A 1.8 Å Resolution Potential Function for Protein Folding

GORDON M. CRIPPEN[1] and MARK E. SNOW[2]

[1]College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109 and [2]Scientific Computation Group, University of Michigan Computing Center, Ann Arbor, Michigan 48103

SYNOPSIS

A general method is presented for constructing a potential function for approximate conformational calculations on globular proteins. The method involves solving a nonlinear program that seeks to adjust the potential's parameters in such a way that a minimum near the native remains a minimum and does not move far away, while any alternative minima shift so as to remain local minima but eventually rise higher than the level of the near-native minimum. Although the potential trades computational speed for detail by representing each amino acid residue as only a single point, correct secondary structural preferences and reasonable tertiary folding can be built into the potential in an entirely routine way. The potential has been parameterized to agree with the crystal structure of avian pancreatic polypeptide (having 36 residues) in the sense that the lowest minimum found (−407 arbitrary units) is reasonably close to the native (1.8 Å rms interresidue distance deviation). In contrast, the lowest nonnative conformation found after extensive searches by a variety of methods was −399 units and 7.5 Å away. Such potentials may prove to be useful in predicting approximate tertiary structure from amino acid sequence, if they can be generalized to apply to more than one protein.

INTRODUCTION

It has long been established that many proteins will fold reversibly in vitro, starting from a rather disordered state, and proceeding on a time scale of seconds or minutes to a unique conformational state indistinguishable from the native conformation.[1,2] The native state must then be a global minimum of the free energy over the region of conformation space accessible to the protein during the course of refolding. This region must be less than the whole of conformation space, because the known rate of rotation about single bonds rules out the exploration of all combinations of all the protein's many torsion angles in the experimentally observed time for refolding. In any case, there must be a correspondence between the amino acid sequence and the native conformation, relatively independent of external influences beyond reasonable solvent composition, pressure and temperature.

A long-standing goal among theoreticians is to compute the folding pathway or at least the native structure, given only the sequence. For small isolated molecules there are good quantum mechanical and empirical potential functions for calculating the internal energy as a function of conformation. Coupled with a thorough *global* search, these computer programs can successfully predict the conformations of a wide variety of molecules in the dilute vapor phase. With much more computational expense, classical molecular mechanics programs are becoming able to account for solvation effects as well. For molecules as large as proteins, quantum mechanical energy evaluations are out of the question, but there are good empirical potential functions available that treat each atom as a separate particle. It is possible to construct a potential function consisting only of attractive forces between atoms such that the potential has a unique global minimum, easily located from any arbitrary starting point by local minimization algorithms,[3,4] but the moment repulsive forces are added (van der Waals contacts, minimal bond length and angle constraints, etc.), the energy surface develops numerous local minima, very few of which have low enough energy to be physically rel-

CCC 0006-3525/90/10-111479-11 $04.00
Biopolymers, Vol. 29, 1479–1489 (1990)

evant. In fact, the number of local minima is thought to increase exponentially with the size of the molecule, such that a global search of the conformation space for even a tetrapeptide is infeasible.[5] After all, even the ultimate special-purpose, microminiaturized, highly parallel, analogue computer—the protein itself—takes shortcuts to solve the problem.

Computational chemists have taken two general approaches to solving the protein folding problems: algorithmic and energetic. Algorithmic development takes the current state-of-the-art energy function as fixed and tries to locate very low-energy structures by cleverly moving the atoms in such a way that the number of energy evaluations is small. The energetics approach seeks to modify the energy function such that the performance of the search algorithm is improved while still favoring the experimentally observed conformations. For example Levitt and Warshel, in their historically significant paper,[6] developed a simple algorithm for hopping from one local minimum to the next, using a potential function that was relatively smooth by simplifying the representation of the amino acid residues. The algorithmic innovation allowed them to proceed from an arbitrary initial conformation to a rather compact one, instead of just locally minimizing the energy and moving only a short distance thereby. Representing each amino acid residue by a single point made the energy surface smooth enough that the algorithm was not faced with an astronomical number of closely spaced local minima to explore. The flaw was that while the potential favored realistically compact structures, it did not discriminate very well between the native conformation and alternative compact conformations.

In general, a major shortcoming in the state of protein folding theory is that while we have amassed a great list of rules (see Refs.[7–11] for a tiny sampling) from the over 100 available high-resolution protein crystal structures regarding how native proteins *should* look, these rules tend not to reject the many alternative conformations that the native state *should not* resemble.[12,13] We have found the discrimination issue to be absolutely essential: not only must the potential favor the native, it must disfavor the alternatives, and the only way to test that is to examine many nonnative conformations.

In this paper, we are concerned primarily with energetic development. Broadly speaking, our goal is to devise a potential function for folding any protein, involving the smallest number of particles consistent with favoring the native conformation over alternative conformations. Regardless of the search algorithm, having to deal with fewer particles would generally be an advantage. There would be fewer degrees of freedom in the problem, and there would be fewer local minima. Clearly there are fewer ways to pack a cubic meter of bowling balls than a cubic meter of golf balls. Most work on the energetics side of the protein folding problem has concentrated on devising very accurate potentials that have a local minimum extremely close to the experimentally determined crystal structure for a number of compounds, as well as agreeing with observed vibrational and rotational spectra and enthalpies of sublimation. We, on the other hand, are willing to sacrifice local accuracy in order to facilitate global searches over conformation space. Note we do not presume that this potential function is necessarily "realistic" in terms of including correct dipole moments, standard van der Waals radii, and so on. Neither do we demand that the potential should be useful in simulating the folding process, but rather only the end result. In order to be applicable to many different proteins, the potential must take into account the amino acid sequence, so the coarsest subdivision of the polypeptide chain should be whole amino acid residues. In addition, the potential must have interactions depending on the amino acid types, as opposed to having a special interaction between residues 5 and 55, for example. The sequence separation between interacting residues can, however, be a factor, so as to enforce reasonable secondary structure, which might differ from long-range packing preferences. In order to keep the polypeptide chain from collapsing to a point, some sort of steric repulsion is required between virtually all pairs of residues. This unfortunately forces the potential to have many local minima, but this is countered to some degree by using only a single particle per residue. To maintain a self-avoiding chain, we must have a potential that at least contains a sum of interactions between most pairs of residues, but some have argued that to properly represent the effects of solvation, interactions between triples of residues must also be included.[14,15] As we shall see, there is no detectable evidence for such interactions in protein crystal structures, so we have stayed with two-body interactions solely.

In the functional form section, we will describe in detail the form of our proposed potential and how we were led to it. Then the parameter determination section explains our general method of adjusting the potential to agree with the native conformation of one small protein. Finally, the results section shows how the potential was tested.

FUNCTIONAL FORM

As discussed above, the simplest representation of a polypeptide chain that is likely to be generally usable for protein folding problems is to let each residue be a single "united atom." We have chosen to center that single particle on the C^α atom, so that the virtual bond length between residues remains nearly constant at 3.8 Å. Although *cis* peptide bonds are known to exist in protein crystal structures, they are rare, and we neglect them. Side chains are represented only to the extent that the residue particles have assigned amino acid types. Moreover, interactions between residues are taken to depend on the distance between them, but not on orientation. Clearly one should be concerned whether this gross simplification of the polypeptide has made an adequate potential function possible. The answer is that for the purposes of establishing approximate secondary structure, such as α-helices, and for tracing out the path of the backbone to a resolution of 1.8 Å, we will see in the results section that it is sufficient.

Three-Body Interactions

Preferences have long been observed in protein crystal structures for, say, hydrophobic residues to be near other hydrophobic residues more frequently than one would expect for random compact conformations of the same molecule.[12,16] Can these preferences be adequately modeled in terms of pairwise interactions, or must one also consider clusters of three residues?

Let $P(a|b)$ be the conditional probability of event a occurring, given that event b has. Denote the 20 different amino acid residue types by letters such as X, Y, and Z, while l indicates two residues have a "longrange" status with respect to each other (sequence separation greater than 7 residues), and c indicates they are in contact (C^α—C^α distance less than 10 Å). Then for any given protein p, we evaluate the conditional probability

$$P(XYlc|XYl;p) = \frac{n(XYlc;p)}{n(XYl;p)} \qquad (1)$$

where $n(XYlc;p)$ is the number of occurrences in protein p of pairs of residues of types X and Y that are longrange in sequence separation but in contact. Unfortunately, Eq. (1) depends on the geometry of protein p. For example, if the chosen contact cutoff distance d (= 10 Å), and $d \gg r_p$, the radius of p, then $P = 1$, whereas $P \to d^3/r_p^3$ for $d \ll r_p$. Now the question at hand is whether $P(XYZlc|XYZl;p)$ is significantly different from what one would expect on the basis of pairwise association conditional probabilities, but for any reasonable accuracy, this must be decided by a survey over many proteins, having a considerable range in sizes and shapes. An approach we used earlier[17] was to fit each protein to a geometric model where one can calculate Eq. (1) analytically, given the amino acid composition of the protein. Here we empirically determine the effect of protein size and shape for each protein, and then average together the adjusted results for all proteins.

It is a general property of conditional probabilities that the probability of three residues of types X, Y, and Z, all longrange from each other, being all in contact with each other is given by

$$\begin{aligned} &P(XYZlc|XYZl;p) \\ &\quad = P(XYlc|XYl;p)P(XZlc|XZl \,\&\, XYlc;p) \\ &\qquad \times P(YZlc|YZl \,\&\, XZlc \,\&\, XYlc;p) \end{aligned} \qquad (2)$$

but we need to express the two complex factors on the right side in terms of conditional probabilities we can evaluate via Eq. (1). Define protein-specific correction factors f_1 and f_2 by the equations

$$\begin{aligned} &P(XZlc|XZl \,\&\, XYlc;p) \\ &\qquad = f_1 P(XZlc|XZl;p) \end{aligned} \qquad (3)$$

and

$$\begin{aligned} &P(YZlc|YZl \,\&\, XZlc \,\&\, XYlc;p) \\ &\qquad = f_2 P(YZlc|YZl;p) \end{aligned} \qquad (4)$$

One would expect $f_1 < 1$ because Z and Y would tend to exclude each other sterically from the neighborhood of X. Similarly, there should be a more pronounced bias for $f_2 > 1$ because Y and Z are already restricted to be near X. Thus we can rewrite Eq. (2) in terms of f_1, f_2, and conditional probabilities we can readily measure:

$$\begin{aligned} &P(XYZlc|XYZl;p) \\ &\quad = P(XYlc|XYl;p)P(XZlc|XZl;p) \\ &\qquad \times P(YZlc|YZl;p) f_1 f_2 \end{aligned} \qquad (5)$$

Now if each native protein conformation is such that apparently only pairwise residue association pref-

erences are important, f_1 and f_2 should depend only on the conformation of p, and not on X, Y, and Z. Therefore for each p, we can estimate f_1 and f_2 by averaging over all residue types:

$$f_{p1} \approx \left\langle \frac{P(XZlc|XZl \,\&\, XYlc; p)}{P(XZlc|XZl; p)} \right\rangle_{XYZ} \quad (6)$$

$$f_{p2} \approx \left\langle \frac{P(YZlc|YZl \,\&\, XZlc \,\&\, XYlc; p)}{P(YZlc|YZl; p)} \right\rangle_{XYZ} \quad (7)$$

The procedure is now straightforward: for each protein p, estimate f_{p1} and f_{p2}, and then for all triples of residue types estimate the right-hand side of Eq. (5), using Eq. (1). Alternatively, one can directly estimate the left-hand side of Eq. (5) by surveying the crystal structures. In any case, we are making the best estimates[18] of the means μ of various observations x_i by the corresponding sample mean over a limited set of n proteins, and of the standard deviations σ by

$$\sigma = \left[\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n(n-1)} \right]^{1/2} \quad (8)$$

over the sampling of proteins. In a survey over 22 proteins from the Brookhaven Protein Data Bank[19] (identification codes 1abp, 1apr, 1ctx, 1lyz, 1pcy, 1rhd, 1sn3, 2adk, 2cna, 2fd1, 2sns, 3cpv, 3mbn, 3pgk, 3rxn, 3tln, 4adh, 4cyt, 4fxn, 4pti, 5cpa, and 8pap), there was never a case where the left and right sides of Eq. (5) differed by more than 2 standard deviations. We therefore conclude there is no compelling evidence at present to include anything more than residue pair interactions in our rough potential function.

Form of Terms

Having now settled on a potential function that consists of a sum of pairwise interaction terms, we must decide on the functional form for each term. The guidelines are simplicity, ease of calculation, and at least minimal simulation of physical reality. If two residues are distant enough in sequence, they can come close together in space, but never closer than 4 Å for obvious steric reasons. On the other hand, when the α-carbons are more than 10 Å apart, there is generally little interaction, and association preferences are not apparent. A simple and easy to calculate interaction term that fulfills these criteria is

$$e(d_{ij}; A, B) = \frac{A}{d_{ij}^m} - \frac{B}{d_{ij}^n} \quad (9)$$

where d_{ij} is the distance between the two interacting residues, $m > n > 0$ are (generally even) integers, and $A, B \geq 0$ are adjustable parameters discussed in the parameter determination section. The choice of A and B determines the optimal separation ρ and the well depth at that separation ϵ. Clearly, $e \to -0$ as $d_{ij} \to \infty$ and $e \to \infty$ as $d_{ij} \to 0$, but the rate of these trends is governed by m and n. When $m = 12$ and $n = 6$, even for the best case of $\rho = 4$ Å, the interaction at 10 Å is 0.8% as strong as at $d_{ij} = \rho$. We have chosen $m = 12$ and $n = 10$, so that the strength at 10 Å is only 0.05%, and only close contacts have significant contributions to the total potential value.

Instead of using A and B, one can easily rewrite Eq. (9) for our choice of m and n in terms of ϵ and ρ directly:

$$e(d_{ij}; \epsilon, \rho) = \epsilon \left[5\left(\frac{\rho}{d_{ij}}\right)^{12} - 6\left(\frac{\rho}{d_{ij}}\right)^{10} \right] \quad (10)$$

As we will detail in the next topic, there will be particular values of ϵ and ρ associated with certain residue types and sequence separations. It is easy to verify that for given protein crystal structures (i.e., fixed d_{ij}s), the choice of one of these ρs that minimizes the sum of all terms where it is used is given by

$$\rho = \left[\frac{\sum d_{ij}^{-10}}{\sum d_{ij}^{-12}} \right]^{1/2} \quad (11)$$

Suppose for a particular class of interaction, we let $\epsilon = 1$ and then adjusted the corresponding ρ so that the sum of the potentials of several crystal structures is a minimum with respect to ρ (not necessarily with respect to varying the residue coordinates!). If there is a wide spread in the d_{ij}s, then ρ will tend to be close to the smallest values of d_{ij}, and many interactions will contribute only small negative values to the total potential. If we could choose the set of interactions such that there is little spread in the d_{ij}s, then the corresponding optimal ρ would tend to leave most interactions with little to be gained by moving the residues of the crystal structures.

There is one other technical detail concerning the functional form of the interaction terms. If we model the chain connectivity, that is, the interaction be-

tween sequentially adjacent residues, with Eq. (10), then if the chain is somehow broken during a calculation with the ends widely separated, there is little force toward rejoining it. Therefore, for these interactions alone we use a simple harmonic function:

$$e_b(d_{i,i+1};\, \epsilon,\, \rho) = \epsilon(d_{i,i+1}^2 - \rho^2)^2 \qquad (12)$$

Residue Classification

Having settled on a functional form for each pairwise interaction term, the problem now is to group interactions according to the residue types involved and their sequence separation so that there will be a different ϵ, ρ pair for each group, the values of which will be determined in the next section. It is desirable to keep the number of adjustable parameters relatively low, both to avoid "overfitting" the data, and also to keep the fitting process from seizing on some statistically unusual feature of the training data set, resulting in a parameter set that misrepresents the folding preferences of proteins in general. Our heuristic is to initially set all ϵs to 1, and then change the interaction groupings so that the potential value resulting from Eq. (11) is minimal.

As in our earlier work,[20] we make the assumption that all interactions between residues i and $i + 1$ for all i fall in the same unique group, regardless of the types of the two residues, and indeed the functional form is the special Eq. (12). Interactions between residues i and $i + 2$ are assumed to depend on the type of residue $i + 1$, resulting in 20 ϵ, ρ pairs to be determined. Interactions $i, i + 3$ depend on the types of residues $i + 1$ and $i + 2$; all longer range interactions are grouped together as "longrange," and depend on the types of the two residues themselves. There would be $20 \times 21/2 = 210$ residue pair types to consider for $i, i + 3$ interactions, and another 210 for longrange, resulting in 420 more ϵ, ρ pairs, an unacceptably high number. If, however, we lump residue types into classes, we could greatly reduce the number of adjustable potential parameters.

Clearly, it would be desirable to have a systematic method for grouping residue types into classes in such a way as to facilitate constructing a good potential function. Our approach, called the minimum potential classification algorithm, is conceptually simple. Consider the crystal structures of 17 proteins (same as the 22 listed above plus 1ppt, except 1abp, 1apr, 2cna, 2fd1, 3pgk, and 8pap each have a few C^α—C^α distances less than 3.0 Å that badly affect our short distance biased averages). Suppose we placed all residues into a single type class and hence only a single pair type class, so that all i, $i + 3$ interactions of these proteins at unit well depth and optimal ρ contribute −827.97 units to the sum of the proteins' potentials. There are $2^{20} = 1{,}048{,}576$ ways to group the 20 residue types into two classes and hence have three pair type classes (1-1, 1-2, and 2-2). The best of these numerous but manageable combinations produces a summed potential of −850.03, a substantial improvement brought about by now having three ρs instead of one, each being more closely representative of the sorts of distances observed in the crystal structures for that pair type class. The algorithm continues on to subdivide each of the two residue classes into two subclasses, and so on. We arbitrarily stopped the process with four residue classes (and hence 10 pair type classes) for $i, i + 3$ interactions and eight for longrange (hence 36 pair type classes). Table I summarizes the optimal residue classes, designating the classes by 0, 1, 2, . . . , and gives the potential contributions.

One should view these residue classes as simply an entirely empirical trick to reduce the number of adjustable parameters in the potential in such a way that these parameters, particularly the ρs, can be estimated in advance, and that proteins would tend to be under relatively low stress according to this potential. For example, we find in the final 8-class subdivision of residues in longrange interactions, that the smallest optimal ρ is 3.99 Å, namely for class 0 to class 0 interactions. The last row of Table I shows that class 0 is {G, P, R}, a rather unlikely combination, but it turns out there are many close GG and GP longrange contacts, and relatively few GR, PP, PR, and RR interactions at all.

To summarize, we have chosen our potential function to be a sum of pairwise interactions, each term given by Eqs. (10) and (12), and for i, $i + 3$ and longrange interactions, the residue types t are mapped into the final classes c given in Table I, and these are in turn mapped into the 10 and 36 pair classes, respectively. The mapping of individual residue class numbers c_i and c_j into pair class p_{ij} is simply

$$p_{ij} = \frac{(c_i + 1)(c_i + 2)}{2} - 1 - c_i + c_j \qquad \text{for} \quad c_i \geq c_j \qquad (13)$$

For each of these $1 + 20 + 10 + 36 = 67$ classes of interactions, there correspond an ϵ and a ρ, the values of which will be determined in the next section.

Table I Grouping Residues into Classes for Minimal Scatter of Distances Within Residue Pair Type Classes (See Text for Further Explanation)

Interaction	Potential	G	A	V	L	I	C	M	F	P	Y	H	W	S	T	K	R	D	N	E	Q
$i, i+3$	−827.97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	−850.03	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
	−866.80	0	2	2	2	0	2	1	0	0	2	0	0	0	0	0	1	2	0	2	2
Final	−874.01	0	2	2	2	3	2	1	3	0	2	3	0	0	3	3	1	2	3	2	2
Longrange	−1078.32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	−1379.84	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1
	−1459.81	0	1	1	2	1	1	2	2	0	1	2	1	1	1	1	0	2	1	2	1
	−1525.05	0	1	3	2	3	3	2	2	0	1	2	3	3	3	1	0	2	1	2	3
	−1550.89	0	1	4	2	3	3	2	2	0	1	2	4	4	3	1	0	2	1	2	4
	−1579.80	0	1	4	2	3	3	5	2	0	1	5	4	4	3	1	0	5	1	5	4
	−1641.65	0	1	4	2	3	3	5	2	0	6	5	4	4	3	6	0	5	1	5	4
Final	−1659.39	0	1	4	2	3	3	5	2	0	6	5	4	4	3	6	0	5	1	5	7

Thus the total potential value of a protein, given its C^{α} coordinates and amino acid sequence, is

$$E = \sum_{i=1}^{n-1} e_b(d_{i,i+1};\, \epsilon_1,\, \rho_1) + \sum_{i=1}^{n-2} e(d_{i,i+2};\, \epsilon_2,\, \rho_2(t_{i+1})) + \sum_{i=1}^{n-3} e(d_{i,i+3};\, \epsilon_3,\, \rho_3(t_{i+1},\, t_{i+2})) + \sum_{i=1}^{n-4} \sum_{j=i+4}^{n} e(d_{ij};\, \epsilon_l,\, \rho_l(t_i,\, t_j)) \quad (14)$$

where the subscripts on the parameters indicate what class of interaction they are involved in, and the dependency on residue type is indicated by ts in parentheses.

PARAMETER DETERMINATION

Exact Local Agreement with Native

The task now is to somehow adjust the 134 ϵs and ρs so that the potential favors in some sense the native conformation of at least one protein, which we have chosen to be avian pancreatic polypeptide[21] (Brookhaven file "1ppt"). Since 1ppt has only 36 residues, it would seem trivial to fit a structure having $3 \times 36 - 6 = 102$ degrees of conformational freedom by adjusting 134 parameters. First, such an argument is valid only locally in a nonlinear fitting problem like this; second, there are some parameters that are involved in many interactions; and third, the interresidue distances in these interactions are linked together in a complicated way by the constraints of three-dimensional geometry. The easiest way to see how this leads to trouble is by way of a simple example. Let $\mathcal{R}^n$ denote n-dimensional space, so that real molecules are in $\mathcal{R}^3$ and the real line is $\mathcal{R}^1$. Suppose we have a molecule in $\mathcal{R}^1$ consisting of three atoms placed in sequence at 0, $x > 0$, and $y > x$. Then our total potential is just $E(x, y) = e(x) + e(y) + e(y - x)$, taking all three interactions to be of the same type. If the form of e is $e(r) = k(r - \rho)^2$, then in order for the native structure x_{nat}, y_{nat} to be a potential minimum, we must have $x_{\text{nat}} = 2\rho/3$ and $y_{\text{nat}} = 4\rho/3$, independent of $k > 0$. Clearly this cannot always be done for an arbitrary native conformation.

To put it more generally, suppose we group the terms in Eq. (14) according to the class of interaction $k = 1, \ldots, 67$ and call the sum of all terms for class k η_k. Thus Eq. (14) becomes

$$E(\mathbf{x}) = \sum_{k=1}^{m} \eta_k \quad (15)$$

where $\mathbf{x}$ is the vector of all coordinates of all residues of the protein, and we have numbered the classes occurring in the protein 1 through m. Consider D, the analog of the rigidity matrix,[4] defined as a Jacobian by

$$(D)_{ij} = \left.\frac{\partial \eta_i}{\partial x_j}\right|_{\mathbf{x}=\mathbf{x}_{\text{nat}}} \quad (16)$$

If the native structure is to be a minimum in the potential, the net force on each residue must be zero, or in other words, all row sums of D must be zero.

In fact, the null space of D is the set of all infinitesimal motions that leave the potential unchanged. These include rigid translations and rotations, but they may include conformational changes as well. In our case with slightly less than 67 rows (some classes of interactions are not present in some small proteins) and 108 columns for 1ppt, we are unable to restrict the null space to only rigid translations and rotations. Our experience over the years with several different native proteins and various classifications of interactions having various functional forms has been that we are never able to force the gradient of the potential to zero at the native conformation without making the potential flat everywhere. Minimizing the magnitude of the gradient of the native 1ppt by adjusting the parameters starting from unit well depths and optimal ρs from the protein survey, results in a set of parameters such that minimization with respect to coordinates starting from the native moves 3.17 Å away. In other words, reducing the gradient at the native does not tend to even produce a perturbed structure near the native.

Approximate Global Agreement with Native

If we cannot adjust the parameters to make exactly the native structure a potential minimum, then there must be a minimum at least near the native that can be reached by local unconstrained optimization starting at the native. Denote this conformation by "perturbed" or "pert." Since steric repulsions produce many local minima that tend to be fairly closely spaced in conformation space, the existence of a perturbed structure is actually rather trivially true. What is much less trivial is that any alternative conformation (denoted by "alt") should have a worse (higher) potential value than the perturbed one. In order to establish a general slope and some broad predictive power to the bumpy potential surface, we further demand that the further away the alternative structure is, the worse its potential should be.

$$E(\text{alt}) - E(\text{pert}) \geq \delta(\text{alt, pert}) \qquad (17)$$

where the measure of distance in conformation space we use is the rms interresidue distance deviation

$$\delta(\text{alt, pert}) = \left[\frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{\text{alt},i,j} - d_{\text{pert},i,j})^2}{n(n-1)/2} \right]^{1/2} \qquad (18)$$

In our earlier work[20] we used many constraints of the form of Eq. (17) as linear inequalities with respect to adjustable A, B parameters in Eq. (9). Although we never produced an infeasible set of inequalities in this way, the system tended to become ill-conditioned as successive alternative conformations added their constraints. In our hands, dealing with ϵ and ρ parameters has proven to be much better behaved. Although E is linear in the ϵs but nonlinear in the ρs, we have found it essential to adjust both,

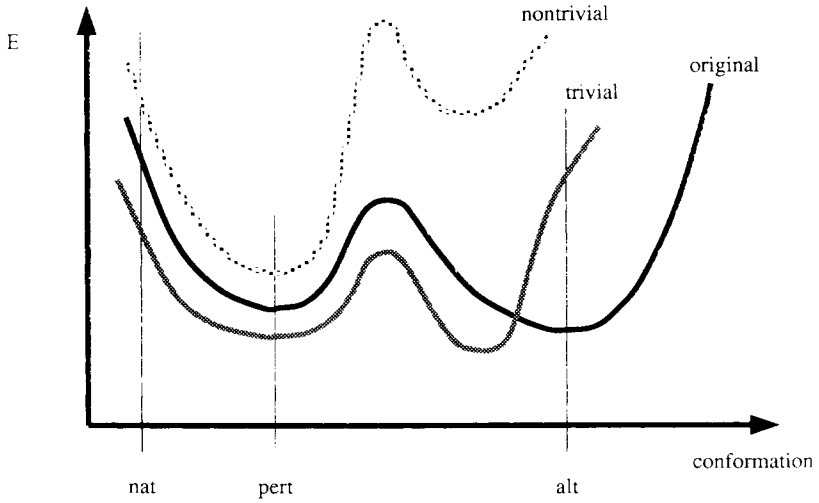

Figure 1 The original potential surface has an alternative conformer that is lower than the perturbed one. A trivial type of parameter change merely shifts to a new violating alternative conformer, whereas a nontrivial parameter change eliminates the offending conformer.

even though that means turning a linear programming problem into a nonlinear one.

Suppose we have some reasonable starting values for the parameters, such as all $\epsilon = 1$, and ρs given by Eq. (11) from a protein survey. Then local minimization of E starting at the crystal structure of 1ppt produces a nearby perturbed structure (δ(nat, pert) = 2.38 Å), and minimization from many random starting conformations may eventually turn up an alternative conformation that violates Eq. (17). It is usually easy to adjust the parameters so that the inequality, viewed as a function of just the parameters for fixed conformations, is no longer violated. However, the result is generally as shown in Figure 1, where the new potential surface has a new alternative conformation that still constitutes a violation. One can iterate this procedure, building up a data base of hundreds of inequalities, a few of which are active in determining the next set of parameters. We have found it much more effective to consider each inequality as a function of the parameters, the coordinates of the alternative conformation, and the coordinates of the perturbed conformation. Then the trivial solution of Figure 1 is avoided, and one can proceed with only a single alternative conformer.

After trying many variations on this theme, we finally hit upon the following nonlinear program as the best formulation of the parameter determination task:

$$\text{minimize } \delta(\text{nat, pert})$$

$$\text{subject to } \begin{cases} E(\text{alt}) - E(\text{pert}) \geq \delta(\text{alt, pert}) \\ \nabla E(\text{pert}) = \mathbf{0} \\ \nabla E(\text{alt}) = \mathbf{0} \\ 1 \leq \epsilon_i \leq 10 \quad \forall i \\ \rho_{\min,i} \leq \rho_i \leq 10 \quad \forall i \end{cases} \tag{19}$$

where the coordinates of the alternative and perturbed structures are variable (but not the coordinates of the native), as are all ϵs and ρs. The $\rho_{\min,i}$ are derived from Eq. (11) and range from 4 to 6.5 Å. The restrictions on the ϵs and ρs serve to avoid absurd and unbounded solutions, but they are not overly constraining since we indeed have found a suitable parameter set. The nonlinear program can be expressed in words as follows: "Find a reasonable set of parameters so that even allowing the alternative and perturbed conformations to shift so as to remain minima, the alternative's energy remains worse than that of the perturbed by a margin at least as great as their difference in conformation space; in addition, try to keep the perturbed conformation close to the native." We have solved Eq. (19) by augmented Lagrangians[22,23] for 1ppt at a cost of about 110 h CPU time on a Sun 4/280. Since

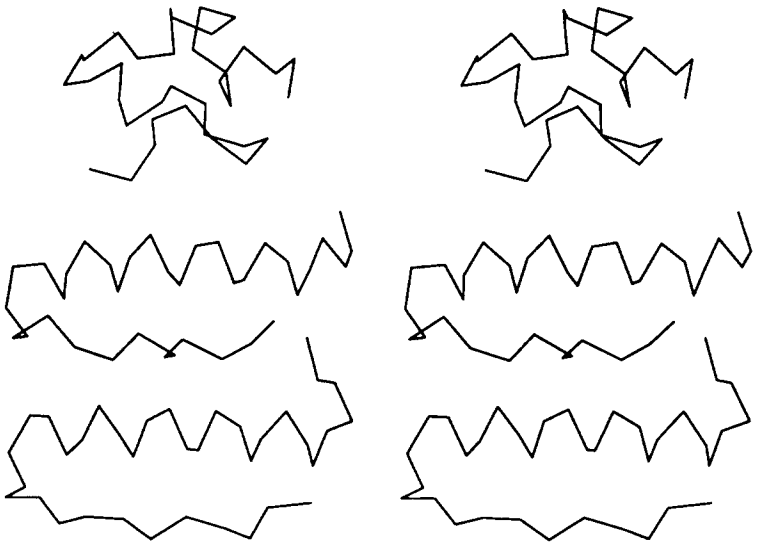

Figure 2 Stereo pairs illustrating the C$^\alpha$ traces for three conformations of 1ppt. At the top is a very compact nonnative structure typically favored by energy embedding (potential = −399.6, rms deviation to the native = 7.52 Å), the middle is the best conformation ever found (−406.9, rms 1.84 Å), and the bottom is the crystal structure of 1ppt. Structures are shown in roughly similar orientations with the N-terminus below, the C-terminus above, and the α-helix running from left to right.

this is a *non*linear program, the result depends on the starting point, and although failure to converge is not proof that the constraints are mutually inconsistent, successfully finding a solution is proof that at least one does exist. One can begin by setting all $\epsilon = 1$ and all ρs from Eq. (11), finding the perturbed conformation by local minimization from the native, and locating the alternative conformation by minimizing from random structures until one is found that violates Eq. (17). Solving Eq. (19) produces a new set of parameters and slightly revised (δ around 0.2–0.3 Å) perturbed and alternative conformers. Another random search might turn up a substantially different troublesome alternative, requiring resolving the nonlinear program with two alternative structures, etc. In practice, we find that simply substituting the new alternative for the old produces a second set of parameters that satisfy both. In fact, we generated a long series of parameter sets by solving other sorts of nonlinear programs, each time using the results of the previous try as the starting parameters for the next calculation. When we finally hit upon Eq. (19), consideration of only two successive alternative conformations produced the final parameter set, starting from the best parameters we had found up until then.

RESULTS

In the course of developing Eq. (19), we produced many alternative conformations of 1ppt, most of them clustered near the native, but many having quite different conformations (δ = 6–7 Å). The final parameter set resulted from solving the nonlinear program, using as reference two successive alternative conformers belonging to this class of quite different structures. Native 1ppt has a long straight helix, a tight bend, and then an antiparallel extended strand packed alongside the helix. In general, the very different alternative structures that the potential function tends to favor have the helix broken in the middle, the two halves packed together as an antiparallel helix-turn-helix cluster, and then the remainder wraps around the core in a rather extended fashion (see Figure 2). The final parameters are given in Table II, where the interactions are labeled as in Eq. (14).

Validating these parameters is necessarily somewhat inductive, since one cannot thoroughly search the entire conformation space of 1ppt. Locally, we can safely say that the native has a potential value of −343.6, and it is not a local minimum. The perturbed (near-native minimum) has depth −388.8 at

Table II Final Parameter Set with Interaction Labels Corresponding to Eq. (14)

Interaction	ϵ	ρ	Interaction	ϵ	ρ	Interaction	ϵ	ρ
1	1.01	3.82	3,1	4.83	8.27	1,13	1.01	9.85
			3,2	1.00	5.30	1,14	1.55	5.87
			3,3	1.00	5.27	1,15	1.12	5.22
2,1	2.51	7.44	3,4	1.00	5.05	1,16	1.01	5.66
2,2	2.22	6.62	3,5	1.00	4.61	1,17	1.10	5.39
2,3	2.25	5.61	3,6	10.00	5.17	1,18	1.18	5.59
2,4	3.32	5.57	3,7	4.69	8.92	1,19	1.09	5.94
2,5	2.11	5.81	3,8	5.53	4.99	1,20	1.00	5.61
2,6	1.00	5.57	3,9	8.66	4.99	1,21	1.00	4.99
2,7	1.00	5.36	3,10	1.00	5.13	1,22	3.03	5.74
2,8	1.00	5.47				1,23	1.02	6.75
2,9	1.88	5.72				1,24	1.96	5.83
2,10	1.95	5.81	1,1	1.02	9.91	1,25	1.04	9.80
2,11	2.29	6.47	1,2	1.89	5.26	1,26	1.01	6.46
2,12	1.00	5.43	1,3	1.00	5.65	1,27	1.00	7.83
2,13	3.58	6.78	1,4	1.00	6.84	1,28	1.05	4.80
2,14	1.06	5.96	1,5	1.04	6.86	1,29	1.01	8.95
2,15	1.00	5.43	1,6	1.02	6.11	1,30	6.74	6.14
2,16	4.71	5.37	1,7	1.00	4.23	1,31	3.59	7.55
2,17	2.09	5.49	1,8	1.04	7.46	1,32	1.02	4.96
2,18	2.21	5.88	1,9	1.21	5.40	1,33	1.72	5.85
2,19	2.00	5.73	1,10	1.00	4.97	1,34	1.42	6.02
2,20	1.90	5.66	1,11	4.27	6.23	1,35	1.94	6.02
			1,12	1.00	5.14	1,36	3.07	10.00

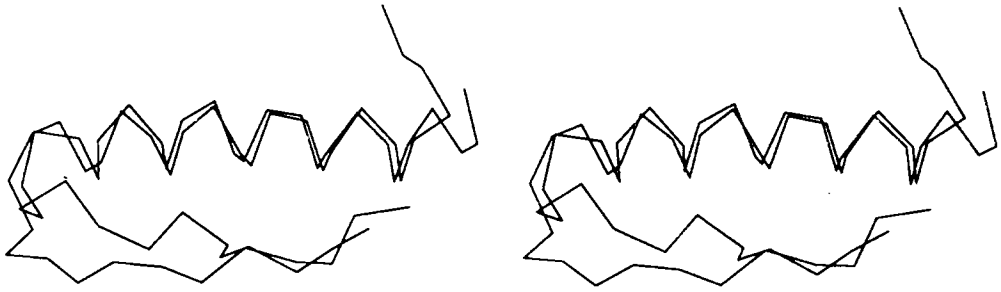

Figure 3 A stereo pair showing the best minimum and the crystal structure of 1ppt (middle and lower structures in Figure 2) with the helices superimposed. In our best minimal structure, the C-terminus of the helix continues an extra turn, and the extended strand hugs the helix more closely.

an rms deviation from the native of only 1.31 Å. Local minimization starting from each one of our library of 337 alternative conformations turned up a best conformation at −406.9 and rms 1.84 Å (Figures 2 and 3). A convenient way to produce many low potential conformers is with the EMBED algorithm,[4] where the upper and lower bounds are both taken to be just the corresponding interaction ρ, except for the longrange upper bounds, which were set to infinity. For 1ppt there are a few violations of the triangle inequality, and even after bound smoothing, trial sets of interresidue distances tend to be far from embeddable, but nonetheless, trial coordinates can be produced. Then instead of refining these by minimizing the violations of the original distance bounds, we simply minimized the potential. Generating 400 additional structures in this way produced at best minima of −380 and rms of 4.5 Å.

Energy embedding[20] proved to be difficult with this potential. The first step involves placing the n residues in $\mathcal{R}^{n-1}$, where there is probably only one minimum, and it will be extremely good because the molecule is enjoying the maximal number of conformational degrees of freedom. For 1ppt, $n = 36$ and the optimal set of trial distances is found by simply choosing the ρ for each interaction. Just as in the EMBED algorithm, we calculate the corresponding 36×36 metric matrix, and now find all 36 eigenvalues and eigenvectors, instead of just the three largest eigenvectors for embedding in $\mathcal{R}^3$. If all eigenvalues are strictly positive, one can calculate coordinates in the usual way that agree completely with the trial distances. We found there was of course one zero eigenvalue (only 35 dimensions are necessary for 36 points), but there were several substantial negative eigenvalues, indicating that the trial distances are not embeddable, i.e., do not correspond to a set of coordinates in any dimensional Euclidean space. Nevertheless, we calculated trial coordinates by using the absolute values of the eigenvalues, thus ensuring the coordinates spanned $\mathcal{R}^{35}$. Subsequent minimization of the potential with respect to all coordinates produced a very low minimum (−638) and a very compact structure (maximum interresidue distance of 11 Å), but it only spanned $\mathcal{R}^{13}$. It is easy to see that this is a feature of nonembeddable ρs, leading to stressed conformations in low-dimensional spaces. For example, suppose we had three points forming a triangle in the plane initially, but that $\rho_{13} > \rho_{12} + \rho_{23}$. Then minimizing the potential results in a linear configuration where point 2 is in the middle, the 1,2 and 2,3 distances are stretched, and the 1,3 distance is compressed. In any event, energy embedding then proceeds from the high-dimensional minimum toward $\mathcal{R}^3$ by gradually driving the fourth, fifth, etc., coordinates toward zero while otherwise keeping the potential as low as possible. The best we could achieve was −399.6 and rms of 7.52 Å (see Figure 2). Rotating the initial structure in $\mathcal{R}^{13}$ in various ways failed to produce either a lower potential value or a better rms deviation from the native.

At this stage of the development, it would be fair to say that we have found a very good potential function for mimicking the native conformation of 1ppt, and that we have tested its *conformational* properties (as opposed to its simulation of vibrational spectra, etc.) extraordinarily thoroughly, particularly in a global sense. In spite of representing entire amino acid residues by single isotropic spheres, we were able to build in quite realistic secondary structural preferences and roughly correct tertiary folding. The form of the potential is applicable to any protein, and it is conveniently quick to

evaluate and minimize, due to the very small number of variables used to represent the conformation. Furthermore, the method used to determine the parameters is rather objective and can be extended to developing a potential that will be useful for more than one protein. Such extensions and testing the predictive power of these potentials are the subject of work currently underway in our laboratory.

In order to present this account of how the problem was solved in the end, a great amount of effort was expended in exploring leads that did not work out. We would particularly like to thank M. Oobatake, V. N. Viswanadhan, and P. K. Ponnuswamy for their hidden but essential contributions to this project over the years. This work was supported by grants from the National Institutes of Health (GM37123), the National Science Foundation (DMB-8705006), and the University of Michigan Program in Protein Structure and Design.

REFERENCES

1. Creighton, T. E. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 5082–5086.
2. Baldwin, R. L. (1987) *Protein Structure, Folding, and Design 2,* UCLA Symposia on Molecular and Cellular Biology; new series vol. 69; Dale L. Oxender, ed.; Alan Liss, New York, pp. 313–320.
3. Crippen, G. M. (1981) *Distance Geometry and Conformational Calculations, Chemometrics Research Studies Series,* Vol. 1, Bawden, D., Ed., Research Studies Press, Wiley, New York.
4. Crippen, G. M. & Havel, T. F. (1988) *Distance Geometry and Molecular Conformation, Chemometrics Research Studies Series,* Bawden, D., Ed., Research Studies Press, Wiley, New York.
5. Crippen, G. M. (1975) *J. Comp. Phys.* **18,** 224–231.
6. Levitt, M. & Warshel, A. (1975) *Nature* **253,** 694–698.
7. Getzoff, E. D., Tainer, J. A. & Olson, A. J. (1986) *Biophys. J.* **44,** 191–206.
8. Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1986) *Biochemistry* **25,** 266–275.
9. Burley, S. K. & Petsko, G. A. (1986) *FEBS Lett.* **203,** 139–143.
10. Edwards, M. S., Sternberg, M. J. E. & Thornton, J. M. (1987) *Protein Eng.* **1,** 173–181.
11. Milner-White, E. J. (1988) *J. Mol. Biol.* **199,** 503–511.
12. Bryant, S. & Amzel, L. (1987) *Int. J. Peptide Protein Res.* **29,** 46–52.
13. Novotny, J., Rashin, A. A. & Bruccoleri, R. E. (1988) *Proteins Struct. Funct. Genet.* **4,** 19–30.
14. Sinha, S. K., Ram, J. & Singh, Y. (1985) *Physica A* **133A,** 247–280.
15. Sinha, S. K., Ram, J. & Singh, Y. (1977) *J. Chem. Phys.* **66,** 5013–5020.
16. Crippen, G. M. (1977) *Biopolymers* **16,** 2189–2201.
17. Oobatake, M. & Crippen, G. M. (1981) *J. Phys. Chem.* **85,** 1187–1197.
18. Efron, B. (1982) *The Jackknife, the Bootstrap and Other Resampling Plans,* Vol. SIAM Monograph No. 38, Regional Conference Series in Applied Mathematics, Soc. Ind. Appl. Math.
19. Bernstein, F. C., Koetzie, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
20. Crippen, G. M. & Ponnuswamy, P. K. (1987) *J. Comput. Chem.* **8,** 972–981.
21. Glover, I., Haneef, I., Pitts, J., Wood, S., Moss, D., Tickle, I. & Blundell, T. (1983) *Biopolymers* **22,** 293.
22. Bertsekas, D. P. (1982) *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York.
23. Crippen, G. M., Smellie, A. S. & Peng, J. W. (1988) *J. Chem. Inf. Comp. Sci.* **28,** 125–128.

Received March 21, 1989
Accepted July 12, 1989