A new conformational search technique and its applications

K. Vengadesan and N. Gautham*

Department of Crystallography and Biophysics, University of Madras, Chennai 600 025, India

The computational identification of the low energy structures of a peptide from its sequence alone has been a problem of major interest for many years. It is not an easy task even for small peptides, due to the multiple-minima problem and combinatorial explosion. A number of conformational search algorithms have been developed in the past for this purpose. We have developed an algorithm that addresses this problem. In statistical experimental design, mutually orthogonal Latin squares (MOLS) are used to systematically sample the space of the variables. This allows the experimenter to conduct the experiment with a relatively small number of runs, instead of examining all possible combinations of values of the variables. We considered whether the problem of searching for minimum energy molecular structure on the potential energy surface could also be similarly solved by MOLS sampling. This has led to the development of a conformational search algorithm. In this article, we briefly review the work carried out in our laboratory using the MOLS search technique over the last decade.

Background

THE computational identification of the optimal three-dimensional fold of a peptide (or protein) from its sequence, particularly if no reference is made to other known structures, is a complex problem and has received a great deal of attention. A number of computational techniques have been developed for this purpose[1–8]. In general, these techniques work on the assumption that the equilibrium structure of the molecule is one that corresponds to the minimum of a suitable potential energy function. Consequently, the techniques have two essential components. The first is a model (representation) of the structure along with an appropriate potential energy function. There are several representations (such as all-atom models, simplified models, etc.) widely used in these energy calculations. The model and energy functions are dependent on each other. In the all-atom models, all the atoms, including hydrogen, are used to represent the structure, and each atom is considered an individual interaction centre in the calculations. Sometimes the hydrogen atoms are removed from the representation, e.g. a methyl group would be modelled as a single 'pseudo-atom' or 'united atom'. In even more simplified models (or virtual atom models), each amino acid residue is represented by one or two interaction centres (e.g. lattice models, bead models, etc.). Several potential energy functions or force fields (AMBER[9], CHARMM[10], ECEPP[11], etc.) have been developed for the all-atom models, and are commonly used. Several other energy functions (for the simplified models), such as the residue–residue effective potential[3], are also in use today.

The second component is a conformational search algorithm for searching the conformational space (or energy hyper surface) of polypeptides and identifying low energy conformations, of which there may be many, or just one with the lowest energy. A number of conformational search algorithms have been developed in the past and periodically reviewed[2–8]. Here we give a brief digest of some of the methods.

In most of these methods, it is assumed that the native (i.e. naturally occurring) conformation corresponds to the point on the potential energy surface with the lowest energy value, i.e. the global energy minimum (GEM). Thus, the main goal of conformational search is to identify this point. This search is generally treated as a global optimization problem in which the potential energy function is the objective function, and the torsion angles or coordinates that are used to represent the conformation of the polypeptide chain are the variables. The task is then to vary the values of these variables and find the point where the objective function has the global minimum value. However, this is not an easy task because the multidimensional energy surface of even small peptides consists of an astronomically large number of points, with several minima, separated by a multiple-scale distribution of energy barriers. It resembles a rugged geographical landscape with many scattered hills (energy barriers) and valleys (minima) having various heights and depths. The computational search for the GEM conformation therefore could get trapped in a local minimum. This is called the 'multiple minima problem'.

In structure prediction and conformational analysis, particularly of small flexible peptides (e.g. enkephalins), it may be important to find all the relatively small number of low energy (local) minima, in addition to the global minimum. This is because the multiple minima may possibly correspond to different conformational sub-states necessary for biological activity. Also, the GEM conformation may not be the (experimentally determined) native structure,

*For correspondence. (e-mail: gautham@unom.ac.in)

because empirical energy functions commonly used are rather approximate, such as, for example, in their treatment of the solvent interactions. Moreover, even with accurate potential functions, the GEM may not be the most highly populated minimum[2], due to the shape of the energy surface – a deep and narrow minimum may be less populated than a broad, but slightly higher, minimum. Under such circumstances, it is necessary to analyse as many different low energy conformations as possible. This is why conformational search methods are usually interested in finding not only the GEM on the PES, but also all minima whose energies are of the same order of magnitude as the GEM.

To locate the GEM or to locate more than one minimum, it is usual practice to generate a large number of starting conformations equally distributed on the energy surface, minimize them using local optimization techniques[12] to the nearest local minima, and then throw away the duplicates. However, even for small peptides of, say, five residues, having about a hundred atoms, finding the GEM or all low energy conformations in a multi-dimensional PES is a computationally demanding problem. An exhaustive search is impractical because the volume of the search space increases exponentially with the number of degrees of freedom (usually the dihedral angles), i.e. with the size of the molecule. From a computational point of view, such problems are related to non-deterministic polynomial time (NP) hard problems[13], in that the total number of possible conformations is an exponential function of the total number of degrees of freedom. It is widely assumed that it requires an exponential amount of time to solve such problems. This phenomenon is commonly known as 'combinatorial explosion'[2]. Therefore, we need some specialized conformational search (optimization) algorithms that can explore the complete conformational space and obtain all the low energy conformations, at tractable computational cost. Several such conformational search algorithms have been developed in the past for peptides and proteins[2,7,8]. These algorithmic strategies are commonly known as conformational search techniques, and they are sometimes used to generate initial starting structures for subsequent conventional minimization, using, say, the conjugate gradient algorithm. Some of these are briefly described below.

Conformational search techniques

Conformational search techniques can be classified broadly into two categories; stochastic and deterministic methods. Stochastic methods (Monte Carlo method, simulated annealing, genetic algorithm, etc.) rely on probabilistic descriptions to aid in locating the global minimum, and there is no natural endpoint to the procedure. Whereas deterministic methods (systematic search method, etc.) provide a certain level of assurance in locating the global minimum and there is a defined endpoint to the procedure. Some of the commonly used conformational search methods are listed here with a brief description.

Systematic (or grid or exhaustive) search methods

In these methods, each dihedral angle involving rotation about a single bond of a molecule is systematically incremented by a fixed amount (e.g. 30°) until all possible combinations of dihedral angles for the chosen increment have been generated. Each combination is subjected to a local energy minimization. The search becomes impractically large for smaller increments, and the method is limited to very small molecules or polypeptide segments[14].

Build-up procedures

Build up procedures are based on the assumption that each fragment is conformationally independent of the other fragments in the molecule. The polypeptide chain is divided into fragments. Ensembles of structures are determined for each of these small fragments, and joined to form the overall conformation[15].

Monte Carlo methods (the Metropolis algorithm)

Monte Carlo methods are stochastic techniques, in which the energy E_0 is calculated first for an arbitrary conformation. By randomly changing the dihedral angles (coordinates in Cartesian space), a new conformation is generated and its energy E is calculated. This new conformation is accepted or rejected depending on the Boltzmann factor

$$\mathrm{BF} = \mathrm{e}^{-(E-E_0)/RT}.$$

The BF value is compared with a random number (RN) between 0 and 1. If BF ≥ RN, this new conformation is accepted; otherwise, it is rejected. This process is continued until a set of low energy conformers has been generated. Several modified Monte Carlo methods have been applied for peptides and proteins[16,17].

Simulated annealing methods

Simulated annealing is based on a connection between statistical mechanics and the process of crystallization. If a physical system is heated until it melts, and then cooled slowly, the entire arrangement can be made to produce the most stable (crystalline) arrangement, and not get trapped in a local minimum. Kirkpatrick and co-workers[18] first applied this strategy to computational optimization of a multivariable function. The same phenomenon can also be applied to the Monte Carlo Metropolis sampling by establishing the correspondence between the energy and the objective function of the optimization problem. The method is widely used for the conformational search problems of molecules[19,20].

Genetic algorithms

The genetic algorithm is derived from the principles of natural evolution, where an initial (parent) population is a set of conformations that are randomly generated for a molecule. The fitness of each member of the population is then calculated using a fitness (energy) function. A new (children) population is then generated from the parent population (with a bias towards the fitter members), commonly using three operators, viz. selection, crossover and mutation. The cycle is repeated until user-specified termination conditions are met and the best fit population is finally obtained. Le Grand and Merz[21], and Schulze-Kremer[22] have reviewed the application of the genetic algorithm in peptides and protein structure prediction.

Distance geometry methods

A molecule is described by a distance matrix whose elements d_{ij} are the distances between atoms i and j. A matrix of upper and lower bounds for each interatomic distance is calculated. Values are then randomly assigned to each interatomic distance between its upper and lower bounds. The distance matrix is converted into a trial set of Cartesian coordinates by a process called embedding and then subjected to a minimization to find a conformer. The application of distance geometry to modelling and prediction of polypeptides has been reviewed by Taylor and Aszódi[23].

Smoothing/deformation methods

Smoothing methods rely on the assumption that the global minimum of a deformed energy hyper surface can be traced back to the global minimum of the original function. In these methods, the energy hyper surface of a molecule is smoothed (or deformed) by successively removing energy minima until only the global minimum remains. This is carried out by applying certain mathematical operators (e.g. diffusion equation). The minimum is easily located using one of the standard minimization methods. However, the position of the single minimum in this deformed function is usually different from that of the global minimum in the original one, and a reversing procedure is employed to trace back from the single minimum in the final deformed function to the related minimum in the original function which is (hopefully) at the global minimum. A number of smoothing procedures have been developed and applied to peptides and proteins[24,25].

Molecular dynamics

In molecular dynamics, one begins with a conformation that is a minimum. The atoms in the molecule are typically constrained using a force field. At regular time intervals, Newton's second law of motion is solved for all atomic degrees of freedom. New positions and velocities of the atoms are calculated, the atoms are moved to these new positions and the cycle is repeated. By performing this process for a number of time steps, the dynamic behaviour of the molecule at the desired temperature can be reproduced. At equal time intervals, conformations are selected from the trajectory and minimized. Use of an elevated temperature allows the molecule to climb over potential energy barriers to reach new regions of the PES that may contain lower-energy minima than the current region. Molecular dynamics is computationally expensive, because it requires a time step of the order of 1 fs, and is often limited to peptides[2,26].

Apart from the above methods, a number of other search methods have been developed[1–8,27]. Some of these are not only limited to polypeptides and proteins, but also can be applied to other systems. Sometimes hybrid search methods, which combine two or more of the above, are used[28].

MOLS conformational search technique

The conformational search method developed in our laboratory searches the PES in an apparently exhaustive manner to locate all the low energy conformers. This method is based on the technique of using mutually orthogonal Latin squares (MOLS) to sample the conformational space as in the field of experimental design. The method has been described elsewhere in detail[29]. Here we briefly describe only the basic idea behind the method.

In the statistical experimental designs[30–32], a variable of specific experimental interest is referred to as a factor, and its possible values are referred to as levels or treatments. Every experiment comprises a set of experimental units (often called plots or runs). The treatments are distributed or applied to the experimental units using a design such as MOLS design and yields of the plots are then analysed using statistical techniques such as analysis of variance (ANOVA), to finally arrive at a strategy to maximize the yield or to compare the effect of different treatments or levels. In other words, MOLS are used to systematically sample the space of the variables. This allows the experimenter to finish the experiment with a relatively small number of runs (M^2) instead of examining all possible combinations of values (M^N) of the variables, where N is the number of variables and M is the number of possible values of each variable. We have cast the problem of searching for the minimum energy molecular structure on the PES as a problem in experimental design. We identify the factors (i.e. variables of interest) as the conformational variables (usually the torsional angles, but other variables are also possible). The treatments or levels of the factors are their possible values (often 0 to 360°). The experimental unit is the computer model of the molecule. The use of MOLS experimental design gives M^2 sub squares (i.e. ex-

periments or plots). Each plot has a set of treatments for the factors of the experimental unit, chosen from all possible combinations of the values of the variables. There are thus M^2 combinations (or molecular conformations) to be calculated. A typical MOLS design is shown in Figure 1. The experiment consists of setting of all the conformational variables of the molecule to a specific set of values and calculating the potential energy. Thus, the response or yield of each 'experiment' in this case is the potential energy of the molecule. We obtain M^2 values of the potential energy corresponding to M^2 conformations systematically chosen to sample the entire conformational space.

In agricultural experiments, the next step in the procedure is to analyse the yields using ANOVA or similar variance analysis procedures, to identify the variables and their values that contribute most significantly to the yield. We have modified this method of analysis. We take Boltzmann weighted averages of different sets of variables to identify the optimum value for each variable, and arrive directly at the best structure. This completes one cycle of calculations which will identify one low energy structure. The structure identified in one run of MOLS may not be the global minimum energy structure, but could be just one of the many possible low energy structures of the molecule. We may repeat the procedure using a different set of MOLS to locate another low energy structure, though often we simply obtain one of the previously obtained structures again. After the procedure is repeated a sufficient number of times (about 1000 for a pentapeptide), we obtain no new structure but only previously obtained structures, indicating that we have identified all the low energy structures of the molecule. The computations may therefore be terminated at this point.

	1	2	3	4	5
I	A 1 α a	B 2 β b	C 3 γ c	D 4 δ d	E 5 ε e
II	B 5 δ c	C 1 ε d	D 2 α e	E 3 β a	A 4 γ b
III	C 4 β e	D 5 γ a	E 1 δ b	A 2 ε c	B 3 α d
IV	D 3 ε b	E 4 α c	A 5 β d	B 1 γ e	C 2 δ a
V	E 2 γ d	A 3 δ e	B 4 ε a	C 5 α b	D 1 β c

Figure 1. An example of a set of mutually orthogonal Latin squares[58], showing four MOLS of order 5. This can be used as a design for an experiment involving four (N) variables each with five (M) values. Symbols in the first Latin square are A, B, C, D, E. Each of these is repeated five times to give a total of 25 symbols, which have been arranged in a Latin square (i.e. each symbol occurs exactly once in each row and exactly once in each column). Similarly, the second, third and fourth Latin squares have been constructed using symbols 1, 2, 3, 4, 5; α, β, γ, δ, ε and a, b, c, d, e respectively, in such a way that each Latin square is orthogonal to the other three Latin squares (i.e. each symbol of one square occurs once, and only once, with each symbol of the other Latin squares). One could use a set of M symbols and construct M-1 MOLS of order M. In the peptide structure optimization experiment, each symbol within the sub square represents a possible value for the corresponding torsion angle, and each sub square represents a possible conformation of the molecule.

As we shall show now, we have successfully applied this technique to identify low energy structures for a variety of small peptides at negligible computational cost. The algorithm was tested on some mathematical functions designed with well-identified optima, as well as on a tetrapeptide and a dinucleotide[33]. The method successfully picked up the optimal values each time. In the following sections we describe the further validations, applications, automation and future directions of this new conformational search technique.

Preliminary validations

In order to verify whether the MOLS search was truly exhaustive, and that all low energy conformers of a molecule could be picked up, it was applied to a test peptide molecule $(Ala)_3$. The results were compared with those from an exhaustive search. A molecule of this size (three residues) was chosen for the comparison, since an exhaustive search (for comparative purposes) becomes impractical for a molecule bigger than this size. At the same time, the MOLS search does not really have any advantage for a molecule smaller than this. (The exhaustive search took 786 min of CPU time, and the MOLS search took only 19 min CPU time, both on a single 650 MHz Pentium III processor.) The comparison (Table 1) corroborated the fact that the MOLS search algorithm was exhaustive, fast and had not missed any low energy structure[29].

The algorithm was next applied to model homopolypeptides (polyalanine and polyglycine) of various lengths. The results for one of them, viz. $(Ala)_5$, are shown here. A total of 1500 structures were generated using the ECEPP/3 force field[11]. This took about 3 h on a single 1.8 GHz Pentium IV processor. The (ϕ, ψ) values of all the structures obtained by the MOLS method fall in the allowed region of the Ramachandran map (Figure 2), which shows that the obtained structures are stereochemically sound. In other words, the method is successful in avoiding unphysical conformations and in identifying optimal ones. The conformation with the lowest energy among 1500 is a right-handed α-helical one. This conformation has been shown as most stable for polyalanine by earlier studies[19,34,35]. The method also identifies other regular structures for this pentapeptide such as right-handed 3_{10} helices, 2.2_7 helices[36], β-turns[37], γ-turns[38] and extended conformations. When classified according to the Lewis criteria[39], about 5% of conformations is helices ($R_{i,i+3} < 6$ Å), 1% is fully extended ($R_{i,i+3} \sim 11$ Å), 58% is extended ($R_{i,i+3} > 7$ Å), and remaining are folded (at least one $R_{i,i+3} < 7$ Å) conformations. $R_{i,i+3}$ is known as characteristic distance (i.e. the distance between the C^α atom of residue i and the C^α atom of residue $i + 3$).

Table 1. Comparison of complete grid (or exhaustive) search with MOLS search for $(Ala)_3$. The last column specifies RMSD between representative structures obtained by complete grid search and MOLS search. It is clear that the two sets of structures are the same (for details see Vengadesan and Gautham[29])

MOLS search			Systematic search			
Cluster rank	Cluster size	Average energy (kcal/mol)	Cluster rank	Cluster size	Average energy (kcal/mol)	Comparison RMSD (Å)
1	48	–7.38	1	82	–7.41	0.01
2	73	–7.17	2	21	–7.25	0.01
3	40	–7.03	3	79	–7.06	0.01
5	64	–6.96	4	287	–6.96	0.04
6	71	–6.84	5	62	–6.85	0.05
8	78	–6.78	6	468	–6.79	0.02
30	1	–6.02	7	1	–6.07	0.91

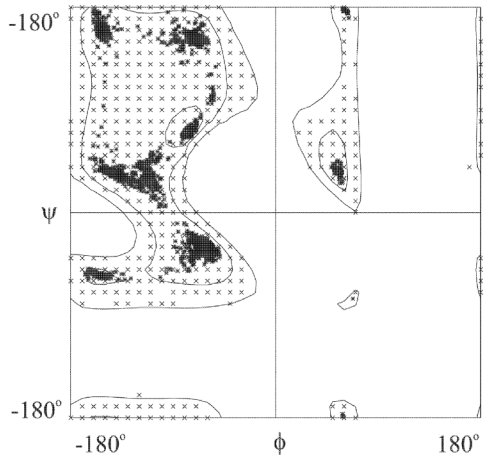

Figure 2. Ramachandran map for $(Ala)_5$, plotted for the 1500 optimal conformations obtained by repeated application of MOLS procedure. Contours in the map were calculated for $(Ala)_2$ using the ECEPP/3 energy function[11]. Three contour levels have been shown corresponding to the energy levels 10.4, 12.4 and 18.4 kcal/mol. Crosses (x) represent discrete conformations picked up by MOLS method. Stars (*) are the same structures after a few cycles of gradient minimization.

Further tests on pentapeptides (Met-enkephalin, Leu-enkephalin, $(Aib)_5$) and decapeptides again showed that the method is able to identify the previously determined experimental and theoretical structures successfully. For $(Aib)_5$, a total of 1500 structure were generated using the ECEPP/3 force field[11]. The Ramachandran map was used to characterize the 1500 structures. Since α-aminoisobutyric acid has two methyl groups attached to the C^{α} atom, the allowed region in the Ramachandran conformational space is more restricted. All the structures fell in the low energy regions of the map. The lowest energy conformation among 1500 is the 3_{10} helical conformation (Figure 3), as observed in crystallographic studies[40,41] and other conformational analysis studies[35]. The residue Aib is known to favour both right-handed and left-handed helices[42]. Both these helices were present among the structures generated. The energy difference between the lowest energy conformation between left and right-handed 3_{10} helices was only 0.04 kcal/mol. α-helices, 2.2_7 helices, β-turns, γ-turns and extended conformations are other conformations identified.

We also examined the sampling quality of the MOLS search algorithm using the sample overlap procedure[43]. According to this procedure, if two conformation samples of the same system generated by two different sampling protocols (e.g. different initial conditions or different methods) overlap and occupy the same conformation space, then the sampling procedure is exhaustive. On the other hand, if the conformation samples do not overlap or have little overlap, then the sampling is incomplete. Evaluation of the overlap between two samples can be carried out with the aid of principal component projections through a joint projection onto the same low-dimensional principal subspace. The two samples generated for $(Ala)_5$ by the MOLS technique occupy almost exactly the same regions as shown in Figure 4, indicating that the conformation sample covers the entire available conformation space.

Applications

Conformational studies of peptides

The MOLS technique was applied for detailed conformational studies of the two neuropeptides: Met-enkephalin and Leu-enkephalin[44]. Numerous studies on these neuropeptides and their analogues have been performed using various techniques to define the biologically active structure or structures, and elucidate the mechanism of action at the receptor site[45–47]. In general, the studies indicate two different structural forms of enkephalins. One is a folded conformation and the other is an extended structure. Both the forms are a large family of structures, the members of which differ in the location and type of bend in the backbone, and in the extent of conformational variability in the side chains.

The ECEPP/3 force field[11] was used in the MOLS procedure. A set (1500) of low energy structures was obtained for each pentapeptide. The set of low energy structures of each peptide was analysed in terms of the minimum energy structures, structural motifs, comparison with crystal structures, backbone hydrogen bonding pattern, probability distribution of dihedral angles, aromatic–aromatic interactions and unique structures[44]. The technique located conformations that had been identified by experimental methods, as well as other theoretical methods. In addition there were some new tightly folded conformations. Figure 5 shows two conformations obtained by the MOLS procedure, which are similar to the GEM structure of Met-enkephalin identified by Li and Scheraga[16]. Some of the low energy structures identified in the MOLS technique corresponded to the fully extended conformations seen in X-ray crystallography studies[48]. The MOLS structure closest to one of these is shown in Figure 6. However, these MOLS structures are not the ones with the lowest energy among the 1500 structures. This may be because the extended conformation is stabilized in crystal packing by intermolecular hydrogen bonds. The force field used in the MOLS procedure has not considered such intermolecular interactions.

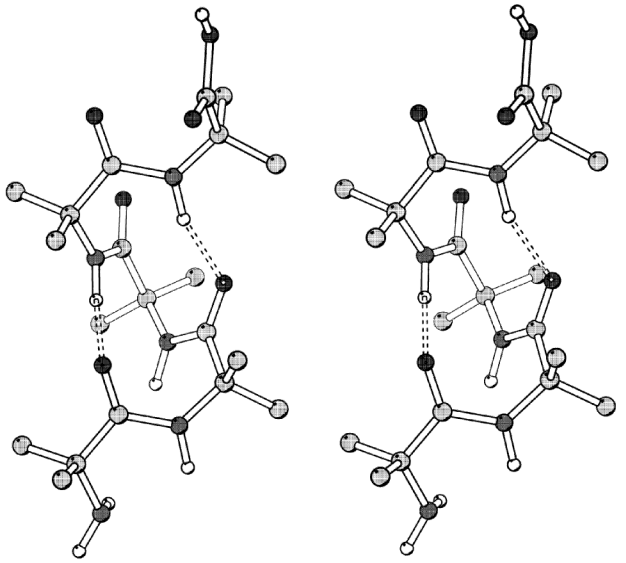

Figure 3. Stereo view of lowest energy structure of $(Aib)_5$ obtained by MOLS technique. Energy of the conformation is –1.2 kcal/mol and structure is right-handed 3_{10} helical stabilized by two hydrogen bonds $Aib^4(NH) \rightarrow Aib^1(CO)$ and $Aib^5(NH) \rightarrow Aib^2(CO)$.

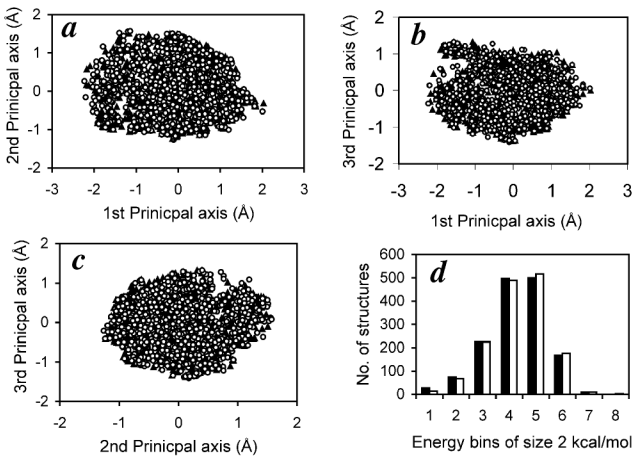

Figure 4. Two different conformation samples (consisting of 1500 conformations each) of $(Ala)_5$ simultaneously projected onto three two-dimensional planes [(*a*), (*b*), and (*c*)], each figure representing a pair from the first three principal axes (which together account for more than 41% of the variance in the data). Open circles represent conformations from the first conformation sample, filled triangles from the second. *d*, Distribution of energies in the two $(Ala)_5$ conformation samples. Number of conformations per energy bin of size 2.0 kcal/mol for conformation sample 1 (black bars) and sample 2 (white bars).

Figure 5. Stereo view of two closest MOLS structures of Met-enkephalin superimposed on the global minimum structure (shown in white) identified by Li and Scheraga[16]. RMSD (for all atom superposition) and energy values of these two structures are 1.8 and 1.5 Å, and –10.1 and –9.0 kcal/mol respectively.

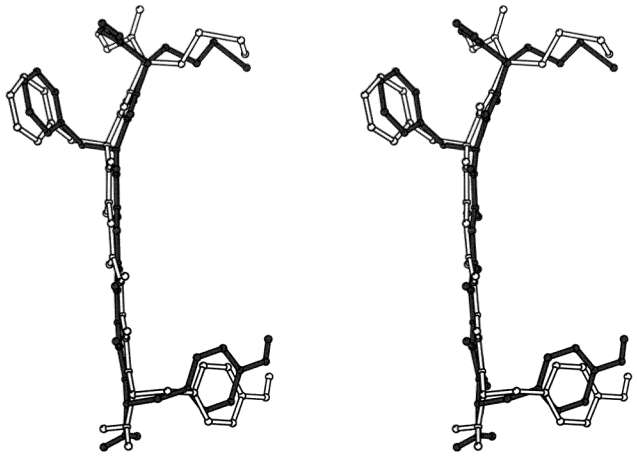

Figure 6. Stereo view of one of the closest extended MOLS structures superimposed on crystal structure of Met-enkephalin (Griffin *et al.*[48], shown in white). RMSD is 1.7 Å and energy of the MOLS structure is 2.6 kcal/mol.

The conformations and their distribution are similar for both enkephalins. Several studies[45] indicate that the folded conformation binds to the μ-receptor and extended conformation to the δ-receptor. The present results have identified both types of conformations. However, the fraction of folded structures was low compared to extended conformations. This may indicate that the latter occur much more frequently than the former, though folded conformations have the lowest energies. The results also indicate the flexibility inherent in the enkephalin molecules, since three different families of conformers, extended, folded and tightly folded, have been observed, all with low energies. Such flexibility is also indicated by the results of other experimental and computational studies[45–47].

Mapping potential energy surface of peptides

The MOLS technique has also been applied to map the potential energy landscape of Met-enkephalin and Leu-enkephalin[49]. Samples obtained from the MOLS technique were utilized to visualize the energy landscapes using various order parameters: principal coordinates projection with a minimal energy envelope procedure[43], Hamming distance[50], and radius of gyration. The positions of experimental structures were also identified on the energy landscape of each peptide. Figure 7 shows the minimal energy envelope of the landscape for Met-enkephalin. The energy landscape of both peptides possessed a significant amount of roughness and resembled a broad funnel. This indicates that there is no deep minimum in the energy landscape, and that the funnel is wide even at the bottom. The partial folded conformations (mixture of 2.2_7 helices, or combinations of γ-turns) could be intermediates during the folding process – they occur in all different energy ranges and in all positions in the landscape.

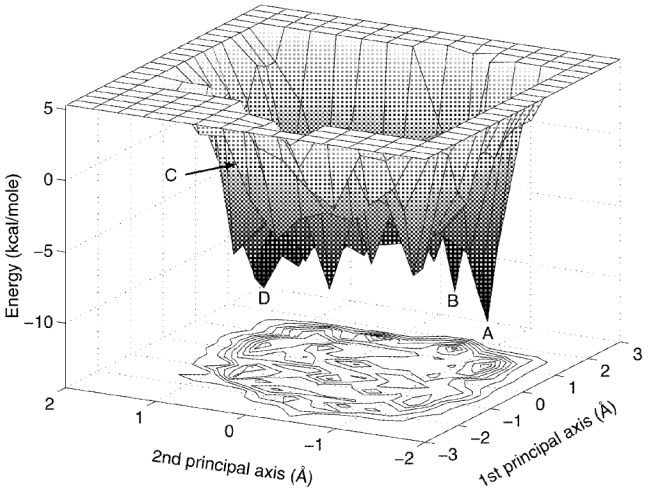

Figure 7. Energy landscape obtained by minimal energy envelope procedure for Met-enkephalin[49]. The two principal axes indicate conformational similarity, and the vertical axis reflects relative energy. Letters A (5 → 2 β-turn), B (4 → 2 β-turn), C (extended), and D (α-helical turn) marked on the energy landscape correspond to the experimental and theoretical minima.

Comparing the effects of different conformational parameters

The MOLS algorithm was also used to compare the effect of the different conformational parameters, and identify the most significant ones[51]. The traditional ANOVA technique[32] is used to analyse the MOLS samples and it helps identify the equality of the effect of torsion angles on the conformational potential energy. The results of the analyses of few pentapeptides and decapeptides show that different torsion angles contribute differently to the conformational energy. In general torsion angles at the ends of the peptide chain do not play a significant role in determining the molecular potential energy corresponding to a particular conformation, while angles in the middle do. In particular for the smaller peptides, and to a much lesser extent for larger ones, torsions that control interactions between bulky amino acid side chains exert a greater influence on the energy. A feature common to all peptides studied is the greater significance of φ angles as compared to ψ angles. For example, the complete analysis of variance for Met-enkephalin experiment is summarized in Table 2. It is clear from Table 2 that the variable ϕ_4 has the greatest effect on the conformational energy. Variations in this angle affect the interactions between the two bulky aromatic side chains of Tyr^1 and Phe^4. Other angles that strongly affect the conformation occur at the centre of the molecule ϕ_2, ψ_2, ϕ_3, ψ_3, ϕ_4, ψ_4, ϕ_5. All these affect the relative orientation of the two bulky aromatic groups. Among these ϕ_2, ϕ_3, ϕ_4 and ϕ_5 show the greatest effect. This is probably related to the fact that in the Ramachandran map, φ has a more restricted range of values than ψ. These results were consistent while the experiment was repeated by choosing a different set of MOLS.

Automation

We have automated the MOLS procedure and written a computer program to carry out the different tasks. The whole procedure consists of five main tasks, viz. building the initial peptide molecule, generation of conformational and energy parameters, generation of optimal conformations by MOLS search technique, minimization and clustering. Programs (PDBGEN, PARGEN, MOLS, MINIMIZ, CLUSTER) have been written in FORTRAN 77 to perform all these tasks. They have been combined with a GUI so that user-interaction is simplified. This program package can be used to generate a complete set of optimal conformations for any linear peptide by just giving few inputs such as sequence, number of cycles (structures), force field option, etc.

Table 2. Analysis of variance for Met-enkephalin experiment[51]. The degree of significance of each variable is specified using *P*-value (small values indicating a greater effect of the corresponding variable) with its significance code (***, $P < 0.001$; *, $P < 0.05$; NS, $P \geq 0.5$) in columns 6 and 7

Source of variation	Degrees of freedom	Sum of squares	Mean square	F_0	*P*-value	Significance code
ϕ_1	36	515.2	14.3	0.6000	9.71×10^{-01}	NS
ψ_1	36	660.8	18.4	0.7696	8.35×10^{-01}	NS
ϕ_2	36	3108.3	86.3	3.6198	1.08×10^{-11}	***
ψ_2	36	1239.1	34.4	1.4430	4.53×10^{-02}	*
ϕ_3	36	2347.5	65.2	2.7338	2.77×10^{-07}	***
ψ_3	36	2142.1	59.5	2.4946	3.59×10^{-06}	***
ϕ_4	36	9238.2	256.6	10.7582	3.33×10^{-16}	***
ψ_4	36	1321.0	36.7	1.5383	2.31×10^{-02}	*
ϕ_5	36	3421.4	95.0	3.9844	1.38×10^{-13}	***
ψ_5	36	958.0	26.6	1.1157	2.95×10^{-01}	NS
Error	1008	24043.8	23.9			
Total	1368	48995.6				

```
Initialize the communications environment
    Build initial model peptide (PDBGEN)
    Generate topological information (PARGEN)
    Allocate data to each processor
    If I am processor 0
        Do MOLS search (MOLS) and minimization (MINIMIZ) for allotted data
        Receive the results of other processors
        Do clustering (CLUSTER)
    Else
        Do MOLS search (MOLS) and minimization (MINIMIZ) for allotted data
        Send the results to processor 0
    End if
Close communications
```

Figure 8. Pseudo code of MOLS procedure parallelization.

The parallelization approach is widely used in conformational search algorithms[52,53] to speed up the conformational search and reduce the computation time. This kind of approach requires a parallel program, a parallel (or cluster) computer and parallel programming paradigm (message passing). The MOLS algorithm may be parallelized at different levels. We adopted the domain decomposition approach (SPMD model) for parallelization of MOLS procedure, in which the data (no. of cycles of the MOLS procedure) are divided into pieces of the same size and then assigned to different processors. Each processor then works only on the portion of the data that is assigned to it. Since the code (i.e. instructions applied to the data) is identical on all processors, the processors can operate independently on large portions of data, with less communication. Since the MOLS search (*MOLS*) and minimization (*MINIMIZ*) are time-consuming tasks in the MOLS procedure, they are parallelized as shown in the pseudo code (Figure 8).

We followed an inexpensive method of Beowulf cluster concept[54,55] (http: //dune.mcs.kent.edu/~parallel/equip/beowulf/) for building a parallel computer with little special hardware. We used Pentium-based PCs with Linux OS they were inter-connected via Local Area Network (LAN) and free message-passing software such as parallel virtual machine (PVM)[55], and several implementations of message-passing interface[55]; MPICH, LAM MPI. Initially we built two-nodes cluster computer and used PVM and MPICH. Later we built five-nodes and nine-nodes cluster computers and used LAM MPI. In the nine-nodes cluster, each node has the following specifications: single 1.8 GHz Pentium IV Processor, 40 GB hard disk, 256 MB RAM and Intel (R) PRO/100 fast ethernet. The Red Hat Linux version 7.3 was installed on all nodes. The nodes are connected through LAN by 16 Port 100Base TX Switch (DX-5016PS). A single monitor, keyboard and mouse are connected to all nodes via a KVM switch (MAXPORT-ACS-1216A).

The performance of the MOLS parallel implementation was examined on the cluster computers for the test case pentapeptide $(Ala)_5$. Five hundred structures were generated, i.e. the number of cycles was set as 500. This job took 4 min 40 s to generate 500 conformations on the nine nodes cluster computer. The same job took 39 min 14 s on a single processor computer. The degree of parallelization was analysed using two measures, time speedup and efficiency (i.e., the percent of parallelization)[55]. For the above test problem speedup = 8.4 and efficiency = 0.93. This shows that the performance of the parallel algorithm is efficient and the parallel program spent much less time for communication. Table 3 shows the performance of parallel algorithm as a function of number of nodes. Initially the running time decreases sharply, but after about six processors the fall is not large, i.e. because with every processor added, the additional resource committed to the task, as a fraction of the entire task, becomes less and less, i.e. 1/7 is not much smaller than 1/6. The quality of the parallelization depends on how evenly the workloads are distributed among different computers and how much time is spent on data transfer between them. From Table 3 it can be clearly seen that the parallelization is high and with greater than 90% efficiency up to nine nodes.

Table 3. Performance of parallel MOLS program as a function of the number of nodes

No. of nodes	Total no. of cycles	No. of cycles per node	Running time** (min)	Speed-up	Efficiency
1	500	500	39.14	1.0	1.0
2	500	250	19.26	2.0	1.0
3	498*	166	13.11	3.0	1.0
4	500	125	10.27	3.8	0.9
5	500	100	8.06	4.8	1.0
6	498*	83	6.59	5.6	0.9
7	497*	71	5.54	6.6	0.9
8	496*	62	5.18	7.4	0.9
9	495*	55	4.4	8.4	0.9

*Total number of cycles is less than 500 to ensure that number of cycles per node is the same integer number for all nodes.
**Largest time taken by any node.

Problems and future of the work

The MOLS algorithm has three main advantages over other structure search techniques.

1. The search is unconstrained and is accomplished at little computational cost. Computational complexity of the algorithm scales (only) as the fourth power of the size of the molecule.
2. The method lends itself easily to parallelization, which would further speed up the computation.
3. The method can be used to find the optimum of a wide variety of functions because there are no major assumptions regarding the form of function.

There are, of course, the following limitations in extending the procedure directly to larger molecules.

1. Each variable must have the same number of levels and the number of levels must be a prime power.
2. The number of levels must be greater than the number of variables.
3. The basic assumption of the MOLS design is that there are no interactions among the variables.

The first two are not the major limitations. The chief problem is the basic assumption of the independence of the factors, which is true only to a limited extent. Thus, though the possibility of extending the method to larger peptides, and to *ab initio* protein structure prediction is obvious, the success of any such scheme would depend on several factors, including the development of an appropriate potential function. Application of the present form of MOLS procedure to several peptides of various lengths showed that the method was successful for peptides of length less than ten residues, and that it has to be refined and modified further for longer peptides. Possible ways to extend the method for longer peptides and other features are now discussed.

Including secondary structure information

The secondary structure of polypeptides can be predicted with high accuracy[56], and it would be useful to utilize these predictions in the MOLS conformational search algorithm in some manner. For example, this information can be used as biasing functions. Alternately, the predicted secondary structures could be first built. MOLS could then be used to position the secondary structure elements optimally.

Utilization of other experimental knowledge

The present MOLS algorithm is purely an *ab initio* technique with no inputs other than the amino acid sequences. Exploring the ways to incorporate partial or full experimental information that we may have about the structure of polypeptides will improve the results. For example, the use of residual dipolar couplings[57] from NMR experiments will help recognize the native conformations.

Build-up approach

The MOLS technique successfully predicts structures for short peptides (<10 residues). A long polypeptide chain can be divided into fragments of length about 5 to 10 residues, and then the set of low energy conformations of each fragment obtained from the MOLS method can be combined to form a complete polypeptide chain conformation by a 'build-up' approach[15].

Developing hybrid algorithms

It has become common to combine two or more conformational search algorithms together to improve the performance of conformational search, because each algorithm has its own strength and weakness[28]. Similarly, the MOLS algorithm can be combined with other available algorithms, such as genetic algorithms[21,22], etc.

Using a simplified model

The present MOLS procedure uses an all-atom representation for the molecule. Simplified models[3] such as united

atom model, bead model, lattice model, etc. are in common use for protein structure prediction. One can use any of these simplified models for longer peptides in the MOLS procedure.

Refinement of the energy function and parameters

In most of the applications of the method described above, even though experimental conformations were present in the set of predicted low energy conformations, they could not be distinguished from other conformations on the basis of the energy values alone. This problem could be overcome by developing a filtering function to discriminate the native from the non-native or near native conformations.

Choice of appropriate temperature

The temperature used in the Boltzmann weighting function was set by default to 3 K. This may not be appropriate for all sizes of peptides.

Inclusion of solvent effect and intermolecular interactions

The present procedure never included any solvation model and intermolecular interactions in the energy function. The empirical energy function used here has a set of parameters calculated from crystal structures of amino acids and has been refined to adjust rotational barriers close to experimental values. Thus, it is expected that these parameters include hydration and intermolecular effects. Inclusion of intermolecular interactions and solvent effects by either an implicit or explicit solvation model might improve the results.

Modelling loop regions in the proteins

Modelling the loop regions of a protein is considered as a bottleneck in the protein structure prediction (specifically in comparative modelling). The MOLS algorithm can be used to predict the conformation of these loop regions. Initial attempts show that the method can predict loop structures with less than 1 Å RMSD with the corresponding crystal structures (V. Kanagasabai and N. Gautham, 2005, unpublished.) However, the predicted structures that were closest to the crystal structures could not be discriminated from the other conformations on the basis of energy values alone.

Application to other molecules

The MOLS algorithm is presently applied to polypeptides. However, the algorithm is not limited to any molecules and thus can be applied to other molecules such as polynucleotides, organic molecules, etc.

Prediction of ligand structures in drug design

The most daunting task in the structure-based drug discovery process is the one of designing or finding a suitable molecule which can bind to the required region of the target. Conventional database searching methods can only suggest drug candidates among known structures. In the absence of known drug structures, the MOLS algorithm can be used to generate libraries of conformations and pick up an optimal 3D model to fit the target.

Latin cubes approach

Latin cubes can be used in the place of Latin squares of the conformational search algorithm.

Application to cyclic molecules

The MOLS method is currently applied to linear molecules only. However, it can be applied to cyclic molecules with a little modification.

Apart from this, the method has many potential applications. For example, the representative models obtained from the MOLS method can be used for solution and refinement of peptide crystal structures. Another possible application is in the area of drug design. The docking of a candidate drug molecule at its target site may be optimized simultaneously with the structure of the drug molecule. Some of these possibilities are being explored further in our laboratory.

1. Howard, A. E. and Kollman, P. A., An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.*, 1988, **31**, 1669–1675.
2. Leach, A. R., A survey of methods for searching the conformational space of small and medium-sized molecules. In *Reviews in Computational Chemistry* (eds Lipkowitz, K. B. and Boyd, D. B.), VCH Publishers, New York, 1991, vol. 2, pp. 1–55.
3. Vásquez, M., Némethy, G. and Scheraga, H. A., Conformational energy calculations on polypeptides and protein. *Chem. Rev.*, 1994, **94**, 2183–2239.
4. Eisenhaber, F., Persson, B. and Argos, P., Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.*, 1995, **30**, 1–94.
5. Böhm, G., New approaches in molecular structure prediction. *Biophys. Chem.*, 1996, **59**, 1–32.
6. Neumaier, A., Molecular modeling of proteins and mathematical prediction of protein structure. *SIAM Rev.*, 1997, **39**, 407–460.
7. Scheraga, H. A., Lee, J., Pillardy, J., Ye, Y. J., Liwo, A. and Ripoll, D., Surmounting the multiple-minima problem in protein folding. *J. Global Optim.*, 1999, **15**, 235–260.
8. Floudas, C. A., Klepeis, J. L. and Pardalos, P. M., Global optimization approaches in protein folding and peptide docking. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* (eds Farach-Colton, M. *et al.*), American Mathematical Society, New Jersey, 1999, vol. 47, pp. 141–171.
9. Weiner, S. J., Kollman, P. A., Nguyen, D. T. and Case, D. A., An all atom force field for simulations of proteins and nucleic acids. *J. Comput. Chem.*, 1986, **7**, 230–252.

10. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. and Karplus, M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 1983, **4**, 187–217.
11. Némethy, G. *et al.*, Energy parameters in polypeptides. 10. Improved geometrical parameters and non-bonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J. Phys. Chem.*, 1992, **96**, 6472–6484.
12. Schlick, T., Optimization methods in computational chemistry. In *Reviews in Computational Chemistry* (eds Lipkowitz, K. B. and Boyd, D. B.), VCH Publishers, New York, 1992, vol. 3, pp. 1–71.
13. Ngo, J. T. and Marks, J., Computational complexity of a problem in molecular structure prediction. *Protein Eng.*, 1992, **5**, 313–321.
14. Bruccoleri, R. E. and Karplus, M., Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 1987, **26**, 137–168.
15. Gibson, K. D. and Scheraga, H. A., Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization. *J. Comput. Chem.*, 1987, **8**, 826–834.
16. Li, Z. and Scheraga, H. A., Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 6611–6615.
17. Ripoll, D. R. and Scheraga, H. A., The multiple-minima problem in the conformational analysis of polypeptides. III. An electrostatically driven Monte Carlo method: Tests on enkephalin. *J. Protein Chem.*, 1989, **8**, 263–287.
18. Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P., Optimization by simulated annealing. *Science*, 1983, **220**, 671–680.
19. Wilson, S. R. and Cui, W. L., Applications of simulated annealing to peptides. *Biopolymers*, 1990, **29**, 225–235.
20. Morales, L. B., Garduno-Juárez, R. and Romero, D., Applications of simulated annealing to the multiple-minima problem in small peptides. *J. Biomol. Struct. Dyn.*, 1991, **8**, 721–735.
21. Le Grand, S. M. and Merz, K. M., The genetic algorithm and protein structure prediction. In *The Protein Folding Problem and Tertiary Structure Prediction* (eds Merz, K. M. and Le Grand, S. M.), Birkhäuser, Boston, 1994, pp. 109–124.
22. Schulze-Kremer, S., Genetic algorithm and protein folding. In *Protein Structure Prediction – Methods and Protocols* (ed. Webster, D. M.), Humana Press, New Jersey, 2000, pp. 175–222.
23. Taylor, W. R. and Aszódi, A., Building protein folds using distance geometry: Towards a general modeling and prediction method. In *The Protein Folding Problem and Tertiary Structure Prediction* (eds Merz, K. M. and Le Grand, S. M.), Birkhäuser, Boston, 1994, pp. 165–192.
24. Piela, L., Kostrowicki, J. and Scheraga, H. A., The multiple-minima problem in the conformational analysis of molecules: Deformation of potential energy hyper surface by diffusion equation method. *J. Phys. Chem.*, 1989, **93**, 3339–3346.
25. Kostrowicki, J. and Scheraga, H. A., Application of diffusion equation method for global optimization to oligopeptides. *J. Phys. Chem.*, 1992, **96**, 7442–7449.
26. Gibson, K. D. and Scheraga, H. A., Variable step molecular dynamics: An exploratory technique for peptides with fixed geometry. *J. Comput. Chem.*, 1990, **11**, 468–486.
27. Saunders, M., Houk, K. N., Wu, Y. D., Still, W. C., Lipton, M., Chang, G. and Guida, W. C., Conformations of cycloheptadecane. A comparison of methods for conformational searching. *J. Am. Chem. Soc.*, 1990, **112**, 1419–1427.
28. Klepeis, J. L., Pieja, M. J. and Floudas, C. A., Hybrid global optimization algorithms for protein structure prediction: Alternating hybrids. *Biophys. J.*, 2003, **84**, 869–882.
29. Vengadesan, K. and Gautham, N., Enhanced sampling of the molecular potential energy surface using mutually orthogonal Latin squares: Application to peptide structures. *Biophys. J.*, 2003, **84**, 2897–2906.
30. Finney, D. J., Randomized blocks and Latin squares. In *Experimental Design and its Statistical Basis*, Cambridge University Press, London, 1955, pp. 45–67.
31. Fisher, R. A., The Latin square. In *The Design of Experiments*, Oliver and Boyd, London, 1960, 7th edn, pp. 70–92.
32. Montgomery, D. C., Randomized blocks, Latin squares, and related designs. In *Design and Analysis of Experiments*, John Wiley, New York, 2000, 5th edn, pp. 126–169.
33. Gautham, N. and Rafi, Z. A., Global search for optimal biomolecular structures using mutually orthogonal Latin squares. *Curr. Sci.*, 1992, **63**, 560–564.
34. Hopfinger, A. J., *Conformational Properties of Macromolecules*, Academic Press, New York, 1973, pp. 82–93.
35. Improta, R., Barone, V., Kudin, K. N. and Scuseria, G. E., Structure and conformational behavior of biopolymers by density functional calculations employing periodic boundary conditions. I. The case of polyglycine, polyalanine, and poly-α-aminoisobutyric acid *in vacuo*. *J. Am. Chem. Soc.*, 2001, **123**, 3311–3322.
36. Ramakrishnan, C. and Ramachandran, G. N., Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys. J.*, 1965, **5**, 909–933.
37. Venkatachalam, C. M., Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers*, 1968, **6**, 1425–1436.
38. Rose, G. D., Gierasch, L. M. and Smith, J. A., Turns in peptides and proteins. *Adv. Protein Chem.*, 1985, **37**, 1–107.
39. Lewis, P. N., Momany, F. A. and Scheraga, H. A., Chain reversals in proteins. *Biochim. Biophys. Acta*, 1973, **303**, 211–229.
40. Shamala, N., Nagaraj, R. and Balaram, P., The 3_{10} helical conformation of a pentapeptide containing α-aminoisobutyric acid (Aib): X-ray crystal structure of Tos-$(Aib)_5$-OMe. *J. Chem. Soc. Chem. Commun.*, 1978, 996–997.
41. Benedetti, E. *et al.*, Solid-state and solution conformation of homo oligo(α-aminoisobutyric acids) from tripeptide to pentapeptide: Evidence for a 3_{10} helix. *J. Am. Chem. Soc.*, 1982, **104**, 2437–2444.
42. Burgess, A. W. and Leach, S. J., An obligatory α-helical amino acid residue. *Biopolymers*, 1973, **12**, 2599–2605.
43. Levy, Y. and Becker, O. M., Energy landscapes of conformationally constrained peptides. *J. Chem. Phys.*, 2001, **14**, 993–1009.
44. Vengadesan, K. and Gautham, N., Conformational studies on enkephalins using the MOLS technique. *Biopolymers*, 2004, **74**, 476–494.
45. Schiller, P. W., Conformational analysis of enkephalin and conformation–activity relationships. In *The Peptides: Analysis, Synthesis, Biology* (eds Udenfriend, S. and Meienhofer, J.), Academic Press, New York, 1984, vol. 6, pp. 219–268.
46. Deschamps, J. R., George, C. and Flippen-Anderson, J. L., Structural studies of opioid peptides: A review of recent progress in X-ray diffraction studies. *Biopolymers*, 1996, **40**, 121–139.
47. van der Spoel, D. and Berendsen, H. J. C., Molecular dynamics simulations of Leu-enkephalin in water and DMSO. *Biophys. J.*, 1997, **72**, 2032–2041.
48. Griffin, J. F., Langs, D. A., Smith, G. D., Blundell, T. L., Tickle, I. J. and Bedarkar, S., 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*, 1986, **83**, 3272–3276.
49. Vengadesan, K. and Gautham, N., The energy landscape of Met-enkephalin and Leu-enkephalin drawn using mutually orthogonal Latin squares sampling. *J. Phys. Chem. B*, 2004, **108**, 11196–11205.
50. Abdali, S., Jensen, M. Ø. and Bohr, H., Energy levels and quantum states of [Leu]enkephalin conformations based on theoretical and experimental investigations. *J. Phys. Condens. Matter*, 2003, **15**, S1853–S1860.
51. Vengadesan, K., Anbupalam, T. and Gautham, N., An application of experimental design using mutually orthogonal Latin squares in

conformational studies of peptides. *Biochem. Biophys. Res. Commun.*, 2004, **316**, 731–737.
52. Lee, J. *et al.*, Efficient parallel algorithms in global optimization of potential energy functions for peptides, proteins, and crystals. *Comput. Phys. Commun.*, 2000, **128**, 399–411.
53. Morales, L. B., Garduno-Juárez, R., Aguilar-Alvarado, J. M. and Riveros-Castro, F. J., A parallel tabu search for conformational energy optimization of oligopeptides. *J. Comput. Chem.*, 2000, **21**, 147–156.
54. Hargrove, W. W., Hoffman, F. M. and Sterling, T., The do it yourself supercomputer. *Sci. Am.*, 2001, **285**, 72–79.
55. Rajaraman, V. and Siva Ram Murthy, C., *Parallel Computers – Architecture and Programming*, Prentice-Hall of India Private Limited, New Delhi, 2000.
56. Rost, B. and Sander, C. J., Prediction of protein secondary structure at better than 70% accuracy. *Mol. Biol.*, 1993, **232**, 584–599.
57. Prestegard, J. H., Al-Hashimi, H. M. and Tolman, J. R., NMR strucutres of biomolecules using field oriented media and residual dipolar couplings. *Q. Rev. Biophys.*, 2000, **33**, 371–424.
58. Ito, K., Latin squares. In *Encyclopedic Dictionary of Mathematics*, MIT Press, Cambridge, Massachusetts, 1987, vol. 2, pp. 891–892.

ACKNOWLEDGEMENTS. We thank the Council of Scientific and Industrial Research, the Department of Science and Technology, and the Universities Grants Commission, New Delhi for financial support.

Received 12 November 2004; revised accepted 14 February 2005