# Determination of an Empirical Energy Function for Protein Conformational Analysis by Energy Embedding

**Gordon M. Crippen***

*College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109*

**P. K. Ponnuswamy**

*Department of Physics, Bharathidasan University, Tiruchirapalli-620 023, Tamilnadu, India*

*Received 21 March 1986; accepted February 3, 1987*

It is quite easy to propose an empirical potential for conformational analysis such that given crystal structures lie near local minima. What is much more difficult, is to devise a function such that the native structure lies near a relatively deep local minimum, at least in some neighborhood of the native in conformation space. An algorithm is presented for finding such a potential acting on proteins where each amino acid residue is represented by a single point. When the given structure is either an $\alpha$-helical, $\beta$-strand, or hairpin bend segment of pancreatic trypsin inhibitor, the resulting potential function in each case possesses a deep minimum within 0.10 Å of the native conformation. The improved energy embedding algorithm locates a marginally better minimum in each case only 0.1–1.3 Å away from the respective native state. In other words, this potential function guides a conformational search toward structures very close to the native over a wide range of conformation space.

## I. INTRODUCTION

In order to predict the conformation of a molecule, one generally carries out a *local* energy minimization, given some energy function and some reasonable starting conformation. Even if the energy function is very realistic, there are generally very many local minima, so that refining a good initial guess is quite practical, but otherwise one must choose many different starting conformations in order to locate the global minimum of energy (and any other local minima which are only slightly higher and are therefore physically important). As one deals with larger and larger molecules, this "multiple minima" problem worsens exponentially, so that for proteins, energy refinement of a very good structure (<1 Å from the native) is possible, but *a priori* prediction of the conformation is out of the question. Of course there is always the old debate whether the native structure of a protein corresponds to the global minimum of its free energy (even assuming one could calculate such a function realistically enough) or whether the native structure is kinetically determined. In this article, we take the empirical position that however the real protein does it, we would be satisfied with a possibly unrealistic potential function that embodies the correct result to the extent that a conformation near the crystal structure is one of the lowest local minima. Energy embedding[1] is a method that tends to find very low energy minima, but not necessarily the global minimum, subject to whatever additional geometric constraints one might have such as ring closures, radius of gyration, etc. The algorithm is explained in detail in the Methods section, but, basically, it begins by finding a local minimum of the given potential function in a very high dimensional Euclidean space where there are generally fewer local minima. Then it converts this starting conformation into a three-dimensional one, while preserving the geometric constraints and otherwise keeping the energy as low as possible. The trouble is that the energy embedding is almost too good! For example, the method applied to bovine pancreatic trypsin inhibitor (BPTI) using the Oobatake-Crippen potential, produces an unrealistically compact structure having an rms interresidue distance matrix deviation from the native of 7.0 Å, and an energy of −124 units.[2] In comparison, local minimiza-

Journal of Computational Chemistry, Vol. 8, No. 7, 972–981 (1987)

CCC 0192-8651/87/070972-10$04.00

tion in three dimensions ($E_3$) starting at the crystal structure moves only 1.2 Å, and has an energy of −119. Similar calculations with an isolated hairpin loop (residues 14–38 of BPTI) showed that energy embedding consistently reached conformations much better in energy than the native, by forming extremely compact structures with no recognizable secondary structure. The conclusion was that the potential needed special strong local interactions to properly maintain secondary structure under energy embedding, even though they were not necessary in order to agree with the crystal structure under ordinary local minimization.

We subsequently developed the Crippen–Viswanadhan potential, which still approximates each amino acid residue as a point located at the $C^{\alpha}$ atom, and considers only isotropic pairwise interactions.[3] It passed a number of exacting tests, such as having a local minimum near the native structure for several different small proteins, returning to the native by energy minimization after conformational perturbations < 1 Å, and preserving secondary structural features under local minimization. However, once again, energy embedding applied to the 26 residue helical protein mellitin produced an extremely compact structure differing from the native by 10.4 Å, and having an energy 31.5 kcal lower. Detailed analysis of energy embedding on the C-terminal six residue helix of mellitin alone, showed that virtual bond lengths and angles along the chain were correctly maintained, but the overall conformation was too compact, although energetically (−6.1 kcal) much better than the native −4.3 kcal. An exhaustive search over all conformation space for such a six residue freely jointed chain in 30° increments located a conformer with energy = −6.2 kcal, and strong conformational resemblance to the energy embedding result. In other words, the embedding algorithm was performing very well, but the given potential function did not correspond adequately to reality over large regions of conformation space, even though it did very well quite near the native.

Clearly, energy embedding is a powerful method for conformational calculations, but it puts unprecedented demands on the potential function. In order to find a better potential, it is necessary to choose its functional form and then adjust the parameters in it for best performance. Energy embedding must have the conformational energy consist of interactions between isotropic points, but the points could represent atoms or groups of atoms. In this work, the goal is to calculate the general polypeptide tertiary structure, so having one point per residue is probably adequate, and having a smaller number of points reduces computer time. Specifically, we require that the total energy $U$ consist of a sum over all pairwise interactions between the $n$ residues, with each interaction depending on the distance $d_{ij}$ between the residues and possibly other factors $\{p\}$, such as sequence separation and residue types.

$$U = \sum_{i<j}^{n} e(d_{ij}; \{p\}) \tag{1}$$

That excludes in particular any torsional terms. In this work, we have chosen to group the terms of eq. (1) so as to separate the short-range interactions that are important for secondary structure from the long-range interactions:

$$U = \sum_{i=1}^{n-1} e(d_{i,i+1}) + \sum_{i=1}^{n-2} e(d_{i,i+2}; t_{i+1}) + \sum_{i=1}^{n-3} e(d_{i,i+3}; t_{i+1}, t_{i+2}) + \sum_{i=1}^{n-4} \sum_{j=i+4}^{n} e(d_{i,j}, t_i, t_j) \tag{2}$$

where $t_i$ is the type (Gly, Ala, ..., Val) of the $i$-th amino acid residue. Much of the justification for this functional form has already been given.[3] The novel feature is that each interaction term has the same form:

$$e(d_{i,i+1}) = \left(\frac{a_1}{d_{i,i+1}^4}\right) - \left(\frac{b_1}{d_{i,i+1}^2}\right) \tag{3}$$

$$e(d_{i,i+2}; t_{i+1}) = \left(\frac{a_{2,t_{i+1}}}{d_{i,i+2}^4}\right) - \left(\frac{b_{2,t_{i+1}}}{d_{i,i+2}^2}\right)$$

$$e(d_{i,i+3}; t_{i+1}, t_{i+2}) = \left(\frac{a_{3,t_{i+1}} + a_{3,t_{i+2}}}{d_{i,i+3}^4}\right) - \left(\frac{b_{3,t_{i+1}} + b_{3,t_{i+2}}}{d_{i,i+3}^2}\right)$$

$$e(d_{ij}; t_i, t_j) = \left(\frac{a_{l,t_i} + a_{l,t_j}}{d_{ij}^4}\right) - \left(\frac{b_{l,t_i} + b_{l,t_i}}{d_{ij}^2}\right)$$

Note that where two residue types are thought to be important in characterizing the interaction, the corresponding $a$'s are simply

added, and similarly with the $b$'s. This keeps $U$ linear in the $a$'s and $b$'s while requiring only 20 of each (one for every residue type) instead of 400 (one for every residue type *pair*). An alternative way of parameterizing such an interaction is directly in terms of the depth of the well, $w$, at the distance of optimal approach, $d_{opt}$.

$$\left(\frac{A}{d_{ij}^4}\right) - \left(\frac{B}{d_{ij}^2}\right) = w\left[\left(\frac{d_{opt}}{d_{ij}}\right)^4 - 2\left(\frac{d_{opt}}{d_{ij}}\right)^2\right] \quad (4)$$

These two forms are readily interconvertible, and it is especially useful to note that

$$d_{opt} = \left(\frac{2A}{B}\right)^{1/2} \quad (5)$$

Given this somewhat arbitrary choice, dictated in part by a desire to approximate physical reality and in part by demands of mathematical simplicity, it remains to determine the values of the 122 $a$s and $b$s. Starting from the general idea that the native structure should have very low energy, there are several ways to formulate this notion mathematically, remarkably few of which work. For example, a necessary (but not sufficient) condition for the native structure to have minimal energy is that the gradient for that conformation be zero:

$$\|\nabla U\|_{nat} = 0 \quad (6)$$

Unfortunately, this is trivially achieved in terms of eq. (4) by $A, B \to 0$, or equivalently $w \to 0$. The resultant potential is useless because $U \equiv 0$ for all conformation space. Alternatively, if we minimize $U_{nat}$ with respect to the parameters subject to the constraint of eq. (6) while requiring all $w > 1$ and $d_{opt} > 0$, there is in general no solution. If instead we attempt to minimize with respect to parameters the magnitude of the gradient at the native subject to the physically reasonable constraints that all $w > 1$ and $d_{opt} > 0$, then most $w \to 1$ and $d_{opt} \to 0$, once again leaving the potential surface trivially flat. The basic conflict is that the surface must be flat only very near the native conformation, not globally; and there must be some interactions with strong weight in order to differentiate between native and nonnative conformations, but there is no point to arbitrarily scaling up interaction strengths because only the relative energies of conformations are important.

## II. NUMERICAL METHODS

The first order of business is how to formulate and solve the search for energy parameters. From eqs. (3) and (2), we see that $U$ is linear in the $a$s and $b$s. In order that each interaction have a separation of minimal energy, it must also be true that all $a, b \geq 0$. If there is a perturbed (non-native) conformer, we want it to have a higher energy than the native. Presumably, the further away it is from the native, the worse its energy should be. Let $\delta_{pert,nat}$ be the rms deviation between the two. Now if the perturbed structure has a slightly higher energy than the native does, we can trivially make the difference as large as we want by multiplying all $a$s and $b$s by some large positive number: To counterbalance this effect, we seek to keep the parameters small. Precisely formulated, this all amounts to a linear program:

$$\text{minimize } \Sigma a + b \quad (7)$$

subject to

$$U_{pert} - U_{nat} \geq \delta_{pert,nat} \text{ for all } pert$$

and

$$a, b \geq 0$$

It turns out to speed the calculations considerably if some *a priori* constraints are also included:

$$d_{opt,i,i+1} = 3.8 \text{ Å} \quad (8)$$

and

$$4 \leq d_{opt,i,j} \leq 20 \quad \forall\, i,j$$

These amount to assuming always *trans* peptide bonds, requiring that the observed distance (~4 Å) of closest approach between C$^\alpha$s never be violated, and that no interactions be entirely repulsive. This same idea of forcing a rough valley in the energy surface by such linear constraints had been tried earlier[3], but not very successfully because often the inequalities corresponding to important perturbed structures are nearly linearly dependent, and the standard simplex algorithm becomes numerically unstable. Instead, we solved eq. (7) by the highly stable revised simplex algorithm.[4]

The perturbed structures referred to in eq. (7) were generated in two ways. Initially, the native structure was perturbed by incrementing and decrementing in turn by 5%

each $d_{i,i+1}$. Taking all these inequalities in eqs. (7) and (8) into account and solving for the parameters $a_1$, $b_1$, $a_{2,Gly}, \ldots$, $b_{l,Val,Val}$, results in the current potential function, eq. (2). Starting at the native conformation and minimizing the potential with respect to the $C^\alpha$ Cartesian coordinates in $E_3$ generally results in a new perturbed conformation at some distance $\delta$ from the native with an energy lower than the native, which is undesirable. Introducing this new inequality constraint and resolving the linear program produces a new set of parameters. Some may have increased while others decreased, but in general the new sum over all parameters is slightly greater than before. New constraints were introduced in this fashion one by one until $\delta \leq 0.1$ Å. At the end of this process, we have found the flattest potential surface such that *all* the perturbed structures are higher than the native, and indeed, the further they are in conformation from the native, the higher their potential. Some kind of reason has been found for each perturbed structure so that its intramolecular interactions are worse than those of the native, on the whole. These reasons are not stated explicitly by the algorithm, and might be hard to deduce by inspecting the final parameters, because relatively good interactions in one part of a perturbed structure may be cancelled out by bad ones elsewhere.

Parameters were derived for most of the runs listed in Table II by the above procedure. However, a few cases of a nine residue segment employed a slight variation in the way constraints are generated. The reasoning is that when we consider local structures in proteins, such as segments of $\alpha$-helix, $\beta$-strand, and bends, probably the most important factors in determining the conformation of that segment are interactions within it and between it and its immediate surroundings, rather interactions within the remainder of the protein. Denoting the residues of the segment by $S$ and those of the rest of the protein by $\overline{S}$, the total energy

$$U_{tot} = U(S) + U(S,\overline{S}) + U(\overline{S})$$

where $U(S)$ is a sum over all interactions *within* the segment $S$, and $U(S,\overline{S})$ is the sum of interactions between residues in $S$ and the remainder of the protein, $\overline{S}$. Then we claim $U(\overline{S})$ has a negligible influence on the conformation of $S$ compared to $U(S)$ and $U(S,\overline{S})$. Therefore we demand

$$U(S_n) + U(S_n,\overline{S}_n) + \delta(S_n,S_p) \leq U(S_p) + U(S_p,\overline{S}_p) \quad (9)$$

where $S_n$ and $S_p$ represent the native and perturbed states respectively of the segment $S$; and $\delta(S_n,S_p)$ is the rms deviation between these two states of the local segment. Adding such inequalities to the usual ones after each perturbed structure is found, produces parameters which tend to strengthen the segment's preference for its native structure independent of the rest of the protein. Since the inequalities involving the whole protein (eq. 7) are included as before, the resulting parameters have taken into consideration the conformational demands of residues in all parts of the protein.

The energy minimizations were carried out by Newton's method with analytical first and second derivatives acting on one residue at a time, cycling repeatedly down the chain of residues until the magnitude of the gradient was sufficiently small. In practice, this is an extremely satisfactory method of minimization[5], clearly superior to steepest descents, conjugate gradients, and even Newton's method working on all points simultaneously.

The second item of methods is the substantial improvements on the energy embedding algorithm since the technique was last described.[1] The first step is to find coordinates, $c_{i,k}, i = 0,\ldots,n-1, k = 0,\ldots,n-2$, in $E_{n-1}$ for the $n$ residues. For the initial guess, we place the residues at the corners of a regular hyper-tetrahedron with edge $r = 5.0$ Å.

$$c_{i,k} = 0, \qquad k \geq i \quad (10)$$

$$c_{i,k} = \frac{1}{i}\sum_{j=0}^{i-2} c_{j,k}, \qquad k < i-1$$

$$c_{i,i-1} = \left(r^2 - \sum_{k=0}^{i-2} c_{i,k}^2\right)^{1/2}$$

The initial guess is refined by Newton minimization, moving one residue at a time. Although Newton's method calculates both a direction to move the residue and a step size along that direction, when far from the minimum, these estimates may not always lead to an energy decrease. The Armijo linesearch has proven to be very reliable for modifying the Newton step when necessary.[6]

Once the $n$ points are embedded in $E_{n-1}$, it is necessary to project the structure down to

$E_3$. The original method went to some pains to decide what three dimensional subspace should be preserved[1], but subsequent experience has shown that the choice is generally not very clear-cut. Instead, we simply demand that the fourth and higher coordinates of each point approach zero while otherwise trying to keep the energy minimal. To put it more precisely, let the coordinates matrix $C = (c_{ij})$, for points $i = 1, \ldots, n$ and coordinates $j = 1, \ldots, n-1$. Then given the initial minimal energy conformation $C_{init}$ in $E_{n-1}$,

$$\underset{C}{\text{minimize}}\ U(C) \tag{1}$$

subject to

$$c_{ij} = 0 \quad \text{for } i = 1, \ldots, n; \qquad j = 4, \ldots, n-1$$

This is simply a nonlinear constrained optimization problem which can be nicely solved by augmented Lagrangians[7]. For simplicity of notation, let the vector of all unwanted coordinates be $\mathbf{g} = (c_{1,4}, c_{1,5}, \ldots, c_{n,n-1})$. Then in this case the augmented Lagrangian function is

$$L(C, \lambda, w) = U(C) + \boldsymbol{\lambda} \cdot \mathbf{g} + \frac{1}{2} w g^2 \tag{12}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrange multipliers and $w$ is a weight on the constraint terms. Initially in iteration 0, $w^{(0)} = 0.01$ and $\lambda^{(0)} = 0$. Then in iteration $k$, minimize $L(C^{(k)}, \lambda^{(k)}, w^{(k)})$ with respect to $C$ by the usual Newton method, resulting in $C^{(k+1)}$ and $\mathbf{g}^{(k+1)}$. Now update

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} + w^{(k)} g_i^{(k+1)} \tag{13}$$

$$w^{(k+1)} = \begin{cases} 2w^{(k)} & \text{if } \|\mathbf{g}^{(k+1)}\| > \|\mathbf{g}^{(k)}\| \\ w^{(k)} & \text{otherwise} \end{cases}$$

The cycles of minimization (eq. 12) and updating (eq. 13) are repeated until the undesired coordinates are reduced essentially to zero ($\|\mathbf{g}\| < 0.01$ Å). During this process, $w$ generally gradually increases, which converts $L$ from a standard Lagrangian to a penalty function. The method is more robust than the unmodified Lagrangian approach, but by having Lagrange multipliers and by introducing the penalty terms gradually, convergence is faster than simply minimizing $U + wg^2$ for large $w$. If the initial conformation in $E_{n-1}$ was really well minimized, then $U$ increases as $\|\mathbf{g}\| \to 0$ because the molecule cannot make as many favorable interactions in a lower dimensional space. Sometimes, the repeated cycles of minimization of $L$ offer such additional opportunities for reducing $U$, that the energy actually slightly decreases. Meanwhile, the various components of $\boldsymbol{\lambda}$ increase from zero, reflecting the energetic cost of reducing the corresponding $g_i$ to zero, but eventually $\boldsymbol{\lambda} \to 0$ because there is no force component acting in the fourth and higher coordinate directions when all points lie in $E_3$.

In this work, the only constraints are that $\mathbf{g} = 0$, but additional experimental constraints could easily be included in eq. (12) as long as they could be expressed as (nonlinear) equality or inequality constraints on the coordinates. In other words, this augmented Lagrangian approach is a generally useful method for energy embedding.

## III. RESULTS

The objective of this paper is to determine a set of energy parameters such that energy embedding can locate the native structure of at least one molecule, namely the one with which the parameters were derived. (Using the potential to correctly predict the conformations of other molecules is a more ambitious goal for the future). As explained in the **Introduction**, our earlier empirical potentials guided energy embedding toward protein structures that were too compact. Hence, finding parameters that would mimic local structures, such as $\alpha$-helix, $\beta$-strand, and bends, would be of considerable importance. Also, as the computer time for the repeated energy minimizations is roughly proportional to the cube of the number of residues, we restricted the present study to the following four segments of BPTI: (1) residues 47–51 (short $\alpha$-helix), (2) residues 14–20 (short extended structure), (3) residues 22–30 (tight hairpin consisting of the bend and two arms), and (4) residues 47–56 (a longer $\alpha$-helix). In what follows, we first describe the results of the latter three larger segments in detail, and then consider the short $\alpha$-helix to illustrate how the successive addition of linear inequalities raises the energy of all of conformation space above that of the native until at the end, the native is at the global minimum.

**Residues 14–20 (7-mer):** The linear program search for parameters for this mo-

lecular unit began with 39 initial constraints, and then added in turn 97 more constraints because the current parameters allowed energy minimization to move the conformation away from the native by 4.5 Å to 0.08 Å. The last set of parameters formed a local minimum only 0.08 Å from the native, so the procedure halted normally. As one might expect from so small a movement, the native has energy = −179.90 (arbitrary units), while the minimized structure improved only slightly to −179.99 units. According to eqs. 2 and 3, there are only 28 parameters involved in this 7-mer. The energy function derived is conservative in that 24 out of 28 parameters were nonzero, as shown in Table I.

It is informative to compare the optimal distance of approach (eq. 5) for every residue pair with the native distances (eq. 14):

$$D_{opt} = \begin{bmatrix} 0.0 & 3.8 & 5.9 & 17.0 & 11.1 & 11.1 & 11.4 \\ 3.8 & 0.0 & 3.8 & 6.1 & 16.4 & 20.0 & 13.6 \\ 5.9 & 3.8 & 0.0 & 3.8 & 4.0 & 20.0 & 13.6 \\ 17.0 & 6.1 & 3.8 & 0.0 & 3.8 & 5.9 & 20.0 \\ 11.1 & 16.4 & 4.0 & 3.8 & 0.0 & 3.8 & 5.9 \\ 11.1 & 20.0 & 20.0 & 5.9 & 3.8 & 0.0 & 3.8 \\ 11.4 & 13.6 & 13.6 & 20.0 & 5.9 & 3.8 & 0.0 \end{bmatrix} \quad (14)$$

$$D_{native} = \begin{bmatrix} 0.0 & 3.8 & 6.0 & 9.0 & 10.4 & 13.3 & 14.2 \\ 3.8 & 0.0 & 3.8 & 6.5 & 9.1 & 12.0 & 13.9 \\ 6.0 & 3.8 & 0.0 & 3.8 & 6.2 & 9.6 & 11.8 \\ 9.0 & 6.5 & 3.8 & 0.0 & 3.8 & 6.4 & 9.4 \\ 10.4 & 9.1 & 6.2 & 3.8 & 0.0 & 3.8 & 6.2 \\ 13.3 & 12.0 & 9.6 & 6.4 & 3.8 & 0.0 & 3.8 \\ 14.2 & 13.9 & 11.8 & 9.4 & 6.2 & 3.8 & 0.0 \end{bmatrix}$$

Notice how the optimal distance between two residues is not always the native distance, and indeed there is considerable tension in some regions. This implies that there was not an overabundance of parameters compared to conformational degrees of freedom. Indeed, finding parameters would be trivial if there was a unique $a$-$b$ pair for every native distance. In that case, simply choose the $a$ and $b$ so that $d_{opt} = d_{ij}$ for all $i$ and $j$, and thus, the native structure would be the global minimum. We cannot do this because the same interaction term and same parameter pair may be invoked in two different interactions having different native distances. For example, $d_{1,5}$ and $d_{1,6}$ both require $a_{l,Ile}$, $b_{l,Ile}$, $a_{l,Cys}$, and $b_{l,Cys}$.

**Table I.** The Relevant Energy Parameters Used by the Extended Segment BPTI 14–20. Notation as in eq. (3).

| | | | |
|---|---|---|---|
| $a_1$ | = 3284.8062 | $b_1$ | = 460.2432 |
| $a_{2,Ala}$ | = 24344.2343 | $b_{2,Ala}$ | = 1295.6050 |
| $a_{2,Ile}$ | = 16500.7636 | $b_{2,Ile}$ | = 932.3348 |
| $a_{2,Lys}$ | = 32168.2304 | $b_{2,Lys}$ | = 1817.6440 |
| $a_{2,Arg}$ | = 1422.0629 | $b_{2,Arg}$ | = 177.4268 |
| $a_{3,Ala}$ | = 263.7835 | $b_{3,Ala}$ | = 32.9705 |
| $a_{3,Ile}$ | = 7840.8765 | $b_{3,Ile}$ | = 39.1601 |
| $a_{3,Lys}$ | = 15903.8115 | $b_{3,Lys}$ | = 79.4740 |
| $a_{3,Arg}$ | = 12620.0244 | $b_{3,Arg}$ | = 63.1821 |
| $a_{l,Ala}$ | = 0.0808 | $b_{l,Ala}$ | = 0.0101 |
| $a_{l,Ile}$ | = 22.3884 | $b_{l,Ile}$ | = 3854.9218 |
| $a_{l,Cys}$ | = 3969.0701 | $b_{l,Cys}$ | = 497.4216 |
| $a_{l,Lys}$ | = 0.0006 | $b_{l,Lys}$ | = 0.0050 |
| $a_{l,Arg}$ | = 95682.9687 | $b_{l,Arg}$ | = 1027.1456 |

The next step is to locate a good minimum in $E_6$. As explained in the **Methods** section, the molecule is first placed as a regular simplex, where each residue is 5 Å from all the others. Such an arbitrary conformation is not at all favorable, having energy = +406.3, but subsequent minimization by 17 iterations of Newton's method brings the energy down to −179.99, which happens also to be the energy of the near-native minimum in $E_3$. In fact, it has no distance which differs more than 0.2 Å from the corresponding native distance. Although some of the largest coordinates are in the 4th, 5th, and 6th dimensions, this conformer nearly lies in a three dimensional subspace of $E_6$.

Now the energy embedding procedure attempts to drive the unwanted coordinates to zero. In this case, convergence would have been faster if the structure had been rotated in $E_6$ so that the unwanted coordinates were as small as possible. However, in general, the structure tends to be so nearly spherical that such a procedure is of little help. It took 8 cycles (see Eq. 13) to reduce the sum of the squares of the unwanted coordinates $g^2$ to 0.03, while the weight $w$ increased from 0.01 to 0.64. All coordinates changed substantially, and the energy remained constant at −179.99. The rms deviation between the final structure and the native was 0.09 Å, which essentially amounts to the same minimum (Fig. 1).

**Residues 22–30 (9-mer):** The linear programming and the energy embedding results

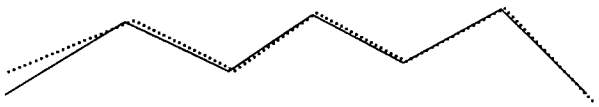


**Figure 1.** Least squares superposition of the native extended fragment (residues 14–20) in dotted lines upon the conformation calculated by energy embedding (solid lines).

for the 9-mer and all other segments studied are presented in Table II.

The 9-mer segment involves an open loop structure connecting two extended parts and, hence, serves as a typical case for bend structures in proteins. We used this segment for four different runs of the linear programming procedure. In the first run, as in the 7-mer study, constraints were added one at a time considering the whole segment while computing the constraint (see eq. 7). In run 2, two constraints were added at a time, the first constraint being made of the whole segment, as in run 1, and the second constraint being computed using only the four central residues of the segment, which correspond to the real bend part of the system (see eq. 9). The latter constraint, thus, gives specific weight to the local bend structure while the whole segment energetic considerations are handled as before by the former. In run 3, again two constraints are added at a time, the second one now being computed using only the first three and the last three residues of the segment. This case corresponds to giving additional weight to the end arms of a bend structure. In run 4, we included only the second constraint of run 3. All four runs began with the same 57 initial constraints, which produced energy parameters that had a minimum 2.1 Å from the native conformation with energy −105.3 units. From there, however, each run progressed characteristically until termination when local minimization from the native moved <0.1 Å. The number of constraints needed and the energies of the final structures also differed accordingly. While the parameters derived from runs 1, 2, and 3 were able to embed their structures with an rms of 0.8 − 1.2 Å, the run 4 potential guided energy embedding to a final conformation with an rms of only 0.5 Å. This is extremely close to the native, as shown in Fig. 2. These results are consistent with the belief that a tight bend forms not because the four bend residues inherently prefer that conformation, but rather because the two strands thus brought together interact favorably. Out of the 42 expected parameters, run 3 produced 36 nonzero values, given in Table III.

We are particularly keen to pursue further the improvement run 4 represents over the

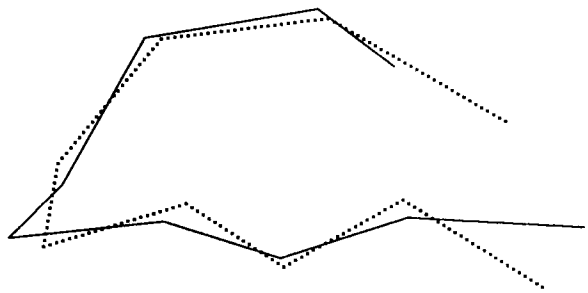


**Figure 2.** Least squares superposition of the native hairpin bend fragment (residues 22–30) in dotted lines upon the conformation calculated by energy embedding (solid lines).

**Table II.** Linear Programming and Energy Embedding Results of the Segments of BPTI

| Residue segment (BPTI) | Linear programming | | | | | Energy embedding | |
|---|---|---|---|---|---|---|---|
| | Run | Status | Constraints | Energy[a] | rms[b] | Energy[a] | rms[b] |
| 22–30 | | Initial | 57 | −105.34 | 2.10 | | |
| | 1 | Final | 147 | −187.02 | 0.09 | −188.39 | 1.19 |
| | 2 | Final | 291 | −164.36 | 0.08 | −165.52 | 1.21 |
| | 3 | Final | 291 | −364.13 | 0.10 | −362.64 | 0.82 |
| | 4 | Final | 197 | −159.71 | 0.08 | −159.65 | 0.51 |
| 47–56 | | Initial | 65 | −159.56 | 3.39 | | |
| | | Final | 232 | −169.44 | 0.09 | −172.60 | 1.09 |
| 14–20 | | Initial | 39 | −66.19 | 4.47 | | |
| | | Final | 145 | −180.00 | 0.08 | −179.99 | 0.09 |
| 47–51 | | Initial | 25 | −36.81 | 0.87 | | |
| | 1 | Final | 44 | −55.71 | 0.07 | −55.74 | 0.36 |
| | 2 | Final | 79 | −62.00 | 0.03 | −61.94 | 0.16 |

[a] arbitrary units
[b] Å

other runs due to including extra constraints on subsets of residues.

**Residue 47–56 (10-mer):** The results for the long $\alpha$-helix are also very encouraging. As seen in Table II, 167 additional constraints were needed to obtain the final structure, which resembled the native in all aspects. The resulting energy parameters, when used by energy embedding, generated a molecule only 1.09 Å away from the native. The superposition of the two structures in Fig. 3 shows how the majority of the difference occurs at the end of the helix, rather like the way a real helix tends to unwind.

**Residues 47–51 (5-mer):** So far we have demonstrated that our linear programming procedure produces potential functions that certainly have local minima, but apparently near the native there is a relatively deep one, and energy embedding converges either on that or a nearby minimum that is nearly as good. Pragmatically speaking, our only concern is that energy embedding should function this well, given the energy parameters. However, with a segment as small as the 5-mer, we can also globally examine conformation space to see whether the native is indeed the global minimum, and what the energy looks like elsewhere. As can be seen in Table I, there is always a strong virtual bond stretching term in these parameter sets ($a_1$ and $b_1$), so obviously any deviation from 3.8 Å is strongly penalized. Therefore, we searched all conformation space viewing the 5-mer as a freely jointed chain, having 3 vicinal bond angles and 2 torsional angles as variables. These five degrees of freedom were searched in 30° intervals, although essentially the same results were found with a 20° step size. The linear programming procedure started with 25 initial constraints and an rms of 0.87 Å, and positioned the minimum energy structure close enough to the native after 19 cycles, when no subsets were used (run 1). The final energy parameters led to an energy embedding result 0.36 Å away from the native. Alternatively, including an extra constraint about a subset consisting of residues 1, 2, and 3 (run 2), required 27 cycles, but the parameters produced led to an energy embedded conformer only 0.16 Å away from the native. As shown in Fig. 4, the successive parameter sets produced in the 19 cycles of run 1 tend to cup the energy surface more and more strongly until finally the global search is unable to find any conformations having lower energy than the native, and there is a very strong positive correlation between higher energy and greater distance in conformation space from the native. Although the fraction of conformations found with energy greater than that of the native progressed from 0.876 to 1, the progression was not entirely monotonic throughout the 19 cycles.

**Table III.** The Relevant Energy Parameters Used by the Hairpin Segment BPTI 22–30. Notation as in eq. (3).

| | | | |
|---|---|---|---|
| $a_1$ | = 2387.2639 | $b_1$ | = 327.5735 |
| $a_{2,Gly}$ | = 5354.3867 | $b_{2,Gly}$ | = 349.6054 |
| $a_{2,Ala}$ | = 1547.9705 | $b_{2,Ala}$ | = 103.7376 |
| $a_{2,Leu}$ | = 630.6378 | $b_{2,Leu}$ | = 3.1521 |
| $a_{2,Tyr}$ | = 12223.8789 | $b_{2,Tyr}$ | = 620.4725 |
| $a_{2,Lys}$ | = 0.0 | $b_{2,Lys}$ | = 0.0 |
| $a_{2,Asn}$ | = 290.3320 | $b_{2,Asn}$ | = 36.2993 |
| $a_{3,Gly}$ | = 236.8990 | $b_{3,Gly}$ | = 1.1572 |
| $a_{3,Ala}$ | = 1.3327 | $b_{3,Ala}$ | = 0.0 |
| $a_{3,Leu}$ | = 0.00043 | $b_{3,Leu}$ | = 0.00346 |
| $a_{3,Tyr}$ | = 136.4423 | $b_{3,Tyr}$ | = 0.0 |
| $a_{3,Lys}$ | = 96.7723 | $b_{3,Lys}$ | = 11.3488 |
| $a_{3,Asn}$ | = 2711.0904 | $b_{3,Asn}$ | = 13.5517 |
| $a_{1,Gly}$ | = 1127.3878 | $b_{1,Gly}$ | = 5.6410 |
| $a_{1,Ala}$ | = 1733.6178 | $b_{1,Ala}$ | = 8.6677 |
| $a_{1,Leu}$ | = 3870.8357 | $b_{1,Leu}$ | = 203.2917 |
| $a_{1,Cys}$ | = 354.1758 | $b_{1,Cys}$ | = 44.3326 |
| $a_{1,Phe}$ | = 421.1347 | $b_{1,Phe}$ | = 2.1081 |
| $a_{1,Tyr}$ | = 403.0842 | $b_{1,Tyr}$ | = 43.2526 |
| $a_{1,Lys}$ | = 0.9359 | $b_{1,Lys}$ | = 0.0135 |
| $a_{1,Asn}$ | = 12081.9453 | $b_{1,Asn}$ | = 775.1811 |

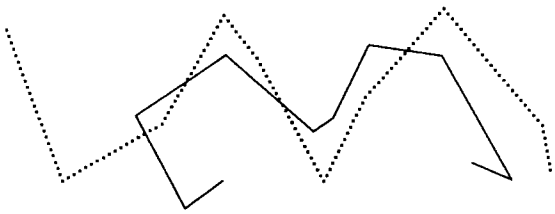
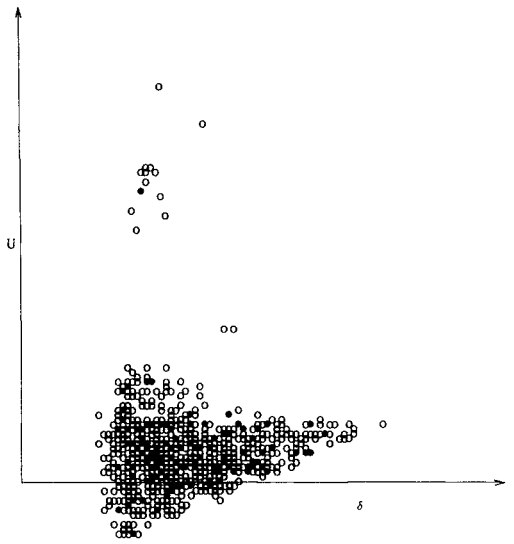


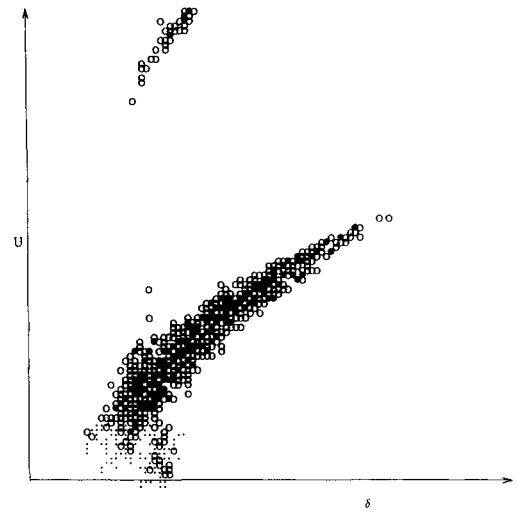
**Figure 3.** Least squares superposition of the native helical fragment (residues 47–51) in dotted lines upon the conformation calculated by energy embedding (solid lines).

## IV. CONCLUSIONS

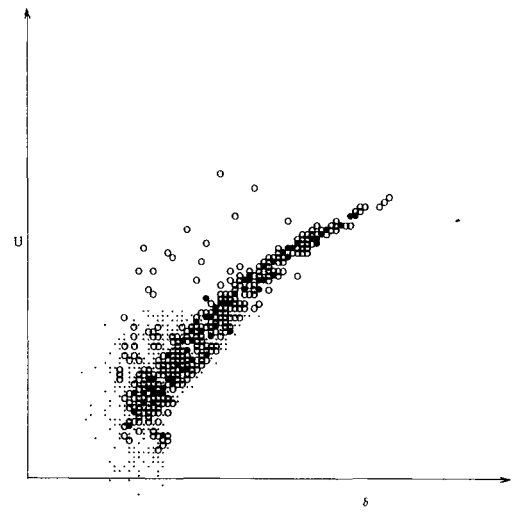
This work adds to the evidence that energy embedding is a very powerful method for locating very low local minima, often nearly as good as the global minimum. The algorithm is clearly not biased toward overly compact structures, but rather it is guided to



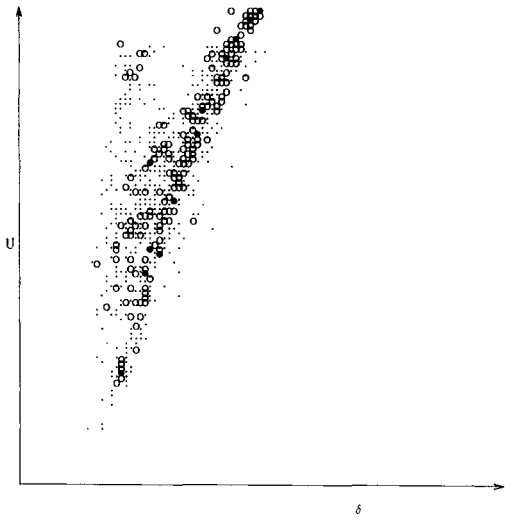

(a)




(b)




(c)




(d)

**Figure 4.** Global survey of the conformation space of the 5-mer, treated as a freely jointed chain. Each conformation is plotted according to its energy, $U$, (where $U_{native} \equiv 0$) and its rms conformation deviation from the native, $\delta$. The $U$ axis runs from 0 to 5 arbitrary energy units, and the $\delta$ axis runs from 0 to 5 Å. The darker the symbol, the more densely the surrounding region of the two-dimensional histogram is populated with conformations. (a) Initial parameter set, 25 inequalities; (b) parameter set from 29 inequalities; (c) parameter set from 39 inequalities; (d) final parameter set, 44 inequalities.

whatever the given potential prefers. Success in conformational analysis using this technique depends critically on having a potential function that tends to favor the correct conformations over a rather wide range of alternative structures. Although the form of the potential chosen in eqs. (2) and (3) would allow it to be applied to any sort of polypeptide chain, one can hardly expect the parameter values in Table I, for example, to be useful for any molecule beyond the particular heptapeptide from which they were derived. Until large numbers of interacting residues are used to derive the parameters, the values shown in Tables I and III are just artifacts of linear programming. If the particular oligopeptide for which they were derived does not have an example of a particular interaction, the corresponding parameters will be zero, such as $a_{2,\mathrm{Lys}}$ and $b_{2,\mathrm{Lys}}$ in Table III. Even parameters which are positive will vary greatly each time a new inequality is added because the new inequality has excluded the vertex of the old feasible region, and the simplex algorithm moves off to a substantially different vertex. What does vary gradually

after most of the inequalities have been added, is the slowly increasing *sum* of all parameters. It is, therefore, not clear at this point what values the parameters will approach when derived from one or more whole proteins.

We think the real import of this paper is the very general method for producing energy parameters. As long as the potential is linear in its parameters, and there are one or more molecules with "desired" conformations as opposed to undesirable ones generated in some fashion, then the linear programming method we have outlined will work. One possible outcome is that the set of inequalities in eq. (7) is inconsistent, which unequivocally demonstrates that the desired potential function cannot be constructed. (Of course, possibly a different functional form would succeed, but this is an arbitrary choice by the investigator outside the scope of the algorithm). The other possible outcome is that further perturbation from the native structure by energy minimization produces very little movement, and the algorithm terminates, as in the test cases shown above. Perhaps very many local minimizations would be required, but by construction, the algorithm must either converge eventually or find inconsistent inequalities. There is no guarantee that satisfying inequalities derived from local perturbations of conformation will eliminate the possibility of some more distant structure being significantly better than the native, although the test cases were fortunate in this respect. When even the experts at devising potential functions confess it is a difficult task, a simple systematic approach such as this may be a great help.

Although this work shows that an empirical function can be constructed to mimic an apparently arbitrary native conformation, clearly the next step is to apply the method to a whole protein, where the number of conformational degrees of freedom exceeds the number of parameters available. There are only 122 parameters altogether, so a real protein having more than 43 residues would have more degrees of freedom than energy parameters, and would thus constitute a somewhat more realistic test. This is a nontrivial programming task because simply repeatedly solving the linear program in these ill-conditioned cases is extremely time consuming for large numbers of constraints. If a set of parameters can be found at all, it must be verified that energy embedding with this potential reproduces the native structure for that same protein. If that should work, the next step would be to use the same potential to embed other proteins, comparing the calculated conformations with their known crystal structures. At least there is the encouraging result that this functional form and single point per residue representation were successful in several cases of various conformations. It remains to be shown that a generally useful potential can be constructed along these lines.

This work was supported by grants from the National Institutes of Health (R01-GM30561) and the National Science Foundation (PCM-8314998). Part of this work was carried out while one of us (GMC) was the recipient of a Fulbright Senior Scholar Fellowship administered by the Australian-American Educational Foundation. We are very grateful for the support of the host institutions, The School of Pharmaceutical Chemistry, Victorian College of Pharmacy Ltd. and the La Trobe University Chemistry Department, both in Melbourne, Australia. PKP thanks the authorities of Bharathidasan University for granting the sabbatical leave to associate with GMC at the University of Michigan.

## References

1. G.M. Crippen, *J. Comput. Chem.*, **3,** 471 (1982).
2. G.M. Crippen, *Biopolymers,* **21,** 1933 (1982).
3. G.M. Crippen and V.N. Viswanadhan, *Int. J. Peptide Protein Res.,* **24,** 279 (1984).
4. D.G. Luenberger, *Introduction to Linear and Nonlinear Programming,* Addison-Wesley, Reading, Massachusetts, 1973.
5. N.L. Allinger, *personal communication* (1984).
6. D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York, 1982, p. 20.
7. D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods,* Academic Press, New York, 1982, pp. 96–156.