Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Biochemical and Biophysical Research Communications 342 (2006) 424–433

BBRC

www.elsevier.com/locate/ybbrc

Protein structure prediction using mutually orthogonal Latin squares and a genetic algorithm

J. Arunachalam, V. Kanagasabai, N. Gautham *

Department of Crystallography and Biophysics, University of Madras, Chennai 600025, India

Received 23 January 2006
Available online 8 February 2006

Abstract

We combine a new, extremely fast technique to generate a library of low energy structures of an oligopeptide (by using mutually orthogonal Latin squares to sample its conformational space) with a genetic algorithm to predict protein structures. The protein sequence is divided into oligopeptides, and a structure library is generated for each. These libraries are used in a newly defined mutation operator that, together with variation, crossover, and diversity operators, is used in a modified genetic algorithm to make the prediction. Application to five small proteins has yielded near native structures.

Keywords: Protein structure prediction; Mutually orthogonal Latin squares; Genetic algorithm

Calculating the three-dimensional structure of a protein from its amino acid sequence remains a central problem in computational biology. The currently available protein structure prediction methods can be categorized into comparative modelling, fold recognition, and ab initio or new fold methods, based on their dependence on previously known structures [1]. Comparative modelling and fold recognition methods use the database of known structures to assist in determination of the structure [2,3]. The methods in the ab initio category aim to predict the correct structure as one with the minimum energy for a potential energy function. This category covers a wide range of methodologies, starting from approaches that introduce tertiary knowledge about the structures [4], to the use of secondary structure information [5,6] to methods that use only sequence information and a potential energy function that may be semi-empirical or derived from a database of known protein structures [7].

A purely physical approach, as in the case of ab initio methods, requires two important issues to be addressed. The first is the definition of a proper energy function or force field that accurately describes the intramolecular interactions as well as the intermolecular interactions with aqueous solvent that, together, stabilize the native folded conformation. Implicit in the choice of a force field is the choice also of an appropriate model for the protein. There are a wide variety of representations discussed in the literature, ranging from all-atom models, such as AMBER force field [8] ECEPP [9] and CHARMM [10], to much simpler coarse-grained models [11].

The second issue to be addressed in structure prediction is the method of searching the large and complex conformational space to arrive at the minimum energy structure, presumed to be the native fold. There are several methods that are reported in the literature [12–14] and they draw upon general global and local optimization techniques. Among these, a number of variants of genetic algorithms have been applied to the protein structure prediction problem. Genetic algorithms are general optimization procedures modelled on the process of natural evolution, with mutations, crossover and replication occurring on a population of strings [15]. 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

* Corresponding author. Fax: +91 44 22352494.
E-mail address: gautham@unom.ac.in (N. Gautham).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.01.162

iterated until the population converges on a single individual with the optimum fitness. Unger and Moult [16] have shown that genetic algorithms perform better than Monte Carlo methods for finding the global minimum energy of simple two-dimensional lattice protein models. Sun [17] has used a genetic algorithm that, with the help of statistical potential and restraints like known native radius of gyration and disulphide bonds (if any), predicts native-like structure. The algorithm of Bowie and Eisenberg [18] uses nine residue short fragment structures and 15–25 residue larger fragment structures from the database of known protein structures for protein structure prediction. Dandekar and Argos [4,19] have used a genetic algorithm that allows only seven possible conformations for each residue, together with appropriate fitness functions to predict the structures of a variety of different classes of proteins.

In this paper, we describe our attempt to overcome the conformational search problem by combining two search techniques, namely, the method of mutually orthogonal Latin squares (MOLS) and a genetic algorithm. The MOLS algorithm was developed in our laboratory to search conformation space exhaustively and build a library of possible low energy local structures for oligopeptides [20]. In the present application, we first divide the protein sequence into short overlapping fragments and then use the MOLS method to build their structural libraries. Next we use a genetic algorithm that exploits the libraries of fragment structures and predicts a single best structure for the protein sequence. In the application of this combined method to some small test proteins, it has predicted their near native structures.

Materials and methods

Since this is the first application of the hybrid MOLS-genetic algorithm method, we restricted our attempts to the following four small α helical proteins, and to one β sheet protein: avian pancreatic polypeptide (APP), villin headpiece (VHP) mellitin (MEL), cMYB (MYB), and tryptophan zipper (TZ) These proteins were chosen because of their small size, well-defined secondary structures, and absence of disulphide bonds. In each case, an all-atom model was used, keeping bond lengths and bond angles fixed at their standard values [21]. The search was therefore carried out in torsion angle space, including the backbone and side chain torsion angles. The conformation of the protein chain was thus specified by n torsion angles θ_r, $r = 1, n$, and the correct structure of the molecule is defined by that set of θ_r that yields the minimum of $V(\theta_r)$ over the entire space, where V is a suitable potential energy function. This potential energy function, or objective function, for minimization was chosen differently at each stage, as explained below. Fig. 1 gives the flowchart of the algorithm. The method operates in two phases.

Phase1: building fragment libraries using MOLS. In phase 1, the sequence was divided into overlapping fragments of nine residues each. A structure library for each fragment was created using the method of MOLS. Thus, for example, in the case of avian pancreatic polypeptide (APP), with a sequence of length 36, there were 28 overlapping nonamers and therefore 28 structure libraries. The MOLS method is explained in full elsewhere [20]. Here we give a very brief summary. The MOLS method, which is a variant of the mean field technique [22], uses mutually orthogonal Latin squares to select n^2 structures from the multi-dimensional conformational space of size m^n, where n is the number of dimensions (i.e., the number of torsion angles) and m specifies the fineness of the search grid. (The step size was taken to be 10° in the present case.) The energy values corresponding to these n^2 structures are calculated and analysed to find the set of torsion angles that specify one low energy structure. This low energy structure is subjected to a few cycles of conjugate gradient minimization [23]. The calculations and analyses are repeated for another set of n^2 structures, chosen using another set of MOLS, to identify another low energy structure. This procedure may be repeated a large number of times, each time identifying one low energy structure. Initial trials on all the nine residue fragments of APP showed that 500 repetitions were sufficient to identify all the relevant low energy structures. Based on this we generated 500 low energy structures for each overlapping nonamer of all the targets. The number of structures in the final library for each peptide was further reduced by clustering together similar structures in the generated set, using the hierarchical clustering algorithm [24] with a cutoff of 1.0 Å RMSD between the backbone atoms. The lowest energy structure in each cluster was chosen as its representative. At the end of this procedure, we obtained libraries of mutually dissimilar structures with 123–327 members for each sequence fragment. Table 1 gives details of structure libraries generated. Table 2 lists the RMSD values (using backbone atoms) of the best structure in each library with the respective experimental structure. It also gives the energy (calculated as described below) of the best structure. Clearly, each library contains at least one structure that is close to the experimental structure and has a low energy value.

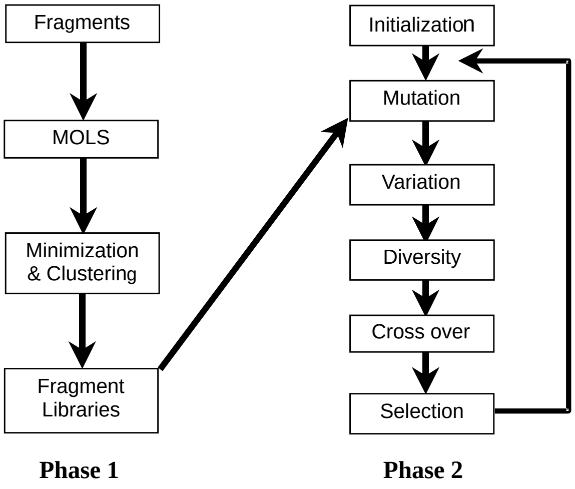

Fig. 1. Flowchart explaining the complete algorithm.

The potential function was chosen based on previous studies on generating structure libraries using the MOLS method [20]. Since the calculations were carried out in torsion angle space, the potential function used included only the electrostatic, van der Waals, and hydrogen bond energy terms from the AMBER force field [25]. There were no bond length, bond angle or torsional energy terms. Besides this, we also included a secondary structure biasing function as described by Crivelli et al. [6]. This function was calculated for each residue in the sequence based on the secondary structure predicted for it by the PHD server [26] accessed through the web site <http://npsa-pbil.ibcp.fr/>. The function,

$$E_{\phi\psi} = \sum_{\text{dihedrals}} k_\phi[1 - \cos(\phi - \phi_o)] + k_\psi[1 - \cos(\psi - \psi_o)]$$

biases the backbone torsion angles of the amino acids of a residue predicted to be α helix or β sheet to be close to their respective ideal values. Here k_ϕ and k_ψ are force constants related to the strength of the secondary structure prediction from the prediction server. ϕ_o and ψ_o are ideal dihedral values of the predicted secondary structure. The secondary structure biasing function is a soft constraint. When used in the MOLS algorithm to create fragment libraries, it also identifies other energetically favourable structures besides the predicted secondary structure. Using the biasing function is therefore a way of ensuring the presence of the predicted

Table 1
Some details of the MOLS-generated structure libraries and the genetic algorithm results

Molecule	Sequence length	Number of torsion angles	Number of fragment libraries	Number of structures in the libraries			Number of generations in genetic algorithm	Average energy value of the final population (kcal/mol)	RMSD of the final structure with the experimental structure (Å)
				Min	Max	Ave			
APP	36	130	28	143	351	243	250	−163.26	4.0
VHP	36	145	28	153	288	176	250	−140.65	5.2
MEL	26	98	18	123	232	182	250	−119.62	4.3
MYB	52	232	44	136	387	213	3000	−211.74	6.1
TZ	16	59	8	151	312	164	1500	−1136.31	1.6

Table 2
Comparison of best structure in the MOLS generated library with their respective experimental structure fragments

Sequence fragment no.	APP		VHP		MEL		MYB		TZ	
	RMSD	Energy	RMSD	Energy	RMSD	Energy	RMSD	Energy	RMSD	Energy
1	1.55	−27.47	1.29	−59.04	0.15	−52.92	1.48	261.81	1.15	−45.71
2	1.34	−36.32	0.92	−64.86	0.22	−46.20	1.46	−47.98	1.74	−49.68
3	1.27	−36.06	0.82	−61.94	0.23	−52.70	1.40	−34.33	1.57	−48.38
4	1.45	−36.37	0.72	5.82	0.42	2.05	1.40	−53.23	1.72	−44.71
5	1.33	−33.59	1.20	45.66	0.84	−51.22	1.54	−60.07	1.87	−45.85
6	1.85	−28.51	1.11	114.04	0.88	−45.36	0.55	−76.96	1.64	−39.03
7	1.61	−36.83	1.38	234.00	0.88	−33.20	0.77	−70.86	1.48	−23.44
8	1.60	−45.38	1.79	370.08	1.37	−37.17	0.12	−77.34	1.21	−42.03
9	1.40	−38.90	1.92	301.61	0.85	−17.08	0.18	−76.91		
10	0.96	−54.25	1.52	306.58	0.78	−15.87	0.18	−70.64		
11	0.44	−57.21	1.85	137.52	1.54	−28.36	0.23	−76.30		
12	0.47	−64.29	1.39	15.62	0.68	−43.49	0.23	−74.48		
13	0.20	−70.71	0.89	3.22	0.29	−52.69	0.30	−73.35		
14	0.17	−60.71	0.83	−52.91	0.23	−33.76	0.74	−60.95		
15	0.14	−58.60	1.57	−37.87	0.21	−55.52	0.98	−58.36		
16	0.24	−66.54	1.77	10.02	0.26	−65.06	0.84	−55.45		
17	0.21	−56.82	1.88	12.32	0.16	−79.40	1.41	−53.67		
18	0.24	−49.84	1.75	296.49	0.28	−78.11	1.81	−52.73		
19	0.22	−43.97	1.27	165.30			1.29	−50.05		
20	0.24	−33.51	1.34	55.36			1.20	−48.14		
21	0.24	−63.08	0.89	29.17			1.68	−34.00		
22	0.18	−65.80	0.37	−79.65			1.30	132.90		
23	0.25	−64.49	0.39	−73.68			1.36	120.91		
24	0.23	−63.87	0.44	−85.56			0.75	127.66		
25	0.73	−70.67	0.45	−80.02			0.64	−52.14		
26	0.85	−65.60	0.89	−72.20			1.01	−35.55		
27	0.77	−55.21	1.25	−63.81			1.06	−48.07		
28	0.94	−56.12	1.18	−59.41			1.16	−32.49		
29							1.53	−50.57		
30							1.80	−32.58		
31							1.44	−39.46		
32							1.61	−53.03		
33							1.76	−49.67		
34							1.38	−41.06		
35							1.09	−58.75		
36							1.25	−68.20		
37							0.51	−76.20		
38							0.24	−71.93		
39							0.44	−83.48		
40							0.65	−81.95		
41							0.58	−70.08		
42							0.87	15.56		
43							0.94	−65.22		
44							1.18	853.26		

RMSD values in Å and energy in kcal/mol.

secondary structure in the fragment library, but not as the sole structure. The time taken to generate a single library varied from 8 to 51 h of CPU time, depending on the sequence, on a single Intel Pentium 4 processor (1.8 GHz) with Red Hat Linux 7.3 as operating system. The wide variation in the running time is chiefly a result of the large difference in time required for the gradient minimization procedure.

Phase 2: the genetic algorithm. Phase 2 of the method uses a genetic algorithm, in combination with the library of structures generated in phase 1, to predict the three-dimensional structure of the protein. The genetic algorithm used in the present application is similar to that of Schulze-Kremer [27]. The protein structure is modelled as a string of torsion angles. An initial random population of ten individual structures is generated. Each new generation of structures is obtained by applying four operators on this population. Unlike other applications of genetic algorithms [4], the operators are configured to operate on numbers (the torsion angles) and not on bit strings. The operators are MUTATE, VARIATE, CROSSOVER, and a newly defined DIVERSITY operator.

In the genetic algorithm as described by Schulze-Kremer [27], the MUTATE operator, when activated for a particular residue, replaces the value of the torsion angles of this residue by a random choice from the ten most frequently occurring values for that residue. The database of most frequently occurring values for all the backbone and side chain torsions was calculated from the structures in the PDB. We have modified this operator as follows. First, the database now consists of the libraries of MOLS structures generated in phase 1. Second, the operator works not on a single residue, but by replacing a randomly selected nine residue fragment by a structure from the respective MOLS-generated structure library. The probability of picking up a particular MOLS-generated structure depends on its energy value, one with lower energy being more probable. Based on initial trial runs, the probability parameter governing the MUTATE operator is set to the constant value of 0.10, leading to a 10% chance for every overlapping nonamer fragment in a molecule to be mutated in each generation.

The VARIATE operator increments or decrements the chosen torsion angle by 1°, 5° or 10°. The initial value of the probability parameter governing VARIATE is set as 0.2. A random number between 0 and 1 is generated for every torsion angle. If the generated random number is less than 0.2, then VARIATE is applied. This probability parameter is dynamic and varies from 0.2 at the beginning of the algorithm to 0.7 at the end. Three different sets of dynamic parameters control the choice between 1°, 5° or 10°. The value of 10° is chosen with a probability of 0.6 at the beginning of the run and 0.0 at the end, for 5° the probabilities are 0.3 and 0.2, respectively, while for 1° they are 0.1 and 0.8, respectively. Another parameter (value 0.5) controls whether the angle is incremented or decremented.

The CROSSOVER operator has two components: the two-point crossover and the uniform crossover. Every time this operation is carried out the individuals in the population are randomly paired. Depending on the CROSSOVER probability, which has a value of 0.7 at the first generation and 0.1 at the last one, each pair is selected or not selected for the CROSSOVER operation. For each selected pair a decision is made between two-point crossover and uniform crossover. This decision is based on the following dynamic parameters. Two-point crossover probability is 0.1 at the first iteration, and 0.9 at the last one, uniform crossover probability is 0.9 and 0.1, respectively. In two-point crossover, two points are randomly selected in the sequence, and the fragment of the structure that lies between these two points are exchanged between the two structures in the selected pair. The uniform crossover operator works on every residue in the sequence. With a probability of 0.5 it exchanges the torsion angle values of the residue with those of its counterpart in the other structure of the pair.

In addition to the three operators above, we introduced a 'DIVERSITY' operator. This operator randomly selects a residue and sets the backbone torsion angles of the next five residues to 180°. The DIVERSITY operator helped avoid premature convergence to a local minimum, in particular during the initial stages of the algorithm. The probability of choosing a given residue was kept constant throughout the run at 0.01.

The initial (zeroth) generation consisted of a population of 100 individuals, with all their variable torsion angles set to 180°. The four operators were applied to all the individuals of a generation (i.e., the 'parent' generation), and the individuals for the next generation (i.e., the 'offspring' generation) were selected by the elitist replacement method [15,27]. In this method, the parent and the offspring generations were merged, and all 200 individuals were sorted on the basis of the fitness function, calculated for each structure as described below. The 100 fittest individuals were selected to constitute the next generation. The cycles of variation and selection were continued until convergence. The criterion for convergence was that all the structures in the final population were similar. This was checked by clustering the population of each generation using a hierarchical clustering algorithm [24] with 3 Å backbone RMSD cutoff. When the clustering routine returned a single cluster with all 100 structures in it, the algorithm was considered to have converged. In all the five attempts, the algorithm converges towards a single structure. The protein MYB with 52 residues in the sequence converged only after nearly 3000 generations, while TZ (16 residues) required 1500 generations to converge. The other three converged in less than 600 generations. The total computation time for phase 2 varied from 1 to 17 h on a Pentium 4 processor, depending on the size of the molecules and the number of generations. Table 1 gives some details regarding this procedure. The structures in the final set were all further subjected to conjugate gradient energy minimization using the program Discover in the Biosym suite [21]. The structure finally used in the subsequent discussion for each target is the one with the minimum energy (as defined by the Discover energy function) in the final, minimized population.

The fitness function used to select the offspring in each generation was the potential function used above to generate the MOLS libraries, but without the secondary structure biasing term. Instead, we included a 'pseudo entropy' term as defined by Schulze-Kremer [27] E_{pe}, calculated as follows:

$$E_{\mathrm{pe}} = 4^{\Delta D},$$

where ΔD is the measured diameter of the structure minus the expected diameter. If the actual diameter is lesser than the expected diameter, then E_{pe} is set to zero. The diameter was measured as the largest distance between any two C^{α} atoms. The expected diameter was calculated as $8 \times (\text{length of the sequence in residues})^{1/3}$. In the case of the tryptophan zipper (TZ), the experimental structure consists of two strands forming a β hairpin. In this case alone, we found it necessary to include the secondary structure biasing term, as well as a rule-based β-strand pairing potential, as described by Kesar and Levit [28]. This is a cooperative term that operates on pairs of hydrogen bonds. A single hydrogen bond does not contribute to the energy, but it favours the formation of other hydrogen bonds belonging to the same pair of strands, which then gives rise to the formation of yet others.

Results

All the targets considered here, except tryptophan zipper (TZ), are α helical proteins. TZ alone contains an antiparallel β sheet. All five targets converge in the reasonable amount of computation time to a single, near native conformation. Avian pancreatic polypeptide (APP), villin head piece (VHP), and mellittin (MEL) converged to a single structure in 250 generations in the genetic algorithm. The RMSD of the best structure in the final population for each, when the backbone atoms are superposed on the respective experimental structures [29–31], is 4.0, 5.2 and 4.3 Å, respectively. The c-Myb protein (MYB) required 3000 generations to converge, and the best structure had a backbone RMSD of 6.1 Å with respect to the experimental structure [32]. TZ converged after 1500 generations to a

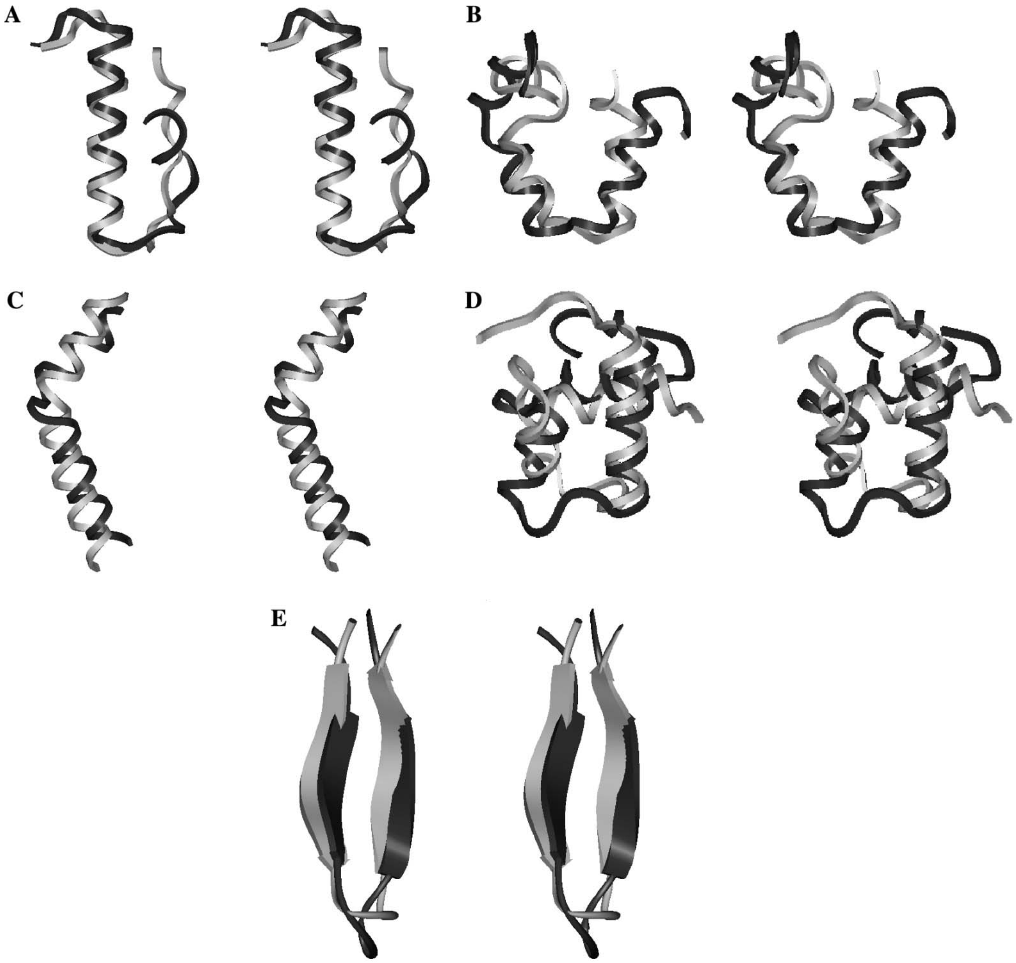

Fig. 2. Stereo view of the predicted structure (dark) superposed on the experimental structure (light) of APP (A), VHP (B), MEL (C), MYB (D), and TZ (E).

single structure with a backbone RMSD of 1.6 Å, when superposed on the experimental structure [33]. The final predicted structures of all five targets, together with the experimental structures, are shown in Fig. 2. We now discuss the results for each of the five structures individually.

Avian pancreatic polypeptide

APP is a 36 residue protein, whose crystal structure (PDB ID: 1PPT) was solved by Blundell et al. [29] at 1.4 Å resolution. The structure consists of a polyproline-type helix from residues 1 to 8 and an α-helix from residues 13 to 31. A turn at residues 9–13 arranges the polyproline helix approximately parallel to the α-helix. Another turn at residues 33 and 34 allows the carboxy terminus to be oriented away from the α-helix. The MOLS library of fragments corresponding to the regions containing α-helix of experimental structure was predicted accurately, and the secondary structure biasing term makes it the lowest energy structure among all the MOLS generated structures. The best predicted structures of the nonamer fragments from residue 10 to 28, which have this secondary structure either completely or partly, had an RMSD below 1 Å with the experimental structure (Table 2). The lowest RMSD was 0.14 Å with energy of −58.6 kcal/mol in the α helical region and 1.3 Å with an energy of −33.5 kcal/mol in the other regions of the sequence. The final structure predicted by the combined MOLS-GA algorithm for this protein has the α helix from residues 14 to 31 and the turn that follows it (Fig. 2A). The polyproline-like helix does not appear in the final prediction though it is one of the low energy structures in the MOLS library. Instead, the predicted structure has possible beta turns with backbone hydrogen bonds between residues G1 and Q4, S3 and T6, and Q4 and Y7. Besides this, the predicted structure has three γ turns with hydrogen bonds between residues Q4 and P2, R33 and V31, and Y36 and H34, not found in the crystal structure. In the experimental structure, the hydrophobic core of APP is formed by the three proline residues (P2, P5, and P8) in the polyproline helix along with F20 and Y27 in the α helix. The predicted structure does not contain this hydrophobic core. Instead, the only contacts between the N-terminal segment and the α helix are the hydrogen bonds between the side chain of R19 with the main chain oxygen of Q4 and Y7 and another hydrogen bond between N23 and main chain oxygen of G1. Though the overall RMSD between the backbone atoms of the final predicted structure and the experimental structure is 4.0 Å, portions of the structure are predicted with much better accuracy (Table 3). As expected, the prediction is the best in the α helical region.

APP has been used earlier as the target protein for structure prediction. Liwo et al. [34] have predicted the structure of APP to an accuracy of 3.8 Å in a two-step algorithm. In this prediction, too, the secondary structure regions are predicted more accurately than the rest of the sequence. Sun has used a reduced representation model and a genetic algorithm to predict the structure of APP to an accuracy of 3.9 Å RMSD [35]. Their genetic algorithm uses conformational dictionaries generated from non-homologous structures from the PDB to perform the MUTATION operation.

Table 3
Comparison of nonamer fragments of final predicted structures with the experimental structure fragments

	APP	VHP	MEL	MYB	TZ
1	3.51	3.02	0.26	3.47	2.12
2	3.39	2.26	0.25	2.51	1.27
3	4.25	1.51	0.29	2.45	1.77
4	4.39	1.73	0.26	2.56	1.65
5	3.31	1.94	1.80	2.20	1.65
6	3.07	2.50	2.88	1.48	1.85
7	3.27	3.74	2.83	0.64	1.95
8	3.04	3.76	2.95	0.53	2.14
9	2.90	3.15	2.84	0.47	
10	2.92	2.93	2.37	0.48	
11	2.86	3.33	2.44	0.49	
12	1.96	3.07	1.77	0.48	
13	0.43	1.85	0.34	0.31	
14	0.39	1.75	0.33	1.62	
15	0.26	2.15	0.21	1.67	
16	0.24	2.64	0.18	1.95	
17	0.27	2.64	0.27	3.2	
18	0.34	2.62	0.33	3.72	
19	0.35	2.99		3.45	
20	0.33	2.78		3.52	
21	0.33	2.21		3.26	
22	0.42	0.80		2.33	
23	0.57	0.86		2.37	
24	1.08	0.76		1.76	
25	1.40	0.78		0.87	
26	1.88	1.55		0.99	
27	2.37	1.68		1.37	
28	2.43	2.81		1.93	
29				2.44	
30				3.11	
31				3.12	
32				2.91	
33				2.68	
34				3.24	
35				3.68	
36				2.89	
37				2.21	
38				1.37	
39				0.74	
40				0.96	
41				2.24	
42				2.95	
43				3.20	
44				3.66	

The table gives the RMSD in Å, when the fragments are superposed.

Villin head piece

The NMR structure of VHP (PDB ID: 1VII) contains three short helices (residues 4–8, 15–18, and 23–30), which are held together by a loop and a turn [30]. The current method predicts all three helices, packed in a native-like conformation (Fig. 2B). Besides the helices, the predicted structure has a β turn with a possible hydrogen bond between residues F36 and K33, and two γ turns with hydrogen bonds between residues M13 and F11, and L35 and K33, which are absent in the crystal structure. The secondary structure predicted by PHD, which was used in generating the MOLS library structures, predicts the secondary structure of VHP to be a single long helix starting from residue 4 to residue 32. The computed preference of the sequence for this secondary structure is reflected weakly in the fact that the helices are longer, from residues 4 to 10, 13 to 19, and 22 to 32 in the predicted structure, as compared to the helices in the experimental structure. The three short helices in the structure surround a hydrophobic core consisting of three phenylalanine residues F7, F11, and F18. A one microsecond molecular dynamics simulation of VHP with explicit water model [36] yielded a stable cluster of structures that had an RMSD of 4.5 Å RMSD when superposed on the experimental structure. In this simulation, the hydrophobic core formed simultaneously with the secondary structures during the very early stage of the folding process. In the structure predicted by the MOLS-GA algorithm however, the native-like fold occurs without this hydrophobic core (Fig. 3). Only F7 and F11 with a distance between the aromatic rings being 5.0 Å may be considered an interacting pair, according to the criterion of Burley and Petsko [37], who showed that phenylalanine rings in proteins could be considered to be involved in an aromatic interaction when they are within 7 Å distance of each other. Also, by creating single and double mutants of the three phenylalanine residues F7, F11, and F18, Frank et al. [38] have shown that VHP can attain a native structure even with out these specific aromatic interactions. In our calculations, we have used no explicit terms for the hydrophobic force solvent interactions, either in generating the MOLS libraries or in the genetic algorithm. Thus, local structural preferences appear to contribute substantially to the final structure of the protein, not just in the secondary structural regions, but also in the loop regions. As shown in Table 2, in the loop regions also, the best structure in the MOLS library for each nonamer had low RMSD with the experimental structure, the lowest being 1.1 Å RMSD.

Mellitin

The crystal structure of MEL (PDB ID: 1MLT) has an α-helix from residues 1 to 10, followed by a β-turn, followed

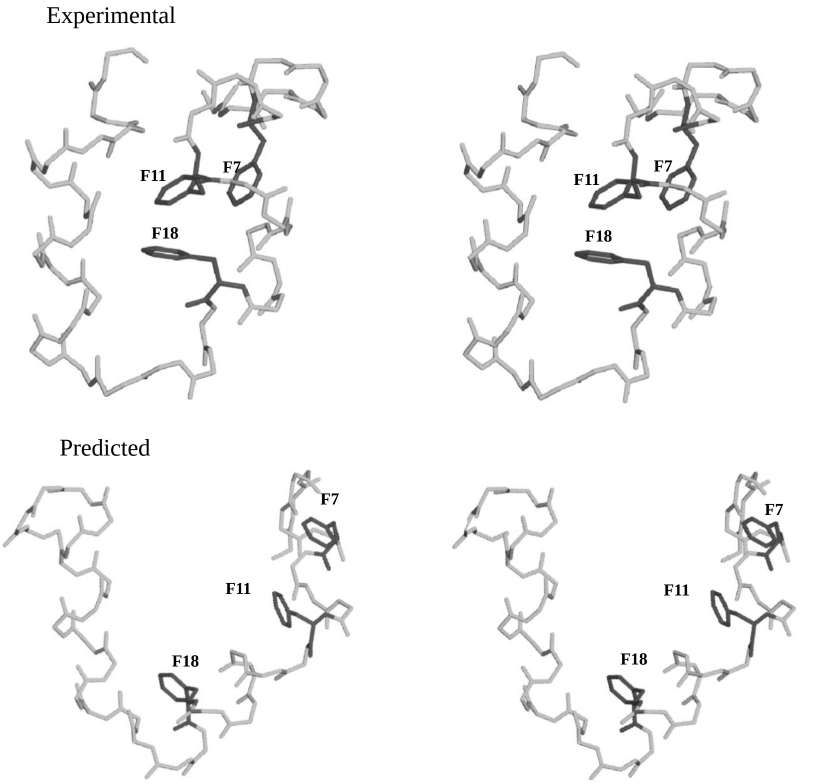

Fig. 3. Experimental and predicted structures of VHP showing the three phenylalanine residues in the hydrophobic core.

by a longer helix from residues 13 to 26. The present algorithm predicts both the helices (from residues 1 to 12 and 14 to 26), as well as the bend (Fig. 2C, overall RMSD 4.3 Å). However, in the predicted structure the bend is formed by a γ-turn between residues 11 and 13. This γ-turn and shortening of the second helix by one residue together makes the predicted structure deviate from the experimental structure. Once again the secondary structures are predicted with a higher accuracy. The MOLS-generated library for the nonamer sequence containing the bend region (residues 12–14) has a structure with an RMSD of 1.3 Å and an energy value of −37.2 kcal/mol (Fig. 4). However, this structure was not selected by the genetic algorithm and does not appear in the final prediction. Table 2 shows that all the nonamer fragments have a structure within 1.5 Å of the respective experimental structure. However, Table 3, which compares the overlapping nonamers of the final predicted structure with the crystal structure fragments, shows that the genetic algorithm apparently failed to retain these local structures in the final prediction.

c-MYB

c-MYB protein (PDB ID: 1GV5) is a 52 residue protein with three α-helices, (residues 7–20, 25–30, and 47–46) packed together mainly by hydrophobic interactions in the crystal structure. There are two specific hydrogen bonds between two of the helices. The predicated structure (overall RMSD 6.2Å, Fig. 2D) retains all these features,

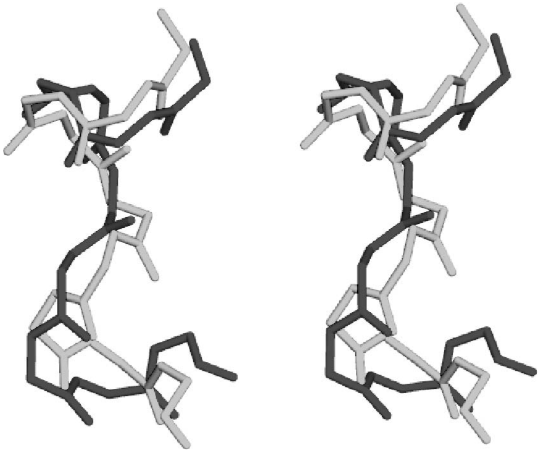

Fig. 4. Stereo view of the superposition of best MOLS prediction (dark) with the crystal structure (light) of the bend region of mellitin (RMSD = 1.3 Å).

except the pair of hydrogen bonds. On comparing the MOLS generated library structures with the crystal structures (Table 2), the helix regions were well predicted. However, in the final prediction (Table 3), while the helix regions remained correct, the structures of the non-helical regions moved away from the experimental structure. This is because of the predicted presence of β-turns between the sets of residues (L1–G4), (I2–P5), (P5–K8), (G35–G38), (G38–C41), and (H48–P51), and γ turns between the triples of residues (G4–W6), (W6–K8), (R36–G38), (G38–Q40), and (N50–E52). All these turns are absent in the experimental structure. Thus, even though the native fold was obtained, the apparent overemphasis on local structure in the prediction leads to its significant deviation from the experimental structure.

Tryptophan zipper

Tryptophan zippers are short structural motifs that form β hairpins. β-Strand structures are notoriously hard to predict and need special functions that help in strand-pairing [39,40]. The 16-residue long tryptophan zipper has been designed to form a β hairpin with two strands and a type IV β turn in between [33]. Initial attempts using the current algorithm unchanged did not produce any folded structures. An analysis of the MOLS library of structures shows that, of the structures for the bend region (residues 10–13) in the library, the best had 1.4 Å RMSD with the experimental structure, but the genetic algorithm did not retain this. In order to drive the structure towards a β sheet structure, the fitness function used in phase 2 of the algorithm was modified to include the secondary structure bias term as well as a special rule-based β pairing potential[28], as described in 'Materials and methods' section. The structure is well predicted after the above inclusion (overall RMSD 1.6 Å, Fig. 2E). The bend in the structure due to the formation of the type IV β-turn between residues 10 and 13 is also predicted exactly. Besides this the predicted structure also has a possible β turn comprising residues (A8–T11). In the experimental structure, the β hairpin motif is stabilized by the aromatic interactions between the tryptophan residues present at both the strands. Such aromatic interactions between the tryptophan residues are not present in our predicted structure. It appears therefore that in this case, the β-sheet pairing function is sufficient to ensure the proper fold. The psuedoentropy term appears ineffective in this case.

Discussion

The algorithm presented here is a combination of a fragment structure generation method, coupled to a genetic algorithm that puts the fragment structures together to arrive at the best structure. The use of fragment libraries in protein structure prediction is not new, and several successful algorithms have incorporated this technique [12,41–43]. The present algorithm is however unique in that it does not obtain any of the fragment structures from previously known structure databases. The secondary structure biasing term is the only part of the potential that is based on previously known structures—the PHD algorithm that we use to predict the secondary structure is based on them [26]. The rest of the potential we have used consists of well-parameterized atom–atom semi-empirical pair potential functions and is not sequence-specific. Further, during phase 2, i.e., the genetic algorithm, a simple fitness function based again on the AMBER [25] potential field was used. The all-β TZ alone required additional β-strand pairing potential terms. Thus, a larger sequence space is accessible to this algorithm than to others. Another advantage of the present method is that it converges towards a single structure, thereby making it unnecessary to search for the best structure from among a myriad of other energetically favourable structures, as in most other prediction algorithms.

In all the five test cases, we noted that the local interactions are predominantly favoured in the energy function we have used. As described above, several γ turns were predicted (Table 4). However, none of the targets have γ turns in the experimental structures. The exception is MYB, which has one γ turn with hydrogen bond between K23 and G21. The overemphasis on local interactions appears to be an important cause of the errors in prediction, even though the overall fold and orientations of the secondary structures are correctly predicted.

The algorithm appears to efficiently arrive at the minimum of the given fitness function, though this function appears to be only an approximate model of the interaction in the peptides. This is clear when we note that the value of this fitness function is always less for the predicted structures than for the experimental structures (Table 5). The

Table 5
Comparison of the value of the fitness function between the experimentally determined structures and the final predicted structures

Molecule	Values of the fitness function (kcal/mol)	
	Predicted structure	Experimental structure
APP	−177.37	129.53
VHP	−40.17	630.08
MEL	−85.36	459.47
MYB	−246.23	752.91
TZ	−1138.31	−1087.15

Table 4
The list of additional β and γ turns in the predicted structures

APP		VHP		MEL		TZ		MYB	
β-Turns	γ-Turns	β-Turns	γ-Turns	β-Turns	γ-Turns	β-Turns	γ-Turns	β-Turns	γ-Turns
G1–Q4	P2–Q4	K33–F36	F11–M13		T11–L13	D7–K10		L1–G4	G4–W6
S3–T6	V31–33R		K33–L35			A8–T11		I2–P5	W6–K8
Q4–Y7	34H–36Y							P5–K8	R36–G38
T32–R35								G35–G38	G38–Q40
								G38–C41	N50–E52
								H48–P51	

side chain interactions are also not modelled very well by the fitness function, and in all five targets, the predicted side chain–side chain interactions do not match the experimentally determined interactions. Improvement of the fitness function to address these issues, for example, by inclusion of hydrophobic and solvent terms [44] other than the psuedoentropy term, and specific side chain interactions [45], as well as statistical 'contact' potentials that model side chain interactions [46], may improve the performance of the algorithm, both in terms of more accurate predictions, as well as in terms of application to larger sequences.

Acknowledgments

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

References

[1] J. Moult, T. Hubbard, K. Fidelis, J.T. Pederson, Critical assessments of methods of protein structure prediction CASP: round III, Proteins: Struct. Funct. Genet. [Suppl]. 3 (1999) 2–6.

[2] T.L. Blundell, B.L. Sibanda, M.J.E. Stenberg, J.M. Thornton, Knowledge based prediction of protein structures and design of novel molecules, Nature 326 (1987) 347–352.

[3] J. Bajorath, R. Stenkamp, A. Aurofo, Knowledge based model building of proteins: concepts and examples, Protein Sci. 2 (1993) 1798–1810.

[4] T. Dandekar, P. Argos, Folding the main chain of small proteins with the genetic algorithm, J. Mol. Biol. 236 (1994) 844–861.

[5] M. Nanias, M. Chinchio, J. Pillardy, D.R. Ripoll, H.A. Sheraga, Packing helices in a protein by global optimization of a potential energy function, Proc. Natl. Acad. Sci. USA 100 (2003) 1706–1710.

[6] S. Crivelli, E. Eskow, B. Bader, V. Lamberti, R. Byrd, R. Schnabel, T. Head-Gordon, A physical approach to protein structure prediction, Biophys. J. 82 (2002) 36–49.

[7] K.A. Olszewski, L. Piela, H.A. Scheraga, Mean field theory as a tool for intramolecular conformational optimization. 3. Test on Mellitin, J. Phys. Chem. 97 (1993) 267–270.

[8] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, J. Am. Chem. Soc. 117 (1995) 5179–5197.

[9] G. Nemethy, K.D. Gibson, K.A. Palmer, C.N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, H.A. Scheraga, 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 (1992) 6472–6484.

[10] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan, M. Karplus, CHARMM: a program for macromolecular energy, minimization and dynamics calculations, J. Comp. Chem. 4 (1983) 187–217.

[11] A. Colubri, Prediction of protein structure by simulating coarse-grained folding pathways: a preliminary report, J. Biomol. Struct. Dyn. 5 (2004) 625–638.

[12] K.T. Simon, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring function, J. Mol. Biol. 268 (1997) 209–225.

[13] A. Kolinski, Protein modeling and structure prediction with a reduced representation, Acta Biochim. Pol. 51 (2004) 349–371.

[14] D. Kihara, H. Lu, A. Kolinski, J. Skolnick, TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints, Proc. Natl. Acad. Sci. USA 98 (2001) 10125–10130.

[15] D.E. Goldberg, Genetic Algorithms, in Search, Optimization and Machine Learning, Pearson Education, Singapore, 1999.

[16] R. Unger, J. Moult, Genetic algorithms for protein folding simulations, J. Mol. Biol. 231 (1993) 75–81.

[17] S. Sun, A genetic algorithm that seeks native states of proteins, Biophys. J. 69 (1995) 340–355.

[18] J.U. Bowie, D. Eisenberg, 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 (1994) 4436–4440.

[19] T. Dandekar, P. Argos, 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 (1996) 645–660.

[20] K. Vengadesan, N. Gautham, Enhanced sampling of the molecular potential energy surface using mutually orthogonal Latin squares: application to peptide structures, Biophys. J. 84 (2003) 2897–2906.

[21] Biosym/MSI Release 95.0, San Diego, CA 92121-3752, USA, 1995.

[22] K.A. Olszewski, L. Piela, H.A. Scheraga, Meanfield theory as a tool for intramolecular conformational optimization 1. Tests on terminally blocked alanine and met-enkephalin, J. Phys. Chem. 96 (1992) 4672–4676.

[23] W.H. Press, S.A. Tukolsky, W.T. Veterling, B.P. Flanery, Numerical Recipies in Fortran—The Art of Scientific Computing, second ed., Cambridge University Press, 1992, pp. 413–418.

[24] Z. Kriz, P.H.J. Carlsen, J. Koca, Conformational features of linear and cyclic enkephalins. A computational study, J. Mol. Struct. (THEOCHEM) 540 (2001) 231–250.

[25] S.J. Weiner, P.A. Kollman, D.T. Nguen, D.A. Case, An all atom force field for simulations of proteins and nucleic acids, J. Comp. Chem. 7 (1986) 230–252.

[26] B. Rost, C. Sander, Prediction of protein secondary structure with better than 70% accuracy, J. Mol. Biol. 232 (1993) 584–599.

[27] S. Schulze-Kremer, Genetic algorithms and protein folding, in: D. Webster (Ed.), Methods in Molecular Biology, Protein Structure Prediction: Methods and Protocols, 143, Humana Press, New Jersey, 2000, pp. 175–222.

[28] C. Kesar, M. Levit, A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics, J. Mol. Biol. 329 (2003) 159–174.

[29] T.L. Blundell, J.E. Pitts, I.J. Tickle, S.P. Wood, C.W. Wu, X-ray analysis (1.4 Å resolution) of avian pancreatic polypeptide: small globular protein harmone, Proc. Natl. Acad. Sci. USA 78 (1981) 4175–4179.

[30] C.J. McKnight, P.T. Matsudaira, P.S. Kim, NMR structure of the 35-residue villin headpiece sub domain, Nat. Struct. Biol. 4 (1997) 180–184.

[31] T.C. Terwilliger, D. Eisenberg, The structure of mellitin. Structure determination and partial refinement, J. Biol. Chem. 257 (1982) 6010–6015.

[32] T.H. Tahirov, K. Sato, E. Ichikawa-Iwata, M. Sasaki, T. Inoue-Bungo, M. Shiina, K. Kimura, S. Takata, A. Fujikawa, H. Morii, T. Kumasaka, M. Yamamoto, S. Ishii, K. Ogata, Mechanism of c-Myb-C/EBP beta cooperation from separated sites on a promoter, Cell 108 (2002) 57–70.

[33] A.G. Cochran, N.J. Skelton, M.A. Starovasnik, Tryptophan zippers: stable, monomeric β-hairpins, Proc. Natl. Acad. Sci. USA 98 (2001) 5578–5583.

[34] A. Liwo, M.R. Pincus, R.J. Wawak, S. Rackowsky, H.A. Scheraga, Prediction of protein conformation on the basis of a search for

compact structures: test on avian pancreatic polypeptide, Protein Sci. 2 (1993) 1715–1731.

[35] S. Sun, Reduced representation model of protein structure prediction: statistical potential and genetic algorithms, Protein Sci. 2 (1993) 762–785.

[36] Y. Duan, P.A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, Science 282 (1998) 740–744.

[37] S.K. Burley, G.A. Petsko, Aromatic–aromatic interaction: a mechanism of protein structure stabilization, Science 229 (1985) 23–28.

[38] B.S. Frank, D. Vardar, D.A. Buckley, C.J. McKnight, The role of aromatic residues in the hydrophobic core of villin head piece subdomain, Protein Sci. 11 (2002) 680–687.

[39] T.J. Hubbard, J. Park, Fold recognition and ab initio structure predictions using hidden Markov models and beta-strand pair potentials, Proteins 23 (1995) 398–402.

[40] J. Cheng, P. Baldi, Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms, Bioinformatics 21 (2005) i75–i84.

[41] M. Boniecki, P. Rotkiewicz, J. Skolnick, A. Kolinski, Protein fragment reconstruction using various modelling techniques, J. Comput. Aided Mol. Des. 17 (2003) 725–738.

[42] G. Chikenji, Y. Fujitsuka, S. Takada, A reversible fragment assembly method for de novo protein structure prediction, J. Chem. Phys. 119 (2003) 6895–6903.

[43] J. Lee, S.Y. Kim, K. Joo, I. Kim, J. Lee, Prediction of protein structure using PROFESY, a novel method based on fragment assembly and conformational space annealing, Proteins 56 (2004) 704–714.

[44] M. Feig, C.L. Brooks III, Recent advances in the development and application of implicit solvent models in biomolecule simulations, Curr. Opin. Struct. Biol. 2 (2004) 217–224.

[45] Q. Fang, D. Shortle, Enhanced sampling near the native conformation using statistical potential for local side chain and backbone interactions, Proteins 60 (2005) 97–102.

[46] S.H. Bryant, C.E. Lawrence, An empirical energy function for threading protein sequence through the folding motif, Proteins 16 (1993) 92–112.