Abstract—Predicting the 3D native conformation of a protein given the amino acid sequence is known as protein structure prediction (PSP) problem and is one of the greatest challenges of computational biology. The current work uses a real valued Genetic Algorithm (GA), a powerful variate of GA to simulate the PSP problem. This algorithm consists of basic evolutionary operators and a fitness vector. The fitness vector is designed by combining a set of knowledge based biophysical filters viz. persistence length, radius of gyration, packing fraction, hydrophobicity ratio and irregularity index of $\phi$ and $\psi$. This vector converts all these biophysical measures into a real value by using specific weights or factors. Overall goal of the GA is to maximize the fitness value. The algorithm has been validated on a set of known globular protein containing 2-4 secondary structure elements. For all the test protein the algorithm converges rapidly and the converged structure shows a RMSD (root mean square deviation) of 3-6\AA as compared to the native structure.

Index Terms—Real valued genetic algorithm, protein folding, structure prediction, structure evaluation, Bio-physical filters.

I. INTRODUCTION

The native structure of a protein determines its biological functions. With exponentially growing amount of genomic and proteomic data, generated from sequencing projects, demanding functional annotation; and with experimental determination of structures being expensive in terms of both time and resources; the problem of protein structure prediction (PSP) from sequences continues to draw the attention of researchers from diverse computational groups [1]–[5].

The available computational methods of PSP may broadly classified into two which rely on knowledge of structural patterns and rules derived from known structures (homology modeling, threading etc.) and ab-initio folding methods which do not require any prior structural references and depend only on the physicochemical characteristics of the constituents [5], [6].

Ab-initio prediction of protein structure from sequence can be divided into two major sub problems: (1) Sampling the vast conformational space using powerful search techniques such as monte carlo simulation [7], simulated annealing [8], genetic algorithm [9], [10] and various other methods and (2) Designing a scoring function that will distinguish the native from the non-native conformation.

In recent years, significant progress has been made towards ab initio protein structure prediction. BHAGEERATH web server by Jayaram et al. [1] utilizes an energy based methodology for narrowing down the search space of small globular proteins. PROTINFO web server by Samudrala et al. [2] utilizes simulated annealing for generating the structures and different scoring functions for selecting the final five conformers. SCRATCH web server by Baldi et al. [3] predicts the protein tertiary structure by utilizing recursive neural networks, evolutionary information fragment libraries. ASTRO-FOLD web server by Klepeis and Floudas [4] employs local interactions and hydrophobicity for the identification of helices and beta-sheets respectively followed by global optimization, stochastic optimization and torsion angle dynamics. The ROBETTA web server by Bakers group [11], [12] utilizes inter protein structural information from protein sequence data and fragment assembly methodology for structure prediction.

In the current work, a real valued GA has been developed in C language to address the PSP problem. It is different from other existing GA as it operates on real values and not on bit strings. The algorithm takes the amino acid sequence and secondary structure information as inputs from the user.

A population of protein conformation is created taking into account the constraint of Ramachandran plot statistics and secondary structure information. The initial population is gradually improved by selection and application of genetic operators during the optimization. After the algorithm converges the optimized population is further analyzed and the structure showing close resemblance with the native structure is identified.

II. THEORY

A. Representation Formalism

For application of real valued GA one can use Cartesian coordinates, torsion angles [13], [14], rotamers etc. Cartesian coordinates representation has the advantage of being easily converted to and from the 3-dimensional conformation of a protein. However, it has the disadvantage that a mutation operator would in most instances create invalid protein conformations where some atoms lie too far apart or collide. Whereas torsion angle representation provides enough degrees of freedom to represent any native conformation. A small change in the $\phi/\psi$ angle can induce large changes in the...
overall conformation helping us to create variability within a population.

B. Methodology used for random structure generation

In the present work the torsion angle representation is used. The individuals of the population corresponds to different conformation of the same polypeptide represented by specific \( \phi \) and \( \psi \) angles. The program first classify the residues on the basis of secondary structure helix (H), sheet (E) and Coil (C). The allowable ranges of \( \phi \) and \( \psi \) and their respective correspondence with H and E is determined by calculating the average ± standard deviation (\( \delta \)) for the \( (\phi/\psi) \) torsion angles, as determined by Sven Hovmöller et al. [15]. The residue forming coil can pick up any set of \( \phi \) and \( \psi \) value from the allowed regions of Ramachandran plot. In stead of random assignment of values, we have used occurrence densities in a database of protein structure to provide relative weight for different \( \phi/\psi \) combination. The torsion angle of 405 proteins structure corresponds to 82,995 amino acids were studied. A 36 bin structure covering \(-180^\circ\) to \(+180^\circ\) has been created. Each bin corresponds to an interval of 10 degree torsion angle. 40 sets of 36 bin structure were created for the 20 amino acids covering both the torsion angles. The computed torsion angle data was classified into their respective bins. The percentage frequency of each bin has been calculated. The bins showing a significant amount of percentage frequency are considered for further computation. The probability of selecting a particular bin for a residue forming coil depends on its percentage frequency. For \( \omega \) torsion angle constant value of \( 180^\circ \) was used.

C. Genetic Algorithm

Genetic Algorithm is based on the “survival of the fittest mechanism” of evolution [9], [10]. GA is different from other available search methods as it maintains a population of solutions. This population advances through successive generations in which the solutions are evolving via genetic operations, in such a way that it gives better survival chance to fitter solutions and keeps a large diversity within the population [16]. Although GA does not guarantee to find the absolute best solution, but it is powerful enough to select the combination of traits that enables the system to function in a best suitable manner in the given environment [17].

In order to combine individuals of one generation to produce new offsprings, the current GA applies 3 types of genetic operators. The decision about the application of an operator is made during run time and can be controlled by various parameters.

1) Crossover: We have used two types of crossover operator as shown in Fig.1. Parents from the current population are selected by tournament selection and grouped pair wise. For each pair, two independent decision are made whether to apply crossover operator or not and second which one to apply.

2) Mutate: The GA mutation operator helps to increase population diversity. Whether an angle will be modified by mutate operator or not is independently decided. For each angle a random number between 0 and 1 is generated and if this number is greater than the mutate parameter at that time then the angle will be replaced by a random choice of one of the most frequently occurring values.

3) Variate: The variate operator consists of three components: the \( i^\circ \), \( 2^\circ \) and \( 5^\circ \) operator, each one refers to the magnitude of angle variation. If variate operator gets activated for a torsion angle then two decisions are made: which of the three components shall be selected and whether to increment or decrement the angle.

D. Fitness Function

For the protein folding application one has to consider more than one physico-chemical property [13]. One way to handle such situations is to combine the different fitness measures into a real value by using weights or factors. Here we have designed a fitness vector as shown in Fig.2 by combining a set of biophysical measures.

Persistence length (PL) is defined as the distance over which the direction of the polymer segment persists and has been extensively used to describe the rigidity of a synthetic polymer and DNA [18]. The algorithm finds the distance between the N-terminal of the \( i^{th} \) residue and C-terminal of the \( (i+j)^{th} \) residue and keep on comparing the distance with \( (i+j+1)^{th} \) residue; where \( i,j \in [1,n] \) and \( n \) represents the total number of residue in the protein. The above process continues till the last residue. The longest distance obtained is called as the PL of that protein. For globular protein its value varies from 15\( \AA \) to 60\( \AA \) with an averages around 40\( \AA \) [19]. These values can be used to select the native like conformations during the GA run.

Radius of gyration (RG) is an indirect measure of the size of the protein. It describes the overall spread of the molecule in and defined as the root mean square distance of the collection
of atoms from their common center of gravity. RG of the globular protein is proportional to \( N^{3/5} \) (N is the no. of residues) and should satisfy the following equation.

\[
R_g = \alpha N^{3/5} + \beta
\]

(1)

To distinguish the non native structure from the native structure the intercept \( \beta \) of the equation was varied, keeping the slope \( \alpha \) same. The value of \( \alpha \) is fixed to 0.359, whereas beta varies from 4.257 to 11.257 with an average value of 7.257 [19].

Packing fraction (PF) shows the compactness of a protein. It is defined as the fraction of volume in a structure occupied by the atoms. It is given as follows

\[
PF = \frac{\text{Sum of volume occupied by all the atoms}}{\text{Volume of the protein}}
\]

(2)

The formation of secondary and tertiary structures in proteins drives them towards achieving high packing densities. Folded proteins are known to exhibit an average PF of 0.7 and lies between 0.6 and 0.8 within 95% confidence limits [19], [20]. Proteins fold in a way such that all the hydrophobic residues occupy the core region and the hydrophilic residues are relatively more exposed on the surface. This property can be converted into a computational filter, by quantifying the “non-polar in and polar out” property as a ratio called as Hydrophobicity ratio (HR).

\[
HR = \frac{\text{Loss of ASA per atom of nonpolar atoms}}{\text{Loss of ASA per atom polar atoms}}
\]

(3)

Only backbone atoms are considered for the calculation. The ASA was computed using NACCESS [21] software package. The HR for globular proteins varies from 1.0 to 1.9 with very few expectations [19].

It has been observed that loop are made up of alpha-helix-like and beta-sheet-like dihedral, which fall in certain allowed region. The allowed ranges of \( \phi \) is \([-30, -150]\) and the allowed ranges for \( \psi \) are \([-45, +15] \& [+120, +180]\). Values that do not fall into either of the above categories were considered to be irregular. Given a conformation the irregularity index of \( \phi \) and \( \psi \) has been calculated as:

\[
Irr_{\phi/\psi} = \frac{\text{Total number of Irregular } \phi/\psi}{\text{Total number of residues}} \times 100
\]

(4)

The thresholds limits of \( \phi \) and \( \psi \) for most native like structure are 1.5%, 5.0% respectively [22].

For all the structure generated, the above parameters are computed and normalized on a scale of 0-1. The normalization is performed on the basis of closeness to the average value. Specific weight has been assigned to each filter on the basis of their efficiency to select the native structure. RG and PL were found to be more efficient in weeding out the non native like structure as compared to PF and HR [1], [19], and assigned with comparatively higher weight. The irregularity index of \( \phi \) and \( \psi \) has an accuracy of 60% in rejecting the structure and hence assigned with lesser weight. The overall fitness of the individual were computed by using following equation

\[
Fitness = 0.25 * PL + 0.25 * RG + 0.15 * PF + 0.10 * HR + 0.10 * I_{rr\phi} + 0.10 * I_{rr\psi} \tag{5}
\]

The overall objective of this algorithm is to maximize the above obtained fitness value. This fitness value determines the probability of selecting a particular structure to the next generation through elicit replacement.

### III. METHODOLOGY AND RESULTS

There are a number of parameters that control the run time behaviors of the GA. At the start of the run, the probability for a torsion angle to be modified by the mutate operator is 80%, at the end of the run it becomes 20%. In between the probability decreases linearly with the number of generations. In contrast, the probability of applying the variate operator increases from 20% at the beginning to 80% at the end of the run. The \( 5^\circ \) component of the variate operator is dominant at the start of the run (60%), whereas the \( 1^\circ \) component dominate at the end (80%). Likewise, the chance of performing a crossover rises from 10% to 70%. At the beginning of the run mutate and two-point-crossovers are applied mostly so that many different regions of the search space are covered. At the end of the run the \( 1^\circ \) component of the variate operator dominates the scene. This is intended for fine tuning of those conformations that have survived the selection pressure of evolution so far.

We used the SCWRL-3.0 program [23] for the proper side chain placement. It is a side-chain prediction algorithm used for proper placement of the side chains to a fixed backbone. The above described algorithm has been tested on a set of small globular proteins. Their length varies from 17 to 46 residues and total number of helices and strands ranges between 2-4. The population size was fixed to 30 and maximum number of generation was set to 100. For all the test proteins the algorithm converged within 50 generation, which takes \(~10\) min. The final population obtained after convergence of the algorithm is analyzed and structure showing close similarity with the native structure is identified. The result obtained are shown in Table 1. The (v) and (vi) columns of the table shows the lowest backbone RMSD, calculated using Superpose Version 1.0 [24] and the corresponding fitness respectively. It may be noted that most converged individuals of the GA show topological correct structure within an RMSD of \(~3.6\) \AA. Fig.3 shows superimposition of the lowest RMSD structure with the respective native like structure.

### IV. CONCLUSIONS

The proposed algorithm arrives at the tertiary structure of the proteins starting from the primary sequence and secondary structure information for the small globular protein. It is applied on a set of 6 small globular proteins having secondary element with \( \alpha \), \( \beta \) and \( \alpha/\beta \). The algorithm converged in \( <10 \) min for all the 6 protein. The optimized structure shows a close resemblance with the native structure with a RMSD of \(~3.6\) \AA.
Fig. 3. The superimposed lowest RMSD structure for some small globular test proteins used for the validation of the algorithm: [a] 1E0Q; [b] 1WQC; [c] 1Q2K; [d] 1ROO; [e] 1RES; [f] 1CRN. The predicted structure by the algorithm is shown in red and the native is in blue.

TABLE I

<table>
<thead>
<tr>
<th>(i) No.</th>
<th>(ii) PDB ID</th>
<th>(iii) No. of residues</th>
<th>(iv) No of 2° elements</th>
<th>(v) Lowest RMSD Å</th>
<th>(vi) Fitness value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1E0Q</td>
<td>17</td>
<td>2E</td>
<td>2.99</td>
<td>0.8294</td>
</tr>
<tr>
<td>2</td>
<td>1WQC</td>
<td>26</td>
<td>2H</td>
<td>5.33</td>
<td>0.7969</td>
</tr>
<tr>
<td>3</td>
<td>1Q2K</td>
<td>31</td>
<td>1H, 2E</td>
<td>4.86</td>
<td>0.8459</td>
</tr>
<tr>
<td>4</td>
<td>1ROO</td>
<td>35</td>
<td>3H</td>
<td>5.99</td>
<td>0.8194</td>
</tr>
<tr>
<td>5</td>
<td>1RES</td>
<td>43</td>
<td>3H</td>
<td>5.91</td>
<td>0.8678</td>
</tr>
<tr>
<td>6</td>
<td>1CRN</td>
<td>46</td>
<td>2H, 2E</td>
<td>5.69</td>
<td>0.8778</td>
</tr>
</tbody>
</table>

V. ACKNOWLEDGMENTS

The authors would like to acknowledge Prof. B. Jayaram, Indian Institute of Technology, New Delhi, India for his constant support while writing the real valued GA and also wishes to thank DST, Govt. of India for supporting this work.

REFERENCES


[22] Lipi thukural, Sandhya R Shenoy, Kumkum Bhushan and B Jayaram (2007) ProRegIn: A regularity index for the selection of native-like tertiary structures of proteins, J. Biosci., 32(1), 71-81
