Enhanced Sampling of the Molecular Potential Energy Surface Using Mutually Orthogonal Latin Squares: Application to Peptide Structures

K. Vengadesan and N. Gautham
Department of Crystallography and Biophysics, University of Madras, Chennai 600 025, India

ABSTRACT The computational identification of the optimal three-dimensional fold of even a small peptide chain from its sequence, without reference to other known structures, is a complex problem. There have been several attempts at solving this by sampling the potential energy surface of the molecule in a systematic manner. Here we present a new method to carry out the sampling, and to identify low energy conformers of the molecule. The method uses mutually orthogonal Latin squares to select (of the order of) n^2 points from the multidimensional conformation space of size m^n, where n is the number of dimensions (i.e., the number of conformational variables), and m specifies the fineness of the search grid. The sampling is accomplished by first calculating the value of the potential energy function at each one of the selected points. This is followed by analysis of these values of the potential energy to obtain the optimal value for each of the n-variables separately. We show that the set of the n-optimal values obtained in this manner specifies a low energy conformation of the molecule. Repeated application of the method identifies other low energy structures. The computational complexity of this algorithm scales as the fourth power of the size of the molecule. We applied this method to several small peptides, such as the neuropeptide enkephalin, and could identify a set of low energy conformations for each. Many of the structures identified by this method have also been previously identified and characterized by experiment and theory. We also compared the best structures obtained for the tripeptide $(Ala)_3$ by the present method, with those obtained by an exhaustive grid search, and showed that the algorithm is successful in identifying all the low energy conformers of this molecule.

INTRODUCTION

The ab initio prediction of peptide (and protein) structure has received a great deal of attention (Pillardy et al., 2001; Morales et al., 2000; Floudas et al., 1999; Scheraga et al., 1999; Klein et al., 1998; Howard and Kollman, 1988). Often, the prediction is carried out in torsion angle space, thereby reducing the number of degrees of freedom by a third. In torsion angle space, the three-dimensional structure of a peptide chain is specified by the n-torsion angles θ_r, $r = 1, n$, and the optimal structure of the peptide is defined by that set of θ_r that yields the minimum of $V(\theta_r)$ over the entire space, where V is a suitable potential energy function (Halgren, 1995). For the commonly used forms of $V(\theta_r)$, the potential energy surface (PES) has a large number of local minima and singularities. This makes it difficult to effectively use continuous minimization techniques to obtain optimal structures (Piela et al., 1989), if we have no knowledge at all about a structure with which we can start. On the other hand, the computational complexity involved in an exhaustive combinatorial search for the best initial structure from which to start the conventional minimization, by the calculation of $V(\theta_r)$ at every point in the multidimensional torsional space, is unacceptable even for small oligopeptides. This is clear when we consider that for a step size of 360/m along each of the n-torsion angles, the size of the search space is m^n. The complexity thus increases exponentially with the size of the molecule, and even a coarse sampling of the entire space for a small oligomer, say at steps of 30° for a peptide with 10 variable torsion angles, is an ''enterprise of great pitch and moment.'' Many strategies have been designed that attempt to overcome this problem, such as constrained Monte Carlo sampling of the space with subsequent continuous optimization (Li and Scheraga, 1987) and the spot algorithm (Crippen and Scheraga, 1971). In this article we present a new algorithm, which uses mutually orthogonal Latin squares (MOLS) to perform enhanced sampling of the conformational space. The sample so obtained is then analyzed by a procedure similar to the mean field technique (Koehl and Delarue, 1996) to obtain a low energy structure of the molecule. The algorithm, as applied to the peptides studied so far, appears equivalent to an unconstrained global search of the entire PES. It calculates the potential function V at N^2 points in the conformational space (N is of the order of n or m, whichever is greater), which are chosen using MOLS. Then, by means of the procedure described below, it obtains the conformation corresponding to a minimum of the function by an analysis of just these N^2 conformational energies. The cycle is repeated by choosing another set of MOLS, to either identify another minimum, if there are several minima, all of approximately the same value, or to confirm the one already found, if the PES contains only one clearly identifiable minimum. Since the number of atom-atom potentials to be calculated in evaluating the PES does scale quadratically with the size of the molecule, the computational

Submitted August 20, 2002, and accepted for publication January 17, 2003.

Address reprint requests to Namasivayam Gautham, University of Madras, Guindy Campus, Chennai 600 025, India. Tel.: 91-44-2-235-1367, 91-44-2-230-0122; Fax: 91-44-230-2494; E-mail: gautham@unom.ac.in, crystal@giasmd01.vsnl.net.in.

Abbreviations used: MOLS, mutually orthogonal Latin squares; LS, Latin square; PES, potential energy surface; MFT, mean field theory; Aib, α-amino isobutyric acid.

0006-3495/03/05/2897/10 $2.00

complexity of the algorithm scales as the fourth power of N. Nevertheless, this makes it possible to address the structures of peptides of realistic length, using moderate computational resources. For instance, in the case of trialanine, the search for low energy conformers took only 19 min, as compared to 786 min for an exhaustive grid search. MOLS are used in the design of, for example, agricultural or pharmaceutical experiments (Finney, 1955), to project a multidimensional parameter space onto two dimensions. To our knowledge, ours is their first use directly as a combinatorial sampling and search technique to obtain the minima of the biomolecular potential function. We had earlier (Gautham and Rafi, 1992) applied a preliminary version of this method to several simple nonlinear test functions, each designed with a single known optimum in the search space. In each case the method picked up this optimum. An extension to three diverse small biomolecules also gave encouraging results. In this article we present a refined formulation of the method. We first give a general description of the algorithm, then demonstrate its utility in ab initio computation of low energy conformers of several peptides, and finally discuss possible reasons for its efficacy and speculate on whether it may be applied to larger molecules. In the Appendix we show how the method is applied to a specific example.

THE MOLS ALGORITHM

Each cycle of the algorithm consists of four steps (see Fig. 1). The first step is the construction of a set of MOLS. A Latin square of order N is defined as a set of N symbols, arranged in a $N \times N$ square, such that each symbol occurs exactly once in every row and once in every column. Two Latin squares are orthogonal if, when they are superimposed, each symbol of the first square occurs once, and only once, with each symbol of the second square. A set of MOLS is a set of Latin squares, every pair of which is orthogonal (Ito, 1987; see also Fig. 2, this article). It has been shown that if N is a prime power, one can construct $N - 1$ MOLS of order N (Ryser, 1963; Liu, 1968a).

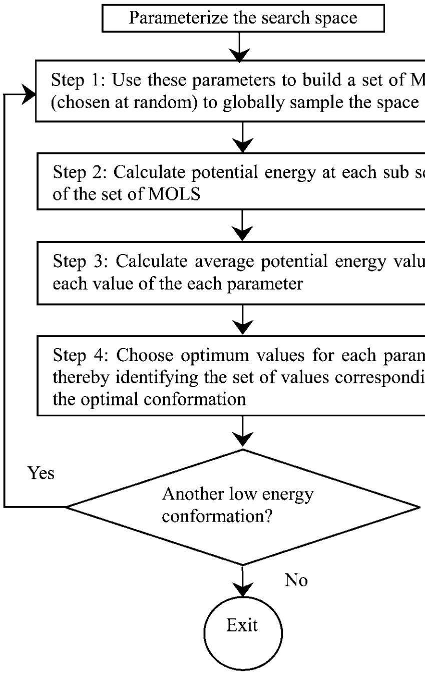
Parameterize the search space
Step 1: Use these parameters to build a set of MOLS (chosen at random) to globally sample the space
Step 2: Calculate potential energy at each sub square of the set of MOLS
Step 3: Calculate average potential energy values of each value of the each parameter
Step 4: Choose optimum values for each parameter, thereby identifying the set of values corresponding to the optimal conformation
Another low energy conformation?
Yes
No
Exit

FIGURE 1 Flow chart for the MOLS procedure.

In the present algorithm we make a correspondence between the symbols used to construct the Latin squares, and the values of conformational variables θ_r, $r = 1, n$, which characterize the peptide conformation. We work in torsion angle space, and specify that each variable torsion angle in the molecule is capable of taking up m different values in a range. Although this range could be restricted by various factors, here it is taken to be 0–360°, the step size thus being 360/m. If r is the index that labels the torsion angles and s is the index that labels the values taken up by each torsion, i.e., the steps along each angle, then $\theta_{r,s}$, $r = 1,n$; $s = 1,m$ are the set of values that define the complete conformational space of the molecule. The sampling has to be carried out among the m^n combinations of these values, each such combination

a1 b1 c1	a2 b2 c2	a3 b3 c3	a4 b4 c4	a5 b5 c5	a6 b6 c6	a7 b7 c7
a2 b7 c6	a3 b1 c7	a4 b2 c1	a5 b3 c2	a6 b4 c3	a7 b5 c4	a1 b6 c5
a3 b6 c4	a4 b7 c5	a5 b1 c6	a6 b2 c7	a7 b3 c1	a1 b4 c2	a2 b5 c3
a4 b5 c2	a5 b6 c3	a6 b7 c4	a7 b1 c5	a1 b2 c6	a2 b3 c7	a3 b4 c1
a5 b4 c7	a6 b5 c1	a7 b6 c2	a1 b7 c3	a2 b1 c4	a3 b2 c5	a4 b3 c6
a6 b3 c5	a7 b4 c6	a1 b5 c7	a2 b6 c1	a3 b7 c2	a4 b1 c3	a5 b2 c4
a7 b2 c3	a1 b3 c4	a2 b4 c5	a3 b5 c6	a4 b6 c7	a5 b7 c1	a6 b1 c2

FIGURE 2 An example of a set of mutually orthogonal Latin squares, showing three MOLS of order 7, i.e., $N = 7$, $n = 3$, and $m = 7$. Symbols in the first Latin square: a1, a2, a3, a4, a5, a6, and a7. Each of these is repeated seven times to give a total of 49 symbols, which have been arranged in a Latin square. Symbols in second Latin square: b1, b2, b3, b4, b5, b6, and b7. The second Latin square is orthogonal to the first. Note that every pairing of a symbol from the first square with one from the second occurs exactly once. Symbols in third Latin square: c1, c2, c3, c4, c5, c6, and c7. This is orthogonal to both the other squares. For clarity in this figure we have used three different sets of N symbols. One could use the same set of N symbols and construct $N - 1$ MOLS of order N. One of subsquares of the set of MOLS has been highlighted; its symbols are a7 of the first Latin square, b1 of the second, and c5 of the third. In the present application, each symbol within the subsquare represents a possible value for the corresponding torsion angle, and each subsquare represents a possible conformation of the molecule. The MOLS method requires the potential function to be evaluated at each of these N^2 points in the conformation space.

specifying one conformation of the molecule. We now use MOLS to pick up N^2 combinations (i.e., N^2 conformations) of these, where N is a prime number greater than the larger of n or m. Once we identify the order N of the Latin squares, it is convenient to set $m = N$, so that now the step size along each of the torsion angles is 360/N, and there are N such steps. To identify the N^2 conformations at which the potential energy calculations are to be carried out, we set each torsion angle to correspond to one Latin square. In other words we arrange the N possible values of each torsion angle in the form of a Latin square. We will have n such Latin squares. We need to further ensure that these form a set of MOLS. The construction of the Latin squares, in a way that ensures they form a set of MOLS, is as follows. We recognize that the set of n MOLS of order N will consist of N^2 subsquares, each containing a set of n torsion angle values corresponding to one conformation of the molecule (see Fig. 2). We will label the subsquares by the indices $u = 1,N$ and $t = 1,N$, and use the symbol $\phi_{\mathrm{r,u,t}}$ to specify the value of the r^{th} torsion angle as found in the subsquare given by the index pair (u,t). To each $\phi_{\mathrm{r,u,t}}$ we assign a value chosen from the set $\theta_{\mathrm{r,s}}$ by putting

$$\phi_{\mathrm{r,u,t}} = \theta_{\mathrm{r,s}},$$

for $r = 1,n$; $s = 1,N$ and $t = 1,N$. The index u is given by

$$u = [(t-1)(r-1) + (s-1)]\mathrm{modulo}(N). \qquad (1)$$

As stated earlier, the value of N is chosen such that 1) $N >$ maximum (m,n), and 2) N is a prime power. The application of this procedure for all values of r, s, and t will result in a set of n MOLS of order N, defined by $\phi_{\mathrm{r,u,t}}$. Equation 1 and the procedure above are adapted from Ryser (1963), and Liu (1968a). Each set $\{\phi_{\mathrm{r,u,t}},\ r = 1,\ n\}$ represents one possible conformation of the polypeptide chain and, therefore, a point in the n-dimensional conformational space. As a working hypothesis, we state that sampling the potential surface of the molecule at the N^2 points specified by these subsquares, out of the possible m^{n}, will enable us to build a map of the entire space, which can then be used to perform a rapid search for the optimum. The hypothesis finds support (but not proof) in fact that, by definition, the set of MOLS implicitly contains every possible two-dimensional projection of the n-dimensional space, i.e., every possible *pairwise* sampling of the torsion angles is present. As shown in Fig. 3, it is possible, in some cases, to obtain a map of a complex three-dimensional object by considering only three two-dimensional projections. The hypothesis above is tantamount to assuming that a similar procedure may be applied to the n-dimensional PES of the molecule. In addition, the use of MOLS allows us to calculate all the $n(n-1)/2$ two-dimensional projections at the same time.

Proceeding on the assumption the hypothesis is true, the second step is to sample the energy hypersurface at these N^2 locations in torsion angle space. This is achieved by calculating the potential energy

$$V_{\mathrm{u,t}} = V(\phi_{\mathrm{r,u,t}}),$$

at each of the N^2 points $u = 1,N$; $t = 1,N$.

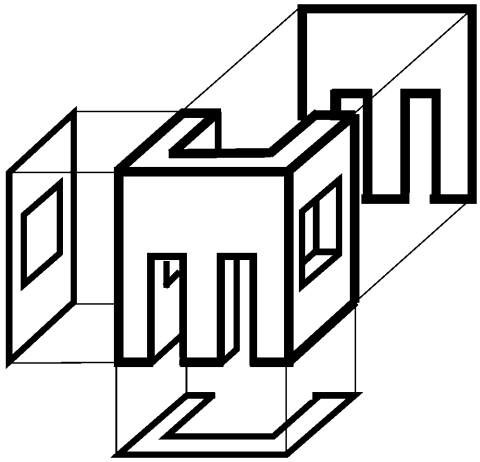

FIGURE 3 A complex three-dimensional object may be reconstructed from three two-dimensional projections.

The third step is to recover the energy map of the conformational space. To accomplish this we construct n one-dimensional representations of the variation of the potential V along each of the torsion angles. The effect of setting a particular torsion to a specific value, regardless of the values of the other torsion angles, is estimated by taking the average of V over those N points in the MOLS where that torsion has been set to that value, i.e.,

$$\langle V\rangle_{\mathrm{r,s}} = \frac{\sum_{\mathrm{t}} V(\varphi_{\mathrm{r,u,t}})\exp(-V(\varphi_{\mathrm{r,u,t}})/kT)}{\sum_{\mathrm{t}} \exp(-V(\varphi_{\mathrm{r,u,t}})/kT)}, \qquad (2)$$

where a Boltzmann weighting function is used to smoothen the potential surface. The temperature, T, has been arbitrarily chosen to be 3.0 K in the present report. Once again, Eq. 1 is used to calculate the value of u for each value of r, s, and t. Since we have set the number of steps m along each angle as equal to N, there will be N such average values for each of the $r = 1,n$ torsion angles. It then follows as a corollary to the hypothesis made above, that the N average values for a given torsion will form a representation of the behavior of the potential V as a function of that torsion. This conjecture has parallels to mean field methods (Koehl and Delarue, 1996). There are, however, conceptual and procedural differences, and these are discussed later in this article.

The fourth and final step is an inspection of each one-dimensional representation, which will therefore reveal the optimum value for the respective torsion. Thus, if

$$V_{\mathrm{r,s}=\omega_{\mathrm{r}}} = \underset{\mathrm{over\ s}}{\mathrm{minimum}}(\langle V\rangle_{\mathrm{r,s}}),$$

then $\theta_{\mathrm{r},\omega_{\mathrm{r}}}$ is the optimum value for the torsion r. The set of optima $\theta_{\mathrm{r},\omega_{\mathrm{r}}}$, $r = 1,n$ will then define a low energy conformation of the peptide.

The indices r and s used in the construction of the MOLS may be assigned to the m values of the n angles in $(m!)^n$ different ways (Liu, 1968b). By our hypothesis, no matter which way the assignment is made, the procedure will result in a low energy structure. Therefore, by choosing different assignments of the angles, the calculations may be repeated to check for the consistency of the results obtained, and to identify other equally energetically favorable structures. The procedure terminates when no new structures are picked up, thus indicating the end of a global search for low energy structures.

We note that, although the above description of the method is in terms of biomolecular structure, there are no assumptions regarding the form of the function V. Therefore, at least under the terms described above, the method may be used to find the optimum of a wide variety of functions.

LOW ENERGY PEPTIDE STRUCTURES

We have applied the method to several small peptides of varying lengths. The potential function used in all these calculations was

$$V=\sum_{i<j}\frac{332q_iq_j}{\varepsilon r_{ij}}-\sum_{i<j}\frac{A_{ij}}{r_{ij}^{6}}+\sum_{i<j}\frac{B_{ij}}{r_{ij}^{12}}+\sum_{i<j}\frac{C_{ij}}{r_{ij}^{12}}+\sum_{i<j}\frac{D_{ij}}{r_{ij}^{10}}, \quad (3)$$

with parameters taken from the Weiner and Kollman forcefield (Weiner and Kollman, 1986). This semiempirical function models the energy as the sum of interatomic electrostatic, van der Waals and hydrogen-bond energies (r_{ij} are the interatomic distances). The method was applied to the neuropeptide [Met[5]]enkephalin (NH_2-Tyr-Gly-Gly-Phe-Met-OH; see Isogai et al., 1977), which has been frequently used as a model system in theoretical studies of biomolecules (Floudas et al., 1999). One thousand five hundred optimal structures were generated by the application of the above procedure to as many different input sets of torsion angles $\theta_{r,s}$. The assignments of the indices r and s to the m-values of the n-angles were made using a random number generator. Fig. 4, *left* and *right*, show two of these optimal structures. One of them has been earlier characterized as the minimum-

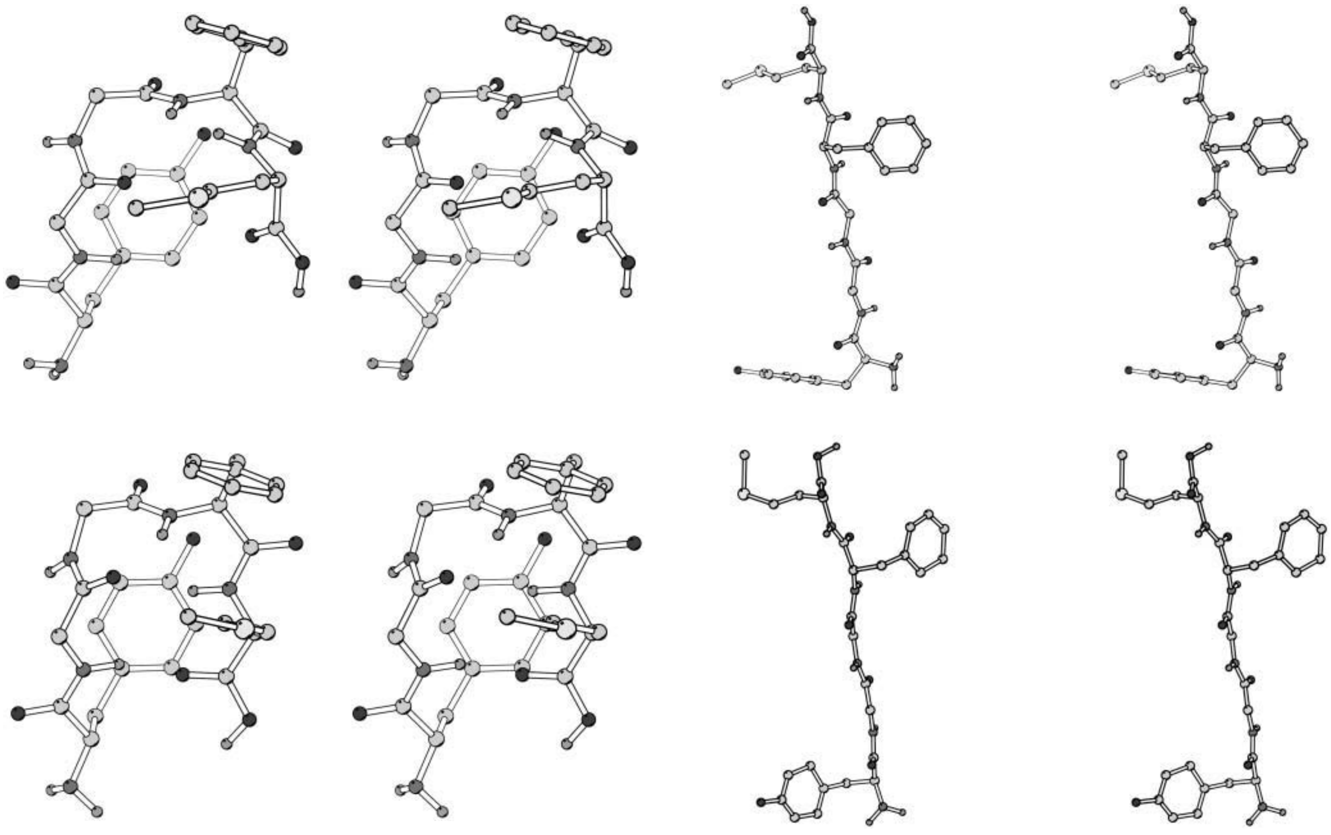

FIGURE 4 (*Left*) Stereo view of one of the low energy structures of [Met[5]]enkephalin obtained by MOLS (*top*) compared with the structure reported by Li and Scheraga (1987) (*bottom*). The energy for the MOLS structure as calculated by Eq. 3 is −16.0 kcal/mol. This pentapeptide has eight backbone torsion angles, and the step size chosen was 10°. The side-chain torsion angles were retained constant at the values given in the energy-minimized structure of Li and Scheraga (1987). Thus $n = 8$ and $m = 36$. The order of MOLS was therefore chosen to be $N = 37$ (the next higher prime number). The assignment of the angle values to the indices to generate a new input set $\theta_{r,s}$, and therefore a new set of MOLS, was carried out afresh for each of the 1500 calculations using a random number generator. The structure shown here is one of these 1500. (*Right*) Another low energy structure identified by MOLS (*top*) for the same pentapeptide (i.e., another of the 1500 structures), compared with its crystal structure (Griffin et al., 1986) (*bottom*). The MOLS energy for this structure is −5.8 kcal/mol.

energy structure by Monte Carlo minimization (Li and Scheraga, 1987). The other is the same as the crystal structure of the peptide (Griffin et al., 1986). We used the Ramachandran map to characterize the rest of the 1500 structures (Fig. 5). All of them fall in the low energy regions of the map. A further indication that these structures are energetically favorable comes from the potential energies corresponding to these structures, as calculated using Eq. 3 on each of the 1500 structures, after it was identified by the MOLS procedure. They are all low, the highest being 2725.4 kcal/mol and the lowest -16.1 kcal/mol, with an average of -5.8 ($SD = 70.9$) kcal/mol and a median value of -8.6 kcal/mol for all 1500. Next we used the SCAR clustering program (Betancourt and Skolnick, 2001), and established that the 1500 conformations were not unrelated but, after subjecting each to a few cycles of gradient minimization (Biosym, MSI Release 95.0, San Diego, CA), could be classified into just 23 closely related structures (Fig. 6). The structures corresponding to these 23 clusters were picked up within the first 300 of the 1500 generated (Fig. 7), indicating that further generation of conformations using MOLS would produce no new low-energy conformation, and that the conformational space had been exhaustively searched. We may state therefore that the conformational space of Met-Enkephalin (as defined in the caption to Fig. 4) consists of just these 23 minima.

Similar results (Table 1) were obtained when the method was applied to [Leu^5]enkephalin (NH$_2$-Tyr-Gly-Gly-Phe-Leu-OH; Isogai et al., 1977), (Aib)$_5$, and the decapeptides NH$_2$-([Leu]$_4$-Aib)$_2$-OH (D1) and NH$_2$-Trp-Ile-Ala-Aib-Ile-Val-Aib-Leu-Aib-Pro-OH (D2). Its application to several other model oligopeptides, including (Ala)$_5$, (Ala)$_{10}$, (Ala)$_{15}$, (Gly)$_5$, (Gly)$_{10}$, and (Gly)$_{15}$ also gave similar results.

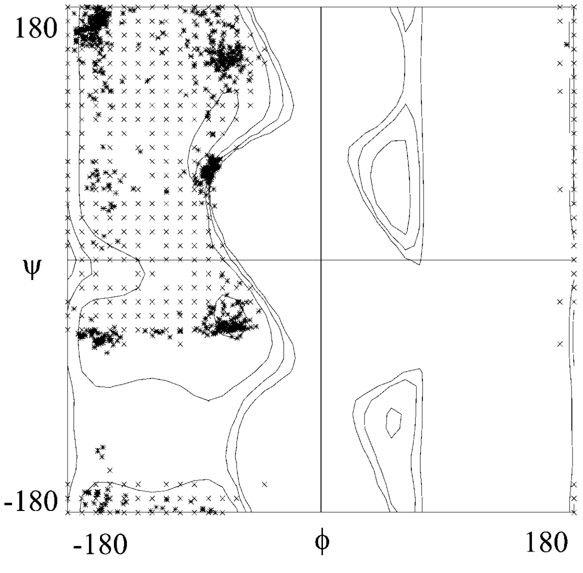

FIGURE 5 Ramachandran map for the nonglycyl residues of Met-enkephalin plotted for the 1500 optimal conformations obtained by repeated application of the four steps of the MOLS method. The contours in the map were calculated for (Ala)$_2$ using Eq. 3 and were drawn at intervals of 2 kcal/mol starting from -3.0 kcal/mol. The crosses (×) represent discrete conformations picked up by the MOLS method. The stars (*) are the same structures after a few cycles of gradient minimization.

To verify whether the search was truly exhaustive, and that all low energy conformers of a molecule could be picked up, we considered the tripeptide (Ala)$_3$. There are six variable torsion angles in this molecule and if we assume the step size to be 19°, we have $n = 6$ and $m = 19$, and the total search space consists of 19^6 (i.e., 47,045,881) points. The potential energy was calculated at every one of these points. One thousand of the lowest energy conformations were saved. Clustering of these, after gradient minimization, yielded seven different structures with energy values ranging from -7.41 to -6.07 kcal/mol. The entire procedure took 786 min of CPU time. The clustering algorithm used was the one suggested by Kříž et al. (2001), rather than the SCAR algorithm used in the other cases above. The latter considers only C^α atoms for clustering. Since there are only three of these in the tripeptide, it yields only a single cluster.

The same conformational space was searched using the MOLS procedure. The order of MOLS was taken to be $N = 19$, with $n = 6$ and $m = 19$. A total of 1000 low energy structures were generated, by running the MOLS procedure 1000 times. When these were clustered together after gradient minimization, they gave rise to 56 unique structures, with energies ranging from -7.38 to -4.53 kcal/mol. The entire calculation consumed only 19 min of CPU time. These structures were compared with those obtained by the complete grid search. As seen in Table 2, the seven low energy conformers found in the complete grid search have also been identified as numbers 1, 2, 3, 5, 6, 8, and 30 by the MOLS method, when ranked according to the average energy of the structures in each cluster. The other 49 low energy conformers picked up by MOLS would presumably have been seen in the complete grid search too, if we had subjected more of the 19^6 structures to cluster analysis, instead of just the best 1000.

DISCUSSION

The MOLS method outlined above is thus a way of sampling all of torsion angle space to identify a library of low-energy three-dimensional structures for any given peptide sequence. The search is unconstrained and is accomplished at very little computational cost. It terminates in polynomial time and, as shown above, may be applied to problems normally classified as NP-hard (Kirkpatrick et al., 1983). The method lends itself easily to parallelization, and this would further speed up the computation.

It is not clear why the method works as well as it does. As mentioned earlier, MOLS are used, for instance, in designing agricultural experiments. To study crop yields, experimental agricultural plots are laid out using Latin squares, the variables being, for example, seed quality, pesticide treatment, and so on. In other words, MOLS are used to sys-

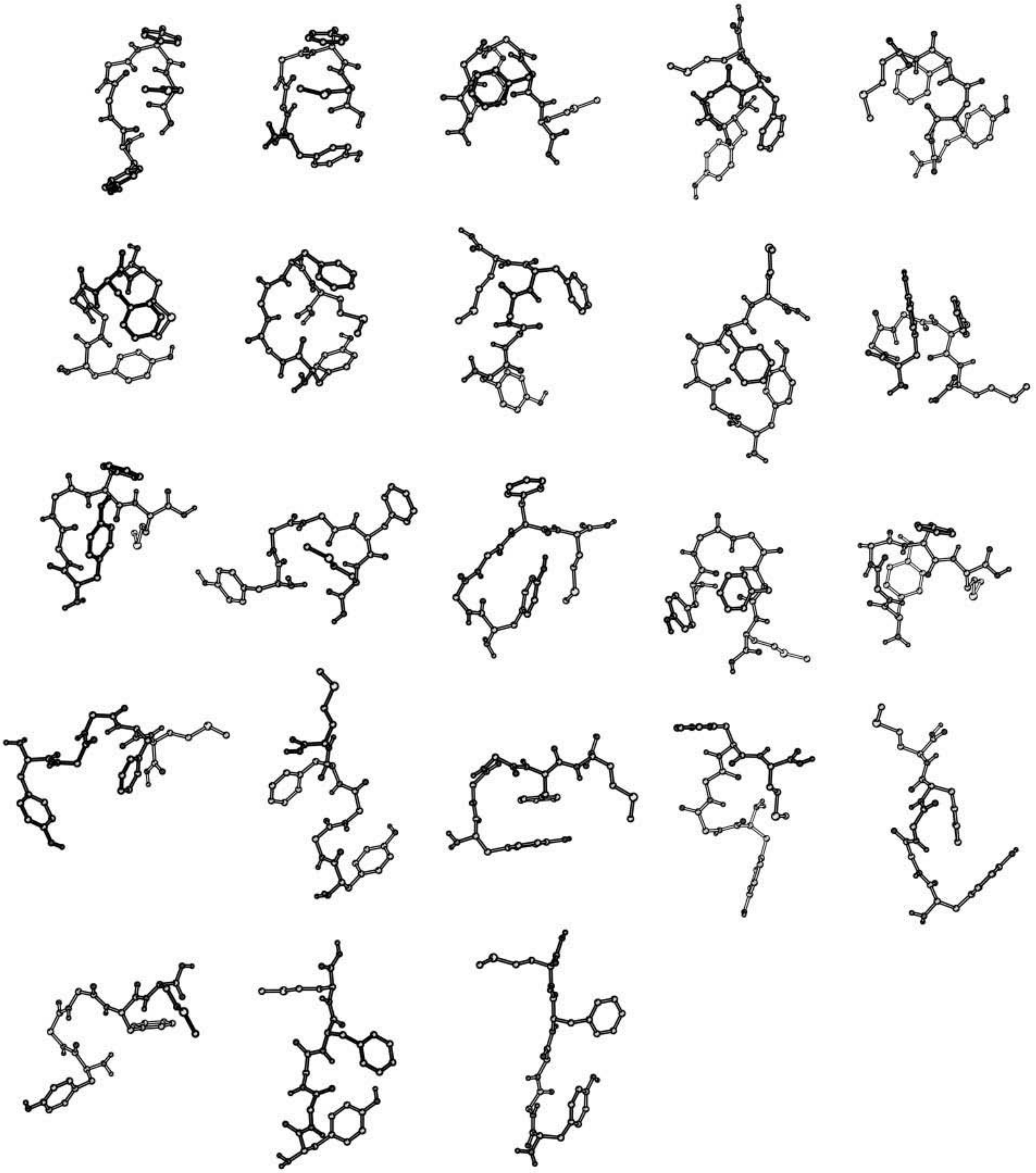

FIGURE 6 The 23 low-energy structures that represent the entire conformational space of [Met^5]enkephalin. Shown here are the centroids of the 23 clusters obtained by applying the SCAR clustering algorithm (Betancourt and Skolnick, 2001) to the 1500 structures, which were first generated by the MOLS method and then subjected to gradient minimization.

tematically sample the space of the variables. The yields of the plots are then analyzed using statistical techniques such as ANOVA, to finally arrive at a strategy to maximize the yield (Finney, 1955). In the present work, we have considered the search for a minimum energy molecular structure as a problem that can be similarly solved by sampling. Each calculation of the conformational energy is considered an 'experiment' that samples the conformational space. We then use MOLS to design the sampling experiments. This is the first innovation. The second one is that, unlike the agricultural experiments, we analyze the sample by taking averages to directly arrive at the best structure. We have as yet no theoretical support for these innovations, except, perhaps, the following.

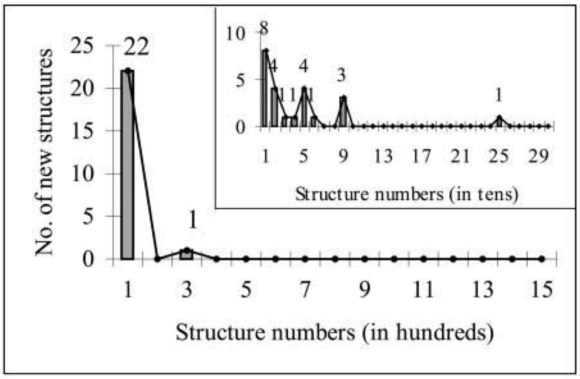

FIGURE 7 The number of mutually dissimilar low energy structures identified versus the number of structures generated using MOLS. After the first 300 structures, no new structures are discovered, showing that all possible low energy conformations of the molecule have already been identified. (*Inset*) The same figure, with an expanded *x*-axis, showing that, in fact, almost all the low energy conformations were identified within the first 90 MOLS structures.

The approximations and assumptions made in the MOLS method find resonance in mean field theory (MFT), especially in the application of the latter as an optimization method for peptide and protein structures (Olszewski et al., 1992). It is therefore instructive to compare the two. In both methods the optimum value of each variable describing the system is found independently of the other variables, in a mean potential or force field constructed using some values for all the others. The two methods however differ conceptually after this point. The mean field in MFT uses a set of values that are either randomly assigned or derived from a known template or ensemble of structures (Koehl and Delarue, 1996). In the MOLS procedure the mean field is built, in Eq. 2 using all possible pairwise combinations of the value of the variable being considered, with all other values of all other variables. Procedurally, too, MFT is commonly an iterative method that is repeated until self-consistency is reached (Olszewski et al., 1992). In contrast, a single application of the MOLS method leads to an energetically favorable structure. The procedure is repeated only when another such structure is sought. Despite these differences, the similarities are strong enough to motivate their further exploration, and this is being undertaken.

The possibility of extending the method to ab initio protein structure prediction is obvious, but the success of any such scheme would depend on several factors, including the development of an appropriate potential function. Such a function would have to possess a deep and fairly wide minimum in conformational space corresponding to the native structure. This is because the method, as discussed above, repeatedly samples each value of each parameter, to determine its effect on the potential, regardless of the other values and the other parameters. If the minimum were not

TABLE 1 Results of the application of the method to oligopeptides

Molecule	(a)	(b)	(c)	(d)	(e)	(f)
Met-enkephalin	1500	23	300	0.2%	105	FABJIB (β-turn and β-strand)
Leu-enkephalin	1000	15	100	0.13%	70	BIXNIF10, FABJEX, GEWWAG, and LENKPH11 (β-turn and β-strand)
(Aib)$_5$	1500	20	300	0.2%	85	BIBDUL and TAIBYM (3_{10}-helix)
D1	1500	60	900	0.0%	590	JUCHUK (α-helix)
D2	1500	96	800	0.0%	490	VINFON, VINFUT, JAXGUK, and DUTLEJ10 (α-helix)

Column (*a*) The total number of structures generated using MOLS. (*b*) The number of unique structures identified. (*c*) The number of generated structures within which all the unique structures were identified. (*d*) The percentage of the (φ,ψ) points among the generated structures that fall outside the allowed regions of the Ramachandran plot, after gradient minimization. (*e*) The time (in min) taken for generating all the structures on a computer based on a single 650-Mhz Pentium III processor. (*f*) The CSD IDs (CSD, V5.16, Cambridge Crystallographic Data Centre, University Chemical Laboratory, Cambridge, UK) of the x-ray structures that are the same as (or similar to) some of the structures picked up by the present method. Also given is a brief description of the structure.

TABLE 2 Comparison of complete grid search with MOLS search for (Ala)$_3$

Complete grid search cluster no.	Cluster size	Average energy (kcal/mol)	MOLS search cluster no.	Cluster size	Average energy (kcal/mol)	RMSD (Å)
1	82	−7.41	1	48	−7.38	0.01
2	21	−7.25	2	73	−7.17	0.01
3	79	−7.06	3	40	−7.03	0.01
4	287	−6.96	5	64	−6.96	0.04
5	62	−6.85	6	71	−6.84	0.05
6	468	−6.79	8	78	−6.78	0.02
7	1	−6.07	30	1	−6.02	0.91

The structures are ranked according to the average energy of all the members in the respective cluster. The cluster size specifies the number of members present. The last column specifies the RMS deviation in atomic positions upon superposition of the lowest energy structure in the cluster obtained by complete grid search upon the lowest energy structure in the cluster obtained by MOLS.

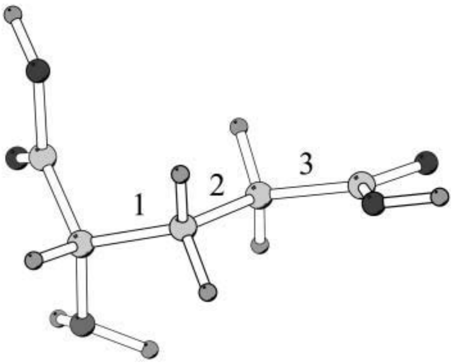

FIGURE A1 Glutamic acid. The variable side-chain torsion angles are numbered.

sufficiently deep, the effect may not be noticeable. It is not clear if the folding funnel (Bicout and Szabo, 2000) that is thought to be representative of the potential energy hyper surface of a protein can be sufficiently well-represented by a simple semiempirical function, such as Eq. 3, to allow a straightforward extension of the MOLS method to larger polypeptide chains. Our experience with the small peptides described above suggests that this may not be the case, since the potential function does not show any single deep minimum.

Apart from this problem, there is a more fundamental objection to the extension of the method, as it stands, to proteins. The method samples a small fraction of the conformational space and derives information about the whole. For a pentapeptide, this fraction is $\sim 10^{-10}$. For proteins the fraction would be virtually zero. At this point in the development of the method, however, it is not clear if this is an insuperable objection. One could, for example, think of applying the method iteratively to larger and larger sections of the chain, each time ensuring that the fraction sampled remains reasonable.

APPENDIX

We illustrate the application of the MOLS procedure to the side chain of Glutamic acid, consisting of just three variable torsion angles (Fig. A1). It is assumed that each angle can take up any one of the six values 0, 60, 120, 180, 240, and 300°, i.e., the range is 0–360° and the step size is 60°. Thus $n = 3$, $m = 6$, and the size of the search space is $6^3 = 216$. We have to choose the order, N, of the Latin squares as a prime number greater than the larger of n and m. Thus $N = 7$, and we need to pick $N^2 = 49$ points out of 216, at which to sample the PES. For convenience we now set $m = N$, i.e., we divide each dimension into seven steps, not six. The values of $\theta_{r,s}$ that now define the search space are given in Fig. A2. We now use the pseudo code routine given in Fig. A3, which is an implementation of Eq. 1, to identify the set of three MOLS, given by $\phi_{r,u,t}$. The result is shown in Fig. A4. Each of the 49 subsquares in this set of MOLS corresponds to a conformation of the molecule. We now calculate the potential energy for each conformation. For example, to calculate the first energy value we build the molecule with torsion angle 1 = 0°, angle 2 = 0°, and angle 3 = 0°, and use Eq. 3. The 49 values of the potential energy so calculated are given in Fig. A5. We now use these values to find the effect of setting each torsion to each of the possible values. Thus to find the average effect of setting torsion angle 1 to a value of 0° we take a Boltzmann weighted average of the energy values in the subsquares (1,1), (2,2), (3,3), (4,4), (5,5), (6,6) and (7,7). Note that these are the seven subsquares in which angle 1 is set to the value of 0°. Also note that in these seven subsquares, the value of 0° for angle 1 occurs once along with each of the values for angle 2. In the same seven subsquares, the same value 0° for angle 1 also occurs once along with each of the values for angle 3. In other words, every possible pairwise combination of the value 0° along with all other values of all other angles is sampled. Generalizing this procedure, to

s r	1	2	3	4	5	6	7
1	0.0	51.4	102.9	154.3	205.7	257.1	308.6
2	0.0	51.4	102.9	154.3	205.7	257.1	308.6
3	0.0	51.4	102.9	154.3	205.7	257.1	308.6

FIGURE A2 The values of $\theta_{r,s}$ for $r = 1, n$ and $s = 1, m$.

```
loop r = 1, n
  loop s = 1, N
    loop t = 1, N
          u = {(t-1)*(r-1) + (s-1)}modulo(N) + 1
          $\phi$ (r,u,t) = $\theta$ (r,s)
    end loop t,s,r
```

FIGURE A3 The pseudo code routine to generate the set of MOLS $\phi_{r,u,t}$ from $\theta_{r,s}$.

u t	1	2	3	4	5	6	7

1	0.0($\phi_{1,1,1}$)
	0.0($\phi_{2,1,1}$)
	0.0($\phi_{3,1,1}$)
		308.6
		257.1
		205.7
			257.1
			154.3
			51.4
				205.7
				51.4
				257.1
					154.3
					308.6
					102.9
						102.9
						205.7
						308.6
							51.4
							102.9
							154.3

2	51.4 51.4 51.4	0.0 308.6 257.1	308.6 205.7 102.9	257.1 102.9 308.6	205.7 0.0 154.3	154.3 257.1 0.0	102.9 154.3 205.7
3	102.9 102.9 102.9	51.4 0.0 308.6	0.0 257.1 154.3	308.6 154.3 0.0	257.1 51.4 205.7	205.7 308.6 51.4	154.3 205.7 257.1
4	154.3 154.3 154.3	102.9 51.4 0.0	51.4 308.6 205.7	0.0 205.7 51.4	308.6 102.9 257.1	257.1 0.0 102.9	205.7 257.1 308.6
5	205.7 205.7 205.7	154.3 102.9 51.4	102.9 0.0 257.1	51.4 257.1 102.9	0.0 154.3 308.6	308.6 51.4 154.3	257.1 308.6 0.0
6	257.1 257.1 257.1	205.7 154.3 102.9	154.3 51.4 308.6	102.9 308.6 154.3	51.4 205.7 0.0	0.0 102.9 205.7	308.6 0.0 51.4
7	308.6 308.6 308.6	257.1 205.7 154.3	205.7 102.9 0.0	154.3 0.0 205.7	102.9 257.1 51.4	51.4 154.3 257.1	0.0 51.4 102.9

FIGURE A4 The values of $\phi_{r,u,t}$, for $r = 1, n$; $u = 1, N$; and $t = 1, N$.

u t	1	2	3	4	5	6	7
1	-2.56	7.22	160.72	-1.69	-1.41	$2.1x10^4$	4.67
2	7.12	-2.69	196.87	0.54	-1.95	-1.73	$7.0x10^7$
3	292.91	1.82	0.56	$5.1x10^6$	-0.61	-1.77	-1.61
4	19.68	22.77	10.23	14.25	8.39	-1.44	-0.81
5	0.38	55.18	4.65	145.83	21.65	43.25	1.48
6	0.40	-1.47	-1.70	9.45	$5.5x10^8$	16.65	1.73
7	19.16	1291.94	-2.18	-1.81	12.71	418.73	-2.69

FIGURE A5 The values (in kcal/mol) of potential energy $V_{u,t}$ for $u = 1, N$ and $t = 1, N$.

```
loop r = 1, n
  loop s = 1, N
    <V(r,s)> = 0.0
    loop t = 1, N
        u = {(t-1)*(r-1)+(s-1)}modulo(N) + 1
        <V(r,s)>  = <V(r,s)> + V(u,t)*weight
  end loop t,s,r
```

FIGURE A6 The pseudo code routine to obtain average energy values $\langle V\rangle_{r,s}$ from $V_{u,t}$.

s r	1	2	3	4	5	6	7
1	2.28	5.52	10.41	0.35	-1.50	-0.34	8.80
2	-0.24	1.77	4.12	7.31	2.31	1.54	2.42
3	0.30	5.75	-1.75	5.08	3.33	0.41	2.50

FIGURE A7 The values of $\langle V\rangle_{r,s}$ calculated from $V_{u,t}$.

s r	1	2	3	4	5	6	7
1	205.7	0.0	257.1	102.9	308.6	154.3	51.4
2	102.9	257.1	205.7	51.4	0.0	154.3	308.6
3	308.6	102.9	154.3	0.0	205.7	51.4	257.1

FIGURE A8 Another set of values of $\theta_{r,s}$.

find the effect of setting angle r to value number s, we use Eq. 2, as implemented in the pseudo code given in Fig. A6. At the end of this procedure, we obtain Fig. A7, which shows the average energy for each setting of each angle. Inspection of this table reveals that the best setting (i.e., one that yields the lowest average potential energy) is $\theta_{1,5}$, i.e., 205.7° for angle 1. Similarly, the best setting is $\theta_{2,1}$ or 0.0° for angle 2 and $\theta_{3,3}$ or 102.9° for angle 3. According to our hypothesis, these three values specify a low energy conformation of the molecule. Indeed, if we set the angles to these values, and calculate the energy of the resulting conformation using Eq. 3 again, we obtain -1.90 kcal/mol. We may recast Fig. A2, since we could associate the indices r and s with the values of the torsion angles in $(7!)^3$ or approximately 1.3×10^{11} different ways. Fig. A8 shows a different association, carried out using a random number generator. If we apply Eq. 1 to this figure, we are led to new set of MOLS. When we carry through the calculations with this set we arrive at a new low energy (-2.67 kcal/mol) conformation for the molecule, described by angle 1 = 0°, angle 2 = 0°, and angle 3 = 205.7°. Further repetitions may lead to other low energy conformers, or to one already identified.

We are grateful to Professor N. Yathindra, University of Madras, for comments.

We thank the Council of Scientific and Industrial Research, the University Grants Commission, and the Department of Science and Technology of the Government of India for financial support.

REFERENCES

Betancourt, M. R., and J. Skolnick. 2001. Finding the needle in a haystack: deducing native folds from ambiguous ab initio protein structure predictions. *J. Comp. Chem.* 22:339–353.

Bicout, D. J., and A. Szabo. 2000. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. *Protein Sci.* 9:452–465.

Crippen, G. M., and H. A. Scheraga. 1971. Minimization of polypeptide energy X. A global search algorithm. *Arch. Biochem. Biophys.* 144:453–461.

Finney, D. J. 1955. Experimental Design and its Statistical Basis. Cambridge University Press, London. pp. 56–67.

Floudas, C. A., J. L. Klepeis, and P. M. Pardalos. 1999. Global optimization approaches in protein folding and peptide docking. DIMACS series in discrete mathematics and theoretical computer science. *Am. Math. Soc.* 47:141–171.

Gautham, N., and Z. A. Rafi. 1992. Global search for optimal biomolecular structures using mutually orthogonal Latin squares. *Curr. Sci.* 63:560–564.

Griffin, J. F., D. A. Langs, G. D. Smith, I. J. Blundell, and S. Bedarker.1986. The crystal structures of [Met^5]enkephalin and a third form of [Leu^5]enkephalin: Observations of a novel pleated β-sheet. *Proc. Natl. Acad. Sci. USA.* 83:3272–3276.

Halgren, T. A. 1995. Potential energy functions. *Curr. Opin. Struct. Biol.* 5:205–210.

Howard, A. E., and P. A. Kollman. 1988. An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.* 31:1669–1675.

Isogai, Y., G. Nemethy, and H. A. Scheraga. 1977. Enkephalin: conformational analysis by means of empirical energy calculations. *Proc. Natl. Acad. Sci. USA.* 74:414–418.

Ito, K. 1987. Encyclopedic Dictionary of Mathematics, Vol. II. MIT Press, Cambridge, Massachusetts. 891–892.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science.* 220:671–680.

Klein, C. T., B. Mayer, G. Köhler, and P. Wolschann. 1998. Systematic stepsize variation: efficient method for searching conformational space of polypeptides. *J. Comp. Chem.* 19:1470–1481.

Koehl, P., and M. Delarue. 1996. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* 6:222–226.

Kříž, Z., P. H. J. Carlsen, and J. Koca. 2001. Conformational features of linear and cyclic enkephalins. A computational study. *J. Mol. Struct.* 540:231–250.

Li, Z., and H. A. Scheraga. 1987. Monte Carlo minimization approach to the multiple minima problem in protein folding. *Proc. Natl. Acad. Sci. USA.* 84:6611–6615.

Liu, C. L. 1968a. Introduction to Combinatorial Mathematics. McGraw-Hill Book Company, New York. pp. 359–370.

Liu, C. L. 1968b. Introduction to Combinatorial Mathematics. McGraw-Hill Book Company, New York. pp. 1–22.

Morales, L. B., R. Garduño-Juárez, J. M. Aguilar-Alvarado, and F. J. Riveros-Castro. 2000. A parallel tabu search for conformational energy optimization of oligopeptides. *J. Comp. Chem.* 21:147–156.

Olszewski, K. A., L. Piela, and H. A. Scheraga. 1992. Mean field theory as a tool for intramolecular conformational optimization. 1. Tests on terminally-blocked alanine and Met-enkephalin. *J. Phys. Chem.* 96:4672–4676.

Piela, L., J. Kostrowicki, and H. A. Scheraga. 1989. The multiple-minima problem in the conformational analysis of molecules. Deformation of the potential energy hypersurface by the diffusion equation method. *J. Phys. Chem.* 93:3339–3346.

Pillardy, J., J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Kamierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y. Ye, and H. A. Scheraga. 2001. Recent improvements in prediction of protein structure by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA.* 98:2329–2333.

Ryser, H. J. 1963. Combinatorial Mathematics. Mathematical Association of America. pp. 79–84.

Scheraga, H. A., J. Lee, J. Pillardy, Y. J. Ye, A. Liwo, and D. Ripoll. 1999. Surmounting the multiple-minima problem in protein folding. *J. Global Optim.* 15:235–260.

Weiner, S. J., and P. A. Kollman. 1986. An all atom force field for simulations of proteins and nucleic acids. *J. Comp. Chem.* 7:230–252.