Abstract—Genetic algorithms have proved to be a successful method for predicting the protein structure. In this paper, we propose a new intermediate selection strategy for genetic algorithms and we implement it for protein folding problem. In a standard genetic algorithm the children replace their parents. The idea behind this is that both parents pass on their good genetic material to their children. In practice however, children can have worse fitness than their parents. We therefore propose another intermediate selection step, which we call as modified keep-best reproduction (MKBR) that ensures that new genetic information is entered into the gene pool, as well as good previous genetic material is being preserved. We have demonstrated the superiority of modified keep-best reproduction on several instances of the protein-folding problem, which not only finds the optimum solution, but also finds them faster than the standard generational replacement schemes.

I. INTRODUCTION

Proteins fold rapidly and reliably to their functional state (native state). Whether the native state is kinetically or thermodynamically controlled remains an open question. The native state can therefore be the global energy minimum or a low-lying meta stable conformer. The energy hypersurface has high dimensionality and complexity. The primary structure of a protein is the amino acid sequence of its polypeptide chain, while the secondary structure is the local arrangement of a polypeptide’s backbone atoms without regard to the conformations of its side chains. Under certain physiological conditions, the primary structure of a protein spontaneously folds into a precise three-dimensional form called its tertiary structure or native state that determines its functional properties. Finding energetically low lying conformations given a sequence of amino acids is termed as "The Protein Folding Problem"[1]. Currently, the primary structures of approximately 40,000 proteins are known. Only a small percentage of these have known native states. Efforts aimed at solving the Protein Folding Problem have involved the optimization of a potential energy function that approximates the thermodynamic state of a protein macromolecule. Since an algorithm using such a potential function does not give insight into how a protein folds, these approaches are instead known as Protein Structure Prediction.

In this study protein conformations and intra molecular interactions are modeled using the simplified HP Bead Model. A genetic algorithm with a new selection strategy has been employed to search the hyper-surface.

II. THE HP MODEL

The hydrophobic-hydrophilic model (HP model) by Dill [3] is a simple abstraction that captures the essence of the important concepts of Protein Structure Prediction. In the HP model, amino acids are divided into two categories: hydrophobic (H) and hydrophilic (P). The primary sequence of a protein is therefore $S \sum \{H, P\}^n$. Using this simplification, optimization models can be developed that seek to maximize interactions between adjacent pairs of hydrophobic amino acids (or hydrophobes). Adjacency is considered only in the cardinal directions of a lattice upon which the sequence is embedded. In an HP lattice, vertices represent amino acids and edges represent connecting bonds. Black squares at the vertices indicate hydrophobes, while white squares indicate hydrophilic amino acids. A lattice can be two or three dimensional, and either square, cubic or triangular. The hydrophobic-hydrophobic (HH) contacts are the basis for the evaluation function. Every pair of hydrophobes that are adjacent on the lattice and not consecutive in the primary sequence is awarded a value $ε$ (usually $−1$). The sum of all such values gives the energy of the conformation. The amino acid sequence is "folded" on a two-Dimensional square lattice on which at each point, the chain can turn $90°$ left or right, or continue ahead. Figure 1 shows a 20-length sequence embedded on a square lattice, HH contacts indicated by dotted arrows.

Fig:1 HP sequence of length 20 on a square lattice with energy $−5$

III. IMPLEMENTATION USING GENETIC ALGORITHM

A. Traditional Genetic Algorithm

Our implementation of the genetic algorithm is same as that described in Unger and Moul [8] but we introduce MKBR as an intermediate selection strategy to improve the
efficiency of the algorithm. The solutions are not encoded as binary strings but rather are the conformations themselves, which are treated directly in the spirit of genetic operators. The process starts with V extended structures. In each generation each structure is subject to a number of mutation steps with rate ranging from 0.01 to 0.20. Each mutation is the same as a single Monte Carlo (MC) step [8] and is subject to similar acceptance criteria as in a MC process. At the end of this MC stage [10], the crossover operation is performed. The chance $p(S_j)$ of a structure being selected for crossover is proportional to its energy value $E_i$, That is

$$p(S_j) = \frac{E_i^N}{\sum_{j=1}^{N} E_j^N}.$$ 

Thus, the lower energy conformations have a higher chance of being selected. For a pair of selected structures a random point is chosen along the sequence and the X-terminal portion of the first structure is connected to the C-terminal portion of the second structure (see Fig. 2). As there are three ways to join the parts together (connecting the chains with angles of 0°, 90° or 270°), these possibilities are tested in a random order to find one that is valid (That is, where no residue from one structure occupies a lattice point used by a residue from the other). If none of the three ways led to a self-avoiding structure, then another pair of structures is selected. Once a valid structure $S_k$ is created, its energy $E_k$ is evaluated and compared to the averaged energy $E_{ij} = (E_i + E_j)/2$ of its "parents" [7,8]. The structure is accepted if $E_k \leq E_{ij}$ or if the energy will be increased based on the decision:

$$Rna < \exp \left[ \frac{E_{ij} - E_k}{c_k} \right].$$

This crossover operation is repeated until $N - 1$ newly accepted hybrid structures have been constructed to constitute the population of the next generation. We allow a higher acceptance rate for bad moves that increase the energy for mutation steps than for crossovers. This strategy maintains the diversity of the population and prevents premature convergence to a few low energy conformations.

The process starts with a population of fully extended structures. Each structure undergoes a MC stage followed by a crossover stage. In the crossover stage, pairs of structures are randomly (based on their energies) cut and pasted. In this example the cut point was randomly chosen to be after residue 14. Joining the first 14 residues of (A) with the last 6 residues of (B) and applying a randomly chosen 270° rotation at the joint achieves the compact structure in (C). In this case, the energy value of the hybrid (C) is -9, lower than the energies -5 and -2 of its "parents". The hybrid is always accepted if its energy is lower than the averaged energies of its parents. or non deterministically accepted according to its energy increase.

B. Modified keep best Reproduction strategy (MKBR)

At the end of each generation we apply an intermediate selection strategy, which we call as Modified Keep Best Reproduction Strategy, which is a modification of KBR [11,12]. In the keep best reproduction strategy only the best of the 2 offspring chromosomes were selected and it was replaced by the best parent. This ensures that both the best offspring and parent chromosome are kept. We identified that there is a potential danger in this approach. In many cases the rejected offspring is better than the next pair of parents selected for evaluations [4]. In such cases the GA would not make significant progress, and also loses valid offspring. They have also shown that the standard selection strategy [10,13], which always keeps both offspring, is shown to have inferior performance on test problems Figure 3 shows a modification of algorithm that incorporates modified keep-best reproduction. It is intuitively clear that KBR has a higher selection pressure than the standard replacement technique of replacing both parents by their two offspring[5,6]. We refer to the latter selection strategy as Standard Selection or short STDS [10]. Both use the same parent selection strategy but MKBR employs an additional selection step on the parents and children in order to decide who will survive into the next generation By keeping the best child we seek to achieve fast convergence. Through controlling the selection pressure by keeping the best parent we seek to prevent premature convergence. MKBR should not be confused with tournament selection. Tournament selection is a parent selection method that randomly chooses s individuals from the population and the best of those s individuals becomes a parent. Here s is the size of the tournament. Increasing s increases the selection pressure [9]. MKBR works locally only on the set of the parents and the set of the children. It does not have the same random component as tournament selection has. Also tournament selection only decides who is chosen for reproduction [2], MKBR decides who will live into the next generation unlike elitist approach.

Figure 2. The crossover operation
C. Results

MKBR not only found the optimal results but it also converged to optimum conformations in lesser number of energy evaluations. Fig (4) shows one of the optimal energy conformations for the sequences of length 20, 24, 25 and 36 respectively. The arrow represents the H-H bond which contribute to the minimum energy. White square represents the Hydrophilic amino acid and a Black square represents hydrophobic amino acid.

Eight sequences for tested for each and GA with MKBR strategy showed better results with less number of energy evaluations, than the others. The sequences tested are given below.

(20)HPHPPHHPPHPHPPHHPP
(24)HPHHPPHHPPHHPPHHPP
(25)PPHPPHHPPHHPPPPPHHHPP
(36)PPPHHPPHHPPPPPHHHHHPPHHPP
(48)PPPHHPPHHPPHHPPHHPPHHPP
(50)PPPHHPPHHPPHHPPHHPPHHPP
(60)PPPHHPPHHPPHHPPHHPPHHPP
(64)PPPHHPPHHPPHHPPHHPPHHPP

IV. COMPARISON OF TRADITIONAL SELECTION AND MODIFIED KEEP BEST REPRODUCTION STRATEGY

The experiment was done on the standard amino acid sequence of length 20,24,25,36,48,50,60,64 with a population size of 200. The mutation rate was 2% to 15%. Each application of a genetic operator is counted as a step. Thus, a generation takes 10 X population size times mutation steps plus the number of crossover trials it take. When a valid conformation is encountered, its energy is evaluated. At the end of each generation the chromosome with worst fitness is replaced with the best parent. The simulation was run for several times. The optimal conformation found for each combination is tabulated. A comparison of the number of energy evaluations of the other methods is also given. It is observed that the GA with MKBR finds optimum solution with minimum number of energy evaluations.

<table>
<thead>
<tr>
<th>Table 1: Energy Evaluations for Protein Folding Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>36</td>
</tr>
<tr>
<td>48</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>64</td>
</tr>
</tbody>
</table>
The chart shows the variation in performance of the compared methods (Figure 5). We use higher mutation rates in combination with crossover. The x-axis shows the length of sequence, while the y-axis shows the number of energy evaluations. The population size was 200. With MKBR we were able to speed up the convergence of the GA by using higher mutation rates. MKBR only keeps the best parent and eliminates the worst child. In case mutation lowers the fitness of the offspring, there is always the good genetic material of the best parent that is kept. So, higher mutation rates are not as disruptive as with traditional GA.

![Chart showing performance variation](chart.png)

Fig(5) The x-axis shows the sequence number while the y-axis shows the number of energy evaluations.

V. CONCLUSION

Modified keep-best reproduction outperforms the standard generational replacement technique significantly on protein folding problems, especially as the problem size increases in terms of time and optimality. The collected data also demonstrates that modified keep-best reproduction are best suited for problems with higher genetic operator probabilities, especially the mutation probability. With standard selection, a high mutation rate usually has a negative effect on the performance of the GA, because of schema disruption. Our selection strategy benefits from higher mutation rates, since we always keep the best parent conformation, thus limiting the disruptive effect that mutation can have.

VI. FURTHER RESEARCH AND DISCUSSION

This paper describes a research on a GA based system for protein folding problem. From initial analysis of the data collected we made the following observations: Even with significantly smaller population sizes, modified keep-best reproduction finds better solutions than standard selection with much larger populations. This means that a better solution can be found with less function evaluations and thus with less total computing time. Another observation we made is that keep-best reproduction is less susceptible to a change in crossover probability and mutation probability and the performance would not so heavily depend on parameter settings such as population size, crossover probability, and mutation probability. We need to test Modified keep-best reproduction on larger amino acid sequences. We are confident however, that on these problems we will still be able to get better results than the standard selection.

APPENDIX-A

The sample output of the actual structure of protein got through simulation for the amino acid sequence PPPHPPPPPPPPHHHHHHPPPPHP of length 36 is given below. The HP square lattice is constructed as the data simulated in the table given below and the structure is constructed as in fig given below.

<table>
<thead>
<tr>
<th>1p</th>
<th>2p</th>
<th>3p</th>
<th>4h</th>
<th>21p</th>
<th>24p</th>
<th>25p</th>
<th>26p</th>
<th>27p</th>
</tr>
</thead>
<tbody>
<tr>
<td>6p</td>
<td>5h</td>
<td>20p</td>
<td>19p</td>
<td>30p</td>
<td>29p</td>
<td>28p</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7p</td>
<td>8p</td>
<td>17p</td>
<td>18p</td>
<td>31p</td>
<td>32p</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9p</td>
<td>16p</td>
<td>15h</td>
<td>34h</td>
<td>33p</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10p</td>
<td>11p</td>
<td>14p</td>
<td>35p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12p</td>
<td>13p</td>
<td>36p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The hydrogen bond corresponding to energy is indicated with pointed arrow.

ACKNOWLEDGMENT

The authors would like to thank Professor K. Murugesan, Department of Mathematics and Computer Applications, National Institute of Technology, Trichy, Tamil Nadu, India, for his valuable ideas and comments given to improve the technical quality of this paper.
REFERENCES


