous (ungapped) segments of length n_s. Regardless of the original sequence of the template protein, the given sequence is applied to each segment, and E is calculated from the template protein's interaction list as described above. Comparing *E(c(s),s)* to threaded nonnative *E(c,s)* for a series of native proteins generates many inequalities (3) that must be satisfied. In addition to satisfying the inequalities, the objective is to minimize $\Sigma_{k=1}^{253} p^2{}_k$, to keep E as smooth as possible. Native proteins used in training are (PDB entry plus optional chain identifier): 135L, 189L, 1AGX, 1AK2, 1APA, 1ARN, 1BTL, 1BW4, 1DRO, 1DSB.A, 1EAL, 1JUD, 1LBA, 1LCL, 1LKI, 1MML, 1NFA, 1PAZ, 1RSY, 1SFE, 1WBC, 2UCE, 8I1B, and 1A2P.A. This last is wild-type barnase with the unresolved

Table 3. Protein Crystal Structures used for the Native State of Barnase

PDB code	Sequence	max RMSD (Å)
1A2P.A	wild type	0.73
1BAN.A	S91A	0.67
1BAO.A	Y78F	0.93
1BNS.A	T26A	0.69
1BRH.A	L14A	0.78
1BRJ.A	I88A	0.78
1BRI.A	I76A	0.83
1BRK.A	I96A	0.64
1BSA.A	I51V	0.70
1BSB.A	I76V	0.68
1BSC.A	I88V	0.73
1BSD.A	I96V	0.93
1BSE.A	L89V	0.61

Suffix .A indicates that the "A" chain of each PDB entry was used. Sequence xNy indicates that residue N of the wild-type sequence was mutated from single letter code type x to y. The maximum RMSD between the given structure and the other twelve is listed.

N-terminal residues added on in an arbitrary extended conformation having little contact with the rest of the protein. Threading templates for each native protein were all the other natives plus 153L, 193L, 1ADE.A, 1ALD, 1AMY, 1BMC, 1BPL.B, 1BTM.A, 1CEM, 1CHD, 1CIU, 1CSE.E, 1DIN, 1DOX, 1DXI.A, 1EUR, 1EZM, 1FRV.A, and one modeled partly unfolded barnase structure.

A second way to adjust the parameters is to fit experimental values of the free energy of unfolding of proteins. In particular, Table 3 in Serrano et al.[14] lists the $\Delta G_u^{H_2O}$ of unfolding in water (pH 6.3 buffer) at 298 K for wild-type barnase and 66 mutants. Most mutations are destabilizing, and one single residue mutation reduces the $\Delta G_u^{H_2O}$ from 8.82 to 4.29 kcal/mol, more than a factor of two. Although the kinetics of barnase unfolding is complicated, the equilibrium thermodynamics can be viewed as a simple two-state process.[15] Suppose the native and denatured states at the fixed temperature of T = 298 K can each be modeled as a single energy level, E, having a degeneracy, w. Then:

$$G = -RT\ln[w\exp(-E/RT)] \tag{7}$$

and

$$\Delta G = E_{denat} - E_{nat} + RT\ln(w_{nat}/w_{denat}). \tag{8}$$

As in the threading procedure, the barnase native conformation is taken to be the full 110-residue chain modeled from the crystal structure (1A2P.A). The denatured state is represented by the single fully extended conformation. These two conformations were used for each of the 67 sequences to calculate E_{nat} and E_{denat}, respectively. Degeneracies of the two states were estimated as explained below, and they enter in to Equation 8 only as their ratio. Then the parameters are adjusted to give a least squares fit between the calculated and observed ΔGs. Since this is overdetermined, the fit becomes perfect, and the remaining degrees of freedom are used to minimize the sum of the squares of the parameters, as in the threading procedure.

The number of walks on a cubic lattice having a given set of contacts can be fit to

$$\begin{aligned}\ln w = 0.3612 + 0.1005k - 0.04462c + 2.422k_0 \\ + 0.2394k_1 \\ + 0.1162k_2 - 3.035k_0/k + 0.4384k_3/k - 1.181f\end{aligned} \tag{9}$$

where k is the number of residues = 1 + the number of steps in each walk, c is the number of contacts (residues on adjacent lattice points but separated in sequence by 3 or more), k_0, k_1, k_2, k_3 are the numbers of residues in the range of zero, one, two, or three contacts respectively, and f is the number of residues on the ends of the chain that are unconstrained by contacts.[16] To estimate the degeneracy w from a protein conformation's interaction list, all sidechain-sidechain interactions are counted that have $|r_i - r_j| > 2$ and $v_{ij} > 3$. This choice of strength cutoff makes $w \approx 2$ for an α-helix, that of a typical domain is larger, the degeneracy of a β-hairpin is larger yet, and the degeneracy of a fully extended chain is vastly larger. Equation 9 estimates the number of cubic lattice walks having strictly a particular set of contacts and no others. Since the denatured state is supposed to represent not only those conformations having few contacts, such as the extended structure, but also all others that differ significantly from the native conformation, the denatured degeneracy is assumed to be the total number of self-avoiding cubic lattice walks, regardless of contacts[16]

$$\ln w = 1.55k - 4.92163. \tag{10}$$

RESULTS

We have had a lot of experience producing potential functions that recognize the correct native fold, so it is no surprise that with 253 adjustable parameters, the all-atom potential described above could be trained to be successful at discriminating between the native and threaded nonnative conformations for several protein sequences. The customary expectation is that with sufficient training over many proteins, the potential will embody generally valid information about protein folding that can be applied to other tasks. Much to our consternation, its agreement with the relative stability of barnase mutants was negligible. Then again, two other potentials performed as poorly, even though they are excellent at recognizing native conformations over a broad set of proteins.[9,10] Either the potential function fails to capture some important effects, or the training was insufficient to restrict the parameters to the neighborhood of general validity. Least squares fitting of the $\Delta G_u^{H_2O}$ data was of course numerically perfect, but then the potential failed at recognizing native conformations. Finally, the barnase mutant data was fitted subject to the constraints of recognizing native folds versus threaded nonnatives. Apparently the functional form is compatible with both tasks, but training for one has little effect on performance at the other.

The final parameters are listed in Table 2. The standard deviation between observed and calculated $\Delta G_u^{H_2O}$ is only 26 cal, no doubt well below the experimental error. There are little discrepancies in the values for interaction between commonly occurring and related sidechains, such as Gln—Leu=243 versus Gln—Ile=−95, leading one to suspect overfitting. Clearly this is strictly a special purpose potential function for barnase equilibrium thermodynamics plus some degree of recognition of correct folds. Apparently a much larger body of data needs to be included before these parameters would reflect a general range of phenomena.

The reason for resorting to such overfitting is that fewer parameters failed to give any sort of decent fit. For example, one can discriminate between native and threaded nonnative conformations for a wide variety of proteins on the basis of a cleverly chosen potential functional form and only 21 adjustable parameters.[10] In that potential, there is one parameter associated with the burial of each type of residue, and the burial is estimated from the C^β coordinates of the sidechains, with no adjustable parameters reflecting what type of surrounding residues are doing the burial. An analogous potential with 22 adjustable parameters failed to give a satisfactory fit to the variations in free energy of unfolding for barnase and 66 mutants. A good fit required pairwise interactions depending on both residue types, such as the result given above.

Because the interaction parameters in Table 2 vary considerably in magnitude and sign, it is possible that the potential is not very smooth in the sense of Equation 5. Using the crystal structure of wild-type barnase (1A2P.A), there are 3,939 interaction terms, and $R = 0.067$, indicating that there is a lot of cancellation of terms producing a sum that is only 15 times greater in magnitude than the absolute value of the largest term. To see how sensitive E is to small variations in conformation, consider the recent crystal structures of wild-type barnase and 12-point mutants listed in Table 3. Naturally, these are all very similar conformations since only 1 out of 110 residues is changed compared with the wild-type sequence. Yet they are also well-packed structures from high-resolution crystallography rather than artificially perturbed structures that might have steric overlaps or looser packing. Let $\sigma/|\mu|$ be a measure of variability of E, where μ is the mean and σ is the standard deviation of E over the 13 structures, using the wild-type sequence for all. For the group interaction potential with its large R and discontinuous atom–atom contact definition, the ratio works out to be 5.3%, which isn't much worse than the 2.9% achieved by our sigmoidally smoothed solvation potential.[10] As an example of how these differences arise, consider the wild-type structure and the L14A mutant listed in Table 3. In the wild-type structure, there happen to be 100 terms involving backbone or sidechain groups of Leu-14. In the L14A mutant, there are 60 matching interactions with Ala-14 having the same v_{ij}. The other 40 terms differ either because the surrounding groups bury a larger fraction of the smaller Ala-14, or because the smaller Ala-14 can't reach out as far as the larger Leu-14 can. Altogether the root mean square deviation in the 100 v_{ij}s is only 0.18, compared with the total range possible of $0 \leq v_{ij} \leq 1$. Out of the 70 interactions in both structures involving residue 15 but not 14, the rms deviation in v_{ij} is 0.09. This is due to slight shifts in conformation, rather than changes in sidechain size.

CONCLUSIONS

It is possible—but by no means trivial—to fit the experimentally determined $\Delta G_u^{H_2O}$ at 298 K for barnase and its 66 point mutants. Only a two-state model is required, but the potential function must be very detailed. Even though the potential has a sharp cutoff for atom–atom contacts, it is fairly smooth compared with residue–residue interaction potentials because there are more atom pairs, and small variations in coordinates don't often break up their interactions. The most valuable outcome from this study is not so much the special barnase potential function, but the realization that recognizing the native conformation over threaded nonnatives is one task for a potential function, and reproducing the free energy of folding is another task. These two tasks are not mutually incompatible, but success in one does not imply success in the other.

ACKNOWLEDGMENTS

This work was supported by grants from the Vahlteich Research Award Fund (College of Pharmacy, University of Michigan), NSF (DBI-9614074), and NIH (1-R01-GM59097-01).

REFERENCES

1 Lazaridis, T., and Karplus, M. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* 2000, **10**, 139–145

2 Dalbec, J. P. Straightening Euclidean invariants. *Ann. Math. Artif. Intell.* 1995, **13**, 97–108

3 Halgren, T. A. Merck molecular force field I. *J. Comput. Chem.* 1996, **17**, 490–519

4 Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. Applied Linear Regression Models, 3rd edn., Irwin, Chicago, 1996

5 Maiorov, V. N., and Crippen, G. M. Size-independent comparison of protein three-dimensional structures. Proteins: Struct. Funct. Genet. 1995, **22**, 273–283

6 Ohkubo, Y. Z., and Crippen, G. M. Determining contact energy function for continuous state models of globular protein conformations. In: Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (Shamir, R., Miyano, S., Istrail, S., Pevzner, P. & Waterman, M., eds.). ACM Press, New York, 2000, pp. 223–230

7 Ohkubo, Y. Z., and Crippen, G. M. Potential energy function for continuous state models of globular proteins. J. Comput. Biol. 2000, **7,** 363–379

8 Huber, T., Torda, A.E., and van Gunsteren, W.F. Structure optimization combining soft-core interaction functions, the diffusion equation method, and molecular dynamics. J. Phys. Chem. 1997, **A101**, 5926–5930

9 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.* 1996, **256**, 623–644

10 Dombkowski, A.A., and Crippen, G.M. Disulfide recognition in an optimized threading potential. Protein Engineering, 2000, **13,** 679–689

11 Huber, T., and Torda, A.E. Protein fold recognition without Boltzmann statistics or explicit physical basis. *Protein Sci.* 1998, **7**, 142–149

12 Crippen, G.M. Easily searched protein folding potentials. *J. Mol. Biol.* 1996, **260**, 467–475

13 Colonna-Cesari, F., and Sander, C. Excluded volume approximation to protein–solvent interaction. The solvent contact model. *Biophys. J.* 1990, **57**, 1103–1107

14 Serrano, L., Kellis, J.T. Jr., 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.* 1992, **224**, 783–804

15 Griko, Y.V., Makhatadze, G.I., Privalov, P.L., and Hartley, R.W. Thermodynamics of barnase unfolding. *Protein Sci.* 1994, **3**, 669–676

16 Crippen, G.M. Enumeration of cubic lattice walks by contact class. *J. Chem. Phys.* 2000, **112,** 11065–11068