MOLS sampling and its applications in structural biophysics

L. Ramya · Shankaran Nehru Viji ·
Pandurangan Arun Prasad · Vadivel Kanagasabai ·
Namasivayam Gautham

Received: 14 October 2010 / Accepted: 19 October 2010 / Published online: 16 November 2010
© International Union for Pure and Applied Biophysics (IUPAB) and Springer 2010

Abstract This review describes the MOLS method and its applications. This computational method has been developed in our laboratory primarily to explore the conformational space of small peptides and identify features of interest, particularly the minima, i.e., the low energy conformations. A systematic “brute-force” search through the vast conformational space for such features faces the insurmountable problem of combinatorial explosion, whilst other techniques, e.g., Monte Carlo searches, are somewhat limited in their region of exploration and may be considered inexhaustive. The MOLS method, on the other hand, uses a sampling technique commonly employed in experimental design theory to identify a small sample of the conformational space that nevertheless retains information about the entire space. The information is extracted using a technique that is a variant of the self-consistent mean field technique, which has been used to identify, for example, the optimal set of side-chain conformations in a protein. Applications of the MOLS method to understand peptide structure, predict the structures of loops in proteins, predict three-dimensional structures of

small proteins, and arrive at the best conformation, orientation, and positions of a small molecule ligand in a protein receptor site have all yielded satisfactory results.

Keywords Peptide and protein conformations · MOLS sampling · Conformational energy landscape · Docking

Introduction

The use of mutually orthogonal Latin squares (MOLS) to systematically sample a very large experimental space is well established (Mandl 1985; Colbourn 1995; Colbourn and Dinitz 2001; Abel et al. 2006). This sampling technique has long been employed in the design of experiments to plan, for example, agricultural or clinical experiments (Fisher 1951; Cox 1958; Finney 1955; Raghavarao 1971; Pearce 1983; Meinert 1986; Mead 1988; Montgomery 2000). Such experiments are typically used to identify the most important variables from among those that influence a given outcome, such as for example the crop yield in an agricultural plot. More ambitiously, the experiments may be designed to specifically identify the particular set of values for the variables that maximizes the outcome. The conceptually simplest way to achieve this end would be to try out all possible combinations of the values of the variables. Since this would obviously lead to combinatorial explosion, with astronomical numbers of possible experiments for even a small number of variables, experimental design strategies have been developed to limit the number of experiments to be performed to a relative small, systematic sample of the entire experimental space, without seriously compromising the amount or quality of the information obtained (Cochran and Cox 1957; Cochran 1977). The use of MOLS is one such strategy.

L. Ramya · S. Nehru Viji · N. Gautham (✉)
Centre of Advanced Study in Crystallography and Biophysics,
University of Madras,
Chennai 600025, India
e-mail: n_gautham@hotmail.com

P. Arun Prasad
Institute of Structural and Molecular Biology and Crystallography,
Department of Biological Sciences, Birkbeck College,
University of London,
London, UK

Present Address:
V. Kanagasabai
Department of Orthopaedic Surgery,
David Geffen School of Medicine, University of California,
Los Angeles, CA, USA

The application of a similar strategy to explore the conformational space of molecules was conceived on the basis of the broad similarity that exists between, for example, the agricultural problem, and that of molecular structure. The conformational space of a molecule may be represented by its potential energy hypersurface (Piela et al. 1989), which is a function of the stereochemical parameters such as bond lengths, bond angles, and torsion angles. A complete mapping of the features (such as the minima) on this hypersurface by calculating the potential energy of every possible conformation of the molecule would require impossibly large amounts of computational resources. We have attempted to solve this problem by applying the methodology of experimental design, in particular that of mutually orthogonal Latin squares (MOLS) sampling, and have obtained encouraging results (Gautham and Rafi 1992). The large conformational space is represented by a small sample that is selected using MOLS. This way of selecting the sample ensures that much of the information about the space is preserved and can be retrieved, usually by statistical analyses of the sample. We have however made a further innovation in this step. Instead of using techniques such as ANOVA (Lindman 1974), the sample is analyzed by a variant of the self-consistent mean field technique (SCMFT, Koehl and Delarue 1994, 1995, 1996; Finkelstein and Reva 1996). Thus we may also describe the method in terms of SCMFT, as an extension or variation of it. In the next section, we briefly describe the technique. In the subsequent sections, we describe various applications of the method, together with further extensions and modifications to the methodology that we have carried out.

The MOLS method

A Latin square of order N consists of N copies of each of N symbols (i.e., a total of N^2 symbols) arranged in an N by N array such that each symbol occurs exactly once in each row and exactly once in each column (Ito 1987). Figure 1a gives an example. A pair of orthogonal Latin squares consists of two separate Latin squares, created according to the above definition, with the additional requirement that when the two squares are superposed, all N^2 possible combinations of the symbols of one square with the symbols of the other are found to occur, and each symbol of the one square occurs once, and only once, with each symbol of the other square (Fig. 1b). This concept of orthogonal pairs of Latin squares can be extended to a set of m mutually orthogonal Latin squares (MOLS), each of order N , every pair of which is orthogonal (Ito 1987) (Fig. 1b). Details of the use of MOLS in experimental design are given in the document http://www.unom.ac.in/Gautham_mols.pdf and in several references therein. Their use in the exploration of molecular conforma-

Fig. 1a, b A Latin square of order 3 (adapted from Cochran and Cox 1957). **a** The Latin alphabet letters a, b, and c are used as symbols for Latin square arrangement. This pattern can be extended to any size. **b** Two mutually orthogonal Latin squares (MOLS) of order 3. The Greek letters α , β , and γ are symbols of the second Latin square, which is orthogonal to the first Latin square (adapted from Fisher 1951)

a)		
a	b	c
b	c	a
c	a	b

b)		
$a\alpha$	$b\beta$	$c\gamma$
$b\gamma$	$c\alpha$	$a\beta$
$c\beta$	$a\gamma$	$b\alpha$

tional space is also extensively discussed in the above document. Here we give a brief summary of this latter application. The MOLS method, as applied to molecular structures, is best approached as a variant of the self-consistent mean field technique applied to protein structure analyses as described in the literature.

The self-consistent mean field technique has been applied to protein structure analyses by several workers to accomplish tasks such as prediction of side chain structures (Koehl and Delarue 1995) and ligand-protein interactions (Liu et al. 2004). Following the description in Koehl and Delarue (1996), the procedure underlying these applications may be summarized as follows. Let Θ be the search space (for example, all possible side chain conformations). This is divided into a number of subspaces θ_i (for example, the conformational spaces of the individual side chains). Each such subspace has a number of states θ_{ij} (for example, the side chain rotamers), each with a probability of occurrence ρ_{ij} . The effective potential due to a state θ_{rs} of a subspace θ_r is given as

$$V^{\text{eff}}(\theta_{rs}) = \sum_{i,j} \rho_{ij} V(\theta_{rs}, \theta_{ij}) \quad (1)$$

where the summation is over all subspaces $i \neq r$, and all the states of these subspaces. $V(\theta_{rs}, \theta_{ij})$ is the potential energy of the interaction between the subspaces θ_r and θ_i , calculated with θ_r set to the state θ_{rs} , and θ_i set to the state θ_{ij} . The procedure starts by assigning $\rho_{ij} = 1/m_i$ (m_i is the number of states of the subspace θ_i , all equally probable). Equation 1 is then repeatedly used to evaluate the effective potential for all states of all subspaces. Next the probabilities ρ are re-evaluated as

$$\rho_{rs} = \exp(-V^{\text{eff}}(\theta_{rs})/RT) / \sum_q \exp(-V^{\text{eff}}(\theta_{rq})/RT) \quad (2)$$

where the summation is over all the states of the subspace r . R is the gas constant and T is the temperature. The newly

determined probabilities are then used to re-determine the effective potentials. This is iterated until the probabilities converge to a set of self-consistent values. The set of most probable states of the subspaces defines the most probable state of the system (e.g., the most probable set of side-chain conformations).

In applying this technique to the conformation of peptides, we define the subspaces as the backbone torsion angles, and the states as the values that these angles can assume. Once again, initially all values are equally probable, and the effective potential due to setting the backbone torsion angle θ_r to the value θ_{rs} is calculated using Eq. 1. However, this extension to torsion angle space is not straightforward, since the potential V in the summation cannot be calculated by simply considering the two backbone torsion angles θ_r and θ_i alone—we need to set all other torsions also to specific values. In other words, the interaction between a pair of subspaces (when the subspaces are the backbone torsion angles) does not depend only on their respective states but depends also on the states of all other subspaces. The expression $V(\theta_{rs}, \theta_{ij})$ is thus not sensible in this context, and the summation in Eq. 1 has to be performed over all possible combinations of the states of all the subspaces except θ_r . This will clearly lead to combinatorial explosion since the number of such combinations is an exponential function of the number of subspaces.

To overcome this problem, we use a small sample of the possible combinations, in other words, a small sample of the conformational space, to calculate the effective potential. We use MOLS to perform this sampling. To use MOLS to systematically sample the conformational space of peptides, we map the symbols used in the set of MOLS to the values of the variable torsion angles in the molecule. Thus, Fig. 1b would represent possible values for one torsion angle in a “molecule” that has only two variable torsion angles. a , b , and c represent possible values of one of them and α , β and γ represent the possible values of the other. If there are n such torsion angles, each with m possible values, then θ_{rs} ($r=1, n; s=1, m$) defines the search space in which the sampling is to be carried out. We build the sample by constructing n MOLS of order N . To calculate the effective potential due to setting θ_r to θ_{rs} we now use the following equations:

$$V^{\text{eff}}(\theta_{rs}) = \sum_q w_q V_q(\theta_{rs} \dots), \text{ for } q = 1, N$$

$$w_q = \exp(-V_q(\theta_{rs} \dots)/RT) / \sum_q \exp(-V_q(\theta_{rs} \dots)/RT) \quad (3)$$

The summation is over all the N points in the MOLS grid at which θ_r has been set to θ_{rs} . V_q is the potential function. The ellipsis in the expression for the potential indicates the setting of all the other torsion angles (except

θ_r), determined using the MOLS algorithm. $V^{\text{eff}}(\theta_{rs})$ is used to evaluate the probability of the value θ_{rs} for θ_r —the value with the lowest effective potential being the most probable. The set of most probable values for the angles defines the lower energy conformation of the peptide. It may be noted that in this formulation the procedure is no longer an iterative one. The weights w_q in Eq. 3 are not the same as the probabilities ρ in Eq. 2. Thus one cycle of MOLS calculations leads to one low energy conformation. To locate another low energy structure, we perform another cycle of calculations, again selecting m^2 points in the conformational space using a different set of MOLS. For n subspaces with m states each, there are $(m!)^n$ different ways of choosing a set of MOLS (Ryser 1963). Using any one of them as the basis for one cycle of calculations would lead to a low energy structure. In principle, therefore, the calculations may be repeated many times to identify all the low-energy conformations. Since many of the conformations generated in subsequent cycles are the same as one generated previously, we use a clustering algorithm (Kříž et al. 2001) to cluster the structures together and identify the unique ones. In practice, we have found that for a pentapeptide it is sufficient to generate 1,500 conformations to identify all its 50 or so low-energy conformations.

Applications of the MOLS technique

This technique of using MOLS in a variant of the SCMFT has been applied to sample the conformational space of peptides (Vengadesan and Gautham 2003, 2004a; Ramya and Gautham 2009), to analyze protein loops (Kanagasabai et al. 2007), to predict protein structure (Arunachalam et al. 2006), and to explore the docking of peptides and other small molecules to receptor sites on proteins (Arun Prasad and Gautham 2008; Viji et al. 2009).

Peptide structure

Met- and Leu-enkephalin are penta-neuropeptides with morphine-like activity. Because of their biological significance, many questions regarding their structures have been investigated. For instance, do the enkephalins assume a single preferred conformation or do they exist in equilibrium involving a number of different conformers? To what extent does the conformation depend on the peptide environment? However despite extensive efforts (Hansmann et al. 1999; Isogai et al. 1977; Kříž et al. 2001), the “true” bioactive conformation of these molecules remains a puzzle, in part due to their high conformational freedom. We have used the MOLS technique to map the conformational space of these molecules. With the ECEPP/3 force field (Vengadesan and Gautham 2003, 2004a, b), 1,500 conformations were

generated for each enkephalin. All of them were found to lie in the allowed regions of the Ramachandran map. Among them were those that corresponded to all the known experimental structures for both molecules: PDB IDs 1PLW.pdb and 1PLX.pdb (Marcotte et al. 2004), and CSD ID FABJIB.pdb (Griffin et al. 1986) for Met-enkephalin; and CSD IDs FABJEX, BIXNIF10, GEWWAG, and LENKPH11 for Leu-enkephalin. The global minimum energy (GEM) structure identified by Scheraga and co-workers (Isogai et al. 1977) was also represented in the collection. In addition there were some new tightly folded conformations (Fig. 2). This was indicative of the exhaustive nature of the conformational exploration, since the struc-

tures, in particular the experimental structures, were a mixture of both extended and well-folded conformations. The sample overlap technique (Levy and Becker 2001) was also used to check for the completeness of the MOLS sample. The multidimensional conformational space of the molecules was reduced by principal coordinate analysis (PCoA) (Becker et al. 2000; Levy and Becker 2001). Figure 3 shows the projection of two samples of 1,500 conformations of Met-enkephalin, each on the two-dimensional space of the first two principal components. The samples overlap almost entirely, indicating that one of them is sufficient to cover the conformational space of the molecules.

Fig. 2 Tightly folded conformations for **a** Met-enkephalin and **b** Leu-enkephalin obtained from MOLS method

Fig. 3 Two different conformation samples (consisting of 1,500 conformations each, represented respectively by *circles* and *triangles*) of Met-enkephalin simultaneously projected onto the plane of the first two principal axes. The distribution of the two Met-enkephalin conformation samples shows their complete overlap

The 1,500 conformations generated for each molecule are not all unique. Therefore a clustering algorithm (Kříž et al. 2001) was used to separate the structures into mutually dissimilar clusters on the basis of a backbone atom rmsd of 1.00 Å between the conformations. There were 50 such clusters for Met-enkephalin and 48 clusters for Leu-enkephalin. Each cluster was represented by the centroid structure, and these centroids were used to visualize the potential as minimum energy envelopes (Fig. 4). This analysis, along with one based on the Hamming distances (Abdali et al. 2003) of the structures from the GEM, shows that the number of low energy states decreases with a decrease in energy, and that compact structures are more favored energetically than extended structures. When the structures were placed in bins of average energy approximately 3 kcal/mol in width, even in the second lowest energy bin, which is only 2.0 kcal/mol higher than the lowest energy bin, the number of mutually different conformations is large. This indicates that there is no deep minimum in the energy landscape. The potential energy landscape is shaped approximately like a funnel, but one that is wide even at the bottom. Also, there are large populations of structures in all the energy bins, with the largest numbers situated in the intermediate energy bins. We conjecture that these could be significant structural intermediates during the folding process.

All the above simulations were carried out without explicit water or other solvent molecules. To test if explicit water would change the landscape, the calculations were repeated with CHARMM 22 force field (MacKerell et al. 1998) in the presence of water molecules (Ramya and Gautham 2009). The results show that the nonsolvated structures were closer to the GEM than the solvated ones,

Fig. 4 The energy landscape obtained by the minimal energy envelope procedure for **a** Met-enkephalin and **b** Leu-enkephalin. The two principal axes indicate conformational similarity, and the vertical axis reflects the relative energy

while the latter were closer to all three experimental structures. This is as expected since the experiments were performed on the solvated molecule and the GEM structure was computed in the absence of explicit solvent. As the experimental structure of Met-enkephalin is an extended conformation, the structures generated by MOLS in the presence of water are far more similar to the experimental structures than when the water molecules are excluded. It is clear that the intramolecular hydrogen bonds in the structures are destroyed by water, which in turn makes it closer to the experimental extended structures. The energy landscapes in presence of solvent (considering all 1,500 conformations), drawn as described earlier, show that solvation increases the width of the “folding funnel” as well as the ruggedness at the bottom of the funnel (Fig. 5). There are many more minima upon solvation than when the peptide is not solvated, though the energy differences

Fig. 5 Comparison of the energy landscapes obtained for the Met-enkephalin in the presence of **a** explicit water and **b** in vacuum. The two principal axes indicate conformational similarity, and the vertical axis indicates the relative energy

between these minima are small. The study shows that inclusion of solvation is important to correctly model the conformational behavior of the molecule. The calculations also indicate that the pentapeptide Met-enkephalin prefers extended structures in an aqueous environment, as opposed to tightly folded structures in the absence of water. Thus, the biologically relevant structure may be the extended structure as seen in the crystallographic and NMR experiments (Marcotte et al. 2004) rather than the GEM structures calculated by various workers (Isogai et al. 1977). In molecular simulation studies, the experimental structure (very often the crystal structure) is considered the native state of the molecule. In the case of the enkephalins, this is a puzzle, since experiments have observed ensembles of native conformational families, from extended to folded, while computational searches have often zeroed in on the GEM structure as the native state. Our results show that the lowest energies are populated by many conformations, as

observed by other workers as well. This multiplicity of favorable structures may be required for the enkephalins to bind multiple receptors.

Protein structure prediction

The CASP experiments (Kryshtafovych et al. 2009) have reviewed the status of protein structure prediction algorithms every 2 years. Currently the most successful *ab initio* algorithms are the following: Rosetta (Raman et al. 2009), MULTICOM-CMFR (Cheng et al. 2009), and MUFOLD-MD (Zhang et al. 2010). However even these are not perfect and the problem is not considered to have found a final solution (Ben-David et al. 2009). We attempted to extend the MOLS method in the direction of protein structure prediction. To accomplish this we have developed a hybrid algorithm combining MOLS and a genetic algorithm (Arunachalam et al. 2006). Genetic algorithms (GA) are general optimization procedures modeled on the process of natural evolution, with mutations, crossover, and replication occurring on a population of strings (Goldberg 1999; Schulze-Kremer 2000). After every round of such operations, a fitness function is used to decide which members of the population recur in the next generation. The procedure is iterated until the population converges on a single individual with the optimum fitness. The method that we developed was based in part on other applications of GA to the prediction of protein structure (Unger and Moult 1993; Bowie and Eisenberg 1994; Sun 1995; Dandekar and Argos 1994, 1996). It operates in two phases. Firstly, the protein sequence is divided into short overlapping fragments. The MOLS method is used to identify about 1,500 low energy structures for each fragment. These structures are clustered together and each cluster is represented by the centroid structure. The centroids formed a library of low energy structures for each fragment. These fragment structure libraries are then used in the second phase of the algorithm. In this phase, which is the GA phase, the protein structure is modeled as a string of torsion angles. An initial population of 100 individual structures is generated for the protein by choosing the torsion angles at random or by setting them all to be 180°. Each new generation of structures is obtained by applying four operators on this population. The operators are MUTATE, VARIATE, Crossover, and a DIVERSITY operator. The MUTATE operator, when activated for a particular residue, replaces the nine-residue fragment flanking the selected residue with a structure from the respective MOLS-generated structure library. The VARIATE operator increases or decreases the chosen torsion angle by a small fixed amount. The Crossover operator randomly pairs individuals in the population. Then the structures of a randomly selected region in the two individuals in the pair are exchanged. The

DIVERSITY operator randomly selects a residue and sets the backbone torsion angles of the next five residues to 180°. After each generation, the parent and the offspring generations are merged, and all 200 individuals are sorted on the basis of the fitness function, calculated for each structure as the conformational potential energy determined using the AMBER force field (Weiner et al. 1986). The 100 fittest individuals are selected to constitute the next generation. In general, in our applications of the above procedure, after about 500 to a few thousands of iterations of this procedure, the population converged to a single structure.

Seven target proteins were selected as the test cases to apply the above method: Avian Pancreatic Polypeptide (Blundell et al. 1981), Villin Head Piece (McKnight et al. 1997), Mellittin (Terwilliger and Eisenberg 1982), the c-Myb protein (Tahirov et al. 2002), Rab4 binding domain of Rabenosyn5 (Ethiraj et al. 2005), engrailed homeo domain (Clarke et al. 1994), and the Tryptophan Zipper (Cochran et al. 2001). These proteins were chosen because of their small size, well defined secondary structures, and absence of disulphide bonds. All the targets considered here, except Tryptophan Zipper (TZ), are α helical proteins. TZ alone contains an anti-parallel sheet. The AMBER force field (Weiner et al. 1986) was used along with a secondary structure biasing function as described by Crivelli et al. (2002). The results of the prediction are shown in Fig. 6.

Exploring the conformational space of protein loops

Loops in protein structures are regions of the polypeptide chain that connect regular secondary structures. The computational prediction of the three-dimensional structure of the loops has received extensive specific attention, and several algorithms have been developed (Sudarsanam et al. 1995; van Vlijmen and Karplus 1997; Wojcik et al. 1999; Fiser et al. 2000; DePristo et al. 2003; Deane and Blundell 2000). There have also been several attempts to classify loop structures (Leszczynski and Rose 1986; Ring et al. 1992; Efimov 1993). We used the MOLS algorithm to map the conformational space of loops (Kanagasabai et al. 2007). The loops were chosen from 3,868 nonredundant high quality protein structures from the PDB database. Loops in these structures were identified and classified according to length and structure. Representative examples were taken from each length category and subjected to the MOLS calculations. The calculations were carried out with the ECEPP/3 force field (Némethy et al. 1992) (without any solvent atoms). [They were repeated with the AMBER energy function (Weiner et al. 1986) to check for bias due the selection of the force field. No such bias was seen.] There were no constraints applied to the end points of the peptide loops. They were free to move and thus the

peptides could assume extended or folded structures that may be different from the ones seen when they are part of their respective proteins.

The results showed that despite the fact that the energy calculations did not take the rest of the protein into account, the MOLS algorithm could identify low energy structures for the loops that were closely similar (less than 2.04 Å backbone atom rmsd) to the native structure. Views of the entire conformational landscapes of these loops revealed that the native structures of the loops fell into one of the many energy minima. We concluded that structure of the loop as seen in the protein is determined primarily by its own sequence, regardless of the interactions it may make with the rest of the protein, with the solvent, or with other, neighboring molecules. One may broaden this statement as follows: Each peptide has 100 or so low energy structures of approximately equal energy value. The precise structure adopted by the loop in the protein may be chosen from among these by considerations arising from the rest of the protein structure and the solvent atoms.

The above calculations were repeated by including in the energy function a statistical mean force potential (Kanagasabai et al. 2007). Such a potential had earlier been constructed (Sippl 1995) by analysis of all known high-resolution protein structures. Since the force field was now being used only for prediction of the structures of loops, we constructed it based only on the experimentally determined loop structures in proteins. This resulted in the following observations. The mean force potential increased the chances of locating the near native structure in the conformational landscape and reduced the amount of incorrect and nonnative secondary structural features in the peptides.

Docking of peptides and small molecules to proteins

One of the crucial steps in computer-aided drug discovery is to identify a range of candidate molecules that bind at an identified site on a receptor, usually a protein, and thereby inhibit its activity (Taylor et al. 2002; Brooijmans and Kuntz 2003). Computer programs that implement in silico docking procedures have helped speed up an otherwise expensive and time-consuming procedure. Such procedures make theoretical predictions regarding the association of a flexible ligand with a protein receptor and require an efficient sampling of the conformational space of the ligand, as well as a way to arrive at the correct “pose” of the molecule in the receptor site. While several docking methodologies are currently available (Morris et al. 1998; Rarey et al. 1996; Jones et al. 1997; Makino and Kuntz 1997), we have extended the MOLS algorithm as a novel “rigid receptor flexible ligand” docking method to simultaneously sample both the “docking space” and the conformational space of the ligand. In extending the

Fig. 6 Stereo view of the superimposed predicted (*dark*) and the experimental structure (*light*) of **a** Rab4 binding domain of Rabenosyn5 (rmsd 3.6 Å), **b** Villin Head Piece (rmsd 5.2 Å), **c** engrailed homeo domain (rmsd 6.5 Å), **d** Avian Pancreatic Polypeptide (rmsd 4.0 Å), **e** the c-Myb protein (rmsd 6.1 Å), **f** Mellittin (rmsd 4.3 Å), and **g** Tryptophan Zipper (rmsd 1.6 Å)

method to the docking problem, the search space was expanded to include the docking space. If the conformation of the peptide is specified by the n torsion angles θ_1 to θ_n , six additional parameters describe its pose in the receptor site, three for the position and three for its orientation, making a total of $n+6$ dimensions in the search space (θ_1 to θ_{n+6}). We define a function (the conformational potential energy of the ligand plus the interaction energy of ligand and receptor) of these $n+6$ variables and use the MOLS method to search for the minima of the function, to simultaneously identify “correct” conformations and poses.

The method was first applied to 56 peptide-protein complexes whose structures were known (Arun Prasad and Gautham 2008). The backbone rmsd of the “best” low

energy structure (i.e., the one closest to the experimental structure) identified by MOLS is within 2.00 Å of the experimental structure in 30 of the 35 cases. In nine of these, the lowest energy pose and structure were the same as the one with the lowest rmsd as compared to the experimental structure. As for the rest of the structures, when the theoretical structures were binned by energy, the one with the lowest rmsd invariably lay in the lowest 20% energy bin. A unique and interesting feature of the algorithm is its ability to identify alternate binding modes, sometimes with lower energies than the “native” binding mode. Some of these alternatively docked structures are shown in Fig. 7. The method was further tested with small organic ligands docked to protein receptors (Viji et al.

Fig. 7 Predicted lowest energy alternate binding modes superposed onto the native structure. Alternate binding mode is shown in *light lines* and the native ligand in *dark lines* with transparent molecular surfaces

2009). Forty-five known protein ligand structures were obtained from the PDB (Berman et al. 2000), and the MOLS method was applied to these. The results of these rests were also on par with the peptide-protein complexes. Not only did the method successfully identify the correct “native” structure and binding mode as among the ones with the lowest energy, but it also identified several alternate binding modes (Fig. 8). Comparisons of the performance of the method with two other popular docking programs, namely AutoDock and GOLD, indicated that the MOLS algorithm generally covers a wider area of the search space. Though it requires greater computer time than GOLD (and approximately the same amount as AutoDock) it performs at least as well in identifying the “correct” docked structure and pose of the ligand.

Conclusion

Apart from the above applications, the MOLS method was used to show that the number of low energy native-like structures for a polypeptide is an exponential function of its

Fig. 8 Superposition of the lowest energy predicted (*light*) structure of 3ERD with the native (*dark*) in the alternative binding mode

sequence length n , i.e., $\exp(bn)/n$, where b is of the order of 1.0. The density of these structures in conformational space remains approximately constant and the increase in the numbers of native-like structures comes from an expansion in the volume of the conformational space (Kanagasabai and Gautham 2008).

We have also attempted to recast the MOLS procedure as a deterministic global optimization technique (Kanagasabai 2008). The procedure as described above is useful for identifying one optimum at a time of the given multivariate function (e.g., the potential energy function) but cannot specify if this is the global optimum. We reformulated the procedure to be an iterative one, such that the MOLS procedure is repeated several times, each time carrying over the results of the previous cycle by using them to adjust the probabilities of the values of the variables. These iterations are continued until the probabilities converge. At this point the values with the highest probabilities are identified as the global optimum. There are several variations in the way the results of the previous cycles are used to decide the probabilities in the current cycle, and one of them gave encouraging results. Further developments in this direction are being carried out.

The MOLS method, originally devised to understand the conformational states of small peptides, has now been developed to address other problems in structural chemistry and biophysics. The method can be parallelized with an efficiency of close to 0.93 (Vengadesan 2004). Since the method is perfectly general, and applicable to any multi-dimensional function, with or without any degree of independence of the variables, we anticipate that its use could become more widespread.

Acknowledgment We thank the Department of Science and Technology, Government of India and UGC, Government of India under Centre of Advanced Study program, for financial support.

References

- Abdali S, Jensen MØ, Bohr H (2003) Energy levels and quantum states of [Leu] enkephalin conformations based on theoretical and experimental investigations. *J Phys Condens Matter* 15: S1853–S1860
- Abel RJR, Colbourn CJ, Dinitz JH (2006) Mutually orthogonal Latin squares (MOLS). In: Colbourn CJ, Dinitz JH (eds) *The CRC handbook of combinatorial designs*, 2nd edn. CRC Press, Boca Raton
- Arun Prasad P, Gautham N (2008) A new peptide docking strategy using a mean field technique with mutually orthogonal Latin square sampling. *J Comput Aided Mol Des* 22:815–829
- Arunachalam J, Kanagasabai K, Gautham N (2006) Protein structure prediction using mutually orthogonal Latin squares and a genetic algorithm. *Biochem Biophys Res Commun* 342:424–433
- Becker OM, Levy Y, Ravitz O (2000) Flexibility, conformation spaces and bioactivity. *J Phys Chem B* 104:2123–2135
- Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y (2009) Assessment of CASP8 structure predictions for template free targets. *Proteins* 77:50–65
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28:235–242
- Blundell TL, Pitts JE, Tickle IJ (1981) X-ray analysis (1.4 Å resolution) of avian pancreatic polypeptide: small globular protein hormone. *Proc Natl Acad Sci USA* 78:4175–4179
- Bowie JU, Eisenberg D (1994) An evolutionary approach to folding small α -helical proteins that uses sequence information and an empirical guiding fitness function. *Proc Natl Acad Sci USA* 91:4436–4440
- Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct* 32:335–373
- Cheng J, Wang Z, Tegge AN, Eickholt J (2009) Prediction of global and local quality of CASP8 models by MULTICOM series. *Proteins* 77:181–184
- Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO (1994) Structural studies of the engrailed homeodomain. *Protein Sci* 3:1779–1787
- Cochran WG (1977) *Sampling techniques*. Wiley, New York
- Cochran WG, Cox G (1957) *Experimental designs*, 2nd edn. Wiley, New York, pp 117–133
- Cochran AG, Skelton NJ, Starovasnik MA (2001) Tryptophan zippers: stable, monomeric β -hairpins. *Proc Natl Acad Sci USA* 98:5578–5583
- Colbourn J (1995) Construction techniques for mutually orthogonal Latin squares. In: Colbourn CJ, Mahmoodian ES (eds) *Combinatorics advances*. Kluwer Academic, Amsterdam, pp 27–48
- Colbourn CJ, Dinitz JH (2001) Mutually orthogonal Latin squares: a brief survey of constructions. *J Stat Plan Inference* 95:9–48
- Cox DR (1958) *Planning of experiments*. Wiley, New York
- Crivelli S, Eskow E, Bader B, Lamberti V, Byrd R, Schnabel R, Head GT (2002) A physical approach to protein structure prediction. *Biophys J* 82:36–49
- Dandekar T, Argos P (1994) Folding the main chain of small proteins with the genetic algorithm. *J Mol Biol* 236:844–861
- Dandekar T, Argos P (1996) Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria for strand regions. *J Mol Biol* 256:645–660
- Deane CM, Blundell TL (2000) A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. *Proteins Struct Funct Genet* 40:135–144
- DePristo MA, DeBakker PIW, Lovell SC, Blundell TL (2003) Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. *Proteins Struct Funct Genet* 51:41–55
- Efimov AV (1993) Patterns of loop regions in proteins. *Curr Opin Struct Biol* 3:379–384
- Ethiraj S, Pan X, Ritacco C, Lambright DG (2005) Structural basis of family-wide Rab GTPase recognition by rabenosyn-5. *Nature* 436:415–419
- Finkelstein AV, Reva BA (1996) Search for the most stable folds of protein chains. I. Application of a self-consistent molecular field theory to a problem of protein three-dimensional structure prediction. *Protein Eng* 9:387–397
- Finney DJ (1955) Randomized blocks and Latin squares. In: *Experimental design and its statistical basis*. Cambridge University Press, London, pp 45–67
- Fiser A, Gian Do RK, Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9:1753–1773
- Fisher RA (1951) *The design of experiments*, 6th edn. Oliver and Boyd, Edinburgh
- Gautham N, Rafi ZA (1992) Global search for optimal biomolecular structures using mutually orthogonal Latin squares. *Curr Sci* 63:560–564
- Goldberg DE (1999) *Genetic algorithms, in search, optimization and machine learning*. Pearson Education, Singapore
- Griffin JF, Langs DA, Smith GD, Blundell TL, Tickle IJ, Bedarkar S (1986) The crystal structures of [Met⁵] enkephalin and a third form of [Leu⁵] enkephalin: observations of a novel pleated β -sheet. *Proc Natl Acad Sci USA* 83:3272–3276
- Hansmann UHE, Okamoto Y, Onuchic JN (1999) The folding funnel landscape for the peptide Met-enkephalin. *Prot Struct Funct Genet* 34:472–483
- Isogai Y, Nemethy G, Scheraga HA (1977) Enkephalin: conformational analysis by means of empirical energy calculations. *Proc Natl Acad Sci USA* 74:414–418
- Ito K (1987) Latin squares. In: *Encyclopedic dictionary of mathematics*. MIT Press, Cambridge, MA, pp 891–892
- Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267:727–748
- Kanagasabai V (2008) Applications of mutually orthogonal latin squares in exploring the conformational space of peptides and proteins. Ph.D Thesis, University of Madras, Chennai
- Kanagasabai V, Gautham N (2008) An estimate of the numbers and density of low-energy structures (or decoys) in the conformational landscape of proteins. *PLoS ONE* 4(4):e5148
- Kanagasabai V, Arunachalam J, Arun Prasad P, Gautham N (2007) Exploring the conformational space of protein loops using a mean field technique with MOLS sampling. *Proteins* 67:908–921
- Koehl P, Delarue M (1994) Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol* 239:249–275
- Koehl P, Delarue M (1995) A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat Struct Biol* 2:163–170
- Koehl P, Delarue M (1996) Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol* 6:222–226
- Kříž Z, Carlsen PHJ, Koca J (2001) Conformational features of linear and cyclic enkephalins. A computational study. *J Mol Struct THEOCHEM* 540:231–250
- Kryshtafovych A, Krysko O, Daniluk P, Dmytriv Z, Fidelis K (2009) Protein structure prediction center in CASP8. *Proteins* 77:5–9

- Leszczynski JF, Rose GD (1986) Loops in globular proteins: a novel category of secondary structure. *Science* 234:849–855
- Levy Y, Becker OM (2001) Energy landscapes of conformationally constrained peptides. *J Chem Phys* 114:993–1009
- Lindman HR (1974) Analysis of variance in complex experimental designs. W. H. Freeman, San Francisco
- Liu Z, Dominy BN, Shakhnovich EI (2004) Structural mining: self-consistent design on flexible protein-peptide docking and transferable binding affinity potential. *J Am Chem Soc* 126:8515–8528
- MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kucsera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kucsera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102:3586–3616
- Makino S, Kuntz ID (1997) Automated flexible ligand docking method and its application for database search. *J Comput Chem* 18(14):1812–1825
- Mandl R (1985) Orthogonal Latin squares: an application of experimental design to compiler testing. *Commun ACM* 28:1054–1058
- Marcotte I, Separovic F, Auger M, Gagne SM (2004) A multidimensional ^1H NMR investigation of the conformation of methionine-enkephalin in fast-tumbling. *Biophys J* 86:1587–1600
- McKnight CJ, Matsudaira PT, Kim PS (1997) NMR structure of the 35-residue villin headpiece subdomain. *Nat Struct Biol* 4:180–184
- Mead R (1988) The design of experiments. Cambridge University Press, Cambridge
- Meinert CL (1986) Clinical trials: design, conduct and analysis. Oxford University Press, New York
- Montgomery DC (2000) Design and analysis of experiments, 5th edn. Wiley, New York
- Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* 19:1639–1662
- Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *J Phys Chem* 96:6472–6484
- Pearce SC (1983) The agricultural field experiment: a statistical examination of theory and practice. Wiley, New York
- Piela L, Kostrowicki J, Scheraga HA (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
- Raghavarao D (1971) Construction and combinatorial problems in design of experiments. Wiley, New York
- Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, Di Maio F, Lange O, Kinch L, Sheffler W, Kim B, Das R, Grishin NV, Baker D (2009) Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins* 77:89–99
- Ramya L, Gautham N (2009) Effects of hydration on the conformational energy landscape of the pentapeptide Met-enkephalin. *J Chem Theory Comput* 5:2180–2190
- Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 261:470–489
- Ring CS, Kneller DG, Langridge R, Cohen FE (1992) Taxonomy and conformational analysis of loops in proteins. *J Mol Biol* 224:685–699
- Ryser HJ (1963) Combinatorial mathematics. Mathematical Association of America, Washington, DC pp 79–84
- Schulze-Kremer S (2000) Genetic algorithms and protein folding. In: Webster D (ed) Methods in molecular biology, protein structure prediction: methods and protocols. Humana, New Jersey, pp 175–222, 143
- Sippl MJ (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5:229–235
- Sudarsanam S, DuBose RF, March CJ, Srinivasan S (1995) Modelling protein loops using a $\phi+1$, ψ dimer database. *Protein Sci* 4:1412–1420
- Sun S (1995) A genetic algorithm that seeks native states of proteins. *Biophys J* 69:340–355
- Tahirov TH, Sato K, Ichikawa-Iwata E, Sasaki M, Inoue-Bungo T, Shiina M, Kimura K, Takata S, Fujikawa A, Morii H, Kumasaka T, Yamamoto M, Ishii S, Ogata K (2002) Mechanism of c-Myb-C/EBP β cooperation from separated sites on a promoter. *Cell* 108:57–70
- Taylor RD, Jewsbury PJ, Essex JW (2002) A review of protein-small molecule docking methods. *J Comput Aided Mol Des* 16:151–166
- Terwilliger TC, Eisenberg D (1982) The structure of mellittin. I. Structure determination and partial refinement. *J Biol Chem* 257:6010–6015
- Unger R, Moulton J (1993) Genetic algorithms for protein folding simulations. *J Mol Biol* 231:75–81
- van Vlijmen HWT, Karplus M (1997) PDB based protein loop prediction: parameters for selection and methods for optimization. *J Mol Biol* 267:975–1001
- Vengadesan K (2004) Sampling the molecular potential energy surface using mutually orthogonal Latin squares and application to peptide sequences. Ph.D Thesis, University of Madras, Chennai
- Vengadesan K, Gautham N (2003) Enhanced sampling of the molecular potential energy surface using mutually orthogonal Latin squares: application to peptide structures. *Biophys J* 84:2897–2906
- Vengadesan K, Gautham N (2004a) Conformational studies on enkephalins using the MOLS technique. *Biopolymers* 74:476–494
- Vengadesan K, Gautham N (2004b) Energy landscape of Met-enkephalin and Leu-enkephalin drawn using mutually orthogonal Latin squares sampling. *J Phys Chem B* 108:11196–11205
- Viji SN, Arun Prasad P, Gautham N (2009) Protein-ligand docking using mutually orthogonal latin squares (MOLSDOCK). *J Chem Inf Model* 49:2687–2694
- Weiner SJ, Kollman PA, Nguyen DT, Case DA (1986) An all atom force field for simulations of proteins and nucleic acids. *J Comput Chem* 7:230–252
- Wojcik J, Mornon J, Chomilier J (1999) New efficient statistical sequence dependant structure prediction of short to medium sized protein loops based on an exhaustive loop classification. *J Mol Biol* 289:1469–1490
- Zhang J, Wang Q, Barz B, He Z, Kosztin I, Shang Y, Xu D (2010) MUFOLD: A new solution for protein 3D structure prediction. *Proteins* 78:1137–1152