



PROTEIN TERTIARY STRUCTURE PREDICTION USING EVOLUTIONARY ALGORITHMS

S. SUDHA^{1*}, S. BASKAR²

Department of Computer Science,
Thiagarajar College of Engineering,
Madurai- 625015,
India.

S. KRISHNASWAMY³

Centre of Excellence in Bioinformatics,
Madurai Kamaraj University,
Madurai-625021,
India.

Abstract: In this paper, Protein Tertiary Structure Prediction using Evolutionary Algorithms (EAs) such as Self-Adaptive Differential Evolution (SaDE) and Real-coded Genetic Algorithm (RGA) are discussed. RGA is implemented with various crossover and mutation operators. The algorithms are tested on a peptide Met-enkephalin. The energy functions used are ECEPP/2 and ECEPP/3 force fields. SaDE and RGA with discrete crossover and boundary mutation produce the best energy values than other crossover and mutation operators of RGA. But, Statistical results of SaDE and RGA show that SaDE outperforms RGA in terms of number of function evaluations, mean energy and success rate. The best results obtained using SaDE and RGA are compared with native structure 1PLW and classical benchmark Scheraga conformation and the corresponding minimum RMSD values are 2.13 Å and 1.45 Å respectively. Comparison of the best results of SaDE and RGA with other reported RGA variants show better performance in terms of energy and computational search efficiency. A set of unique hundred best solutions obtained from both algorithms are clustered using hierarchical cluster algorithm. This gives seven independent clusters suggesting the robustness of these methodologies and the ability to explore the conformational space available and to populate the near native conformations.

Keywords: Protein Structure Prediction, Self-adaptive Differential Evolution, Genetic Algorithm, Met-Enkephalin, Energy function

I. Introduction

Proteins are the most important molecular entities of a living organism and understanding their functions is an important task to treat diseases and synthesize new drugs. Protein is a linear chain of amino acids. These amino acids fold into complex structures based on chemical and physical properties. These folds determine the function of the protein provided that the folding in its native structure, which is the correct three dimensional structure of the protein [1].

Native protein structures have been experimentally determined using X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectrographic methods. Massive amounts of protein sequence data are produced by modern large-scale DNA sequencing efforts such as the Next Generation Sequencing methods. However the number of experimentally determined protein structures is lagging far behind the output of protein sequences[1] due to the time-consuming and relatively expensive nature of X-ray crystallography and NMR. Therefore, computational techniques are needed to predict the protein structure in order to understand structure-function relations.

Two important problems in structure prediction are finding global minimum free energy of a protein and the necessity to explore the large space of possible protein conformations. These problems can be partially bypassed in homology modeling and fold recognition methods, whereas the *de novo* or *ab initio* protein structure prediction method explicitly resolves these problems. These modeling methods seek to build three-dimensional protein models "from scratch", based on physical principles rather than on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e., global optimization of a suitable energy function). Efforts to solve such PSP problem via energy minimization correspond to find the global minimum free energy of a

Protein. PSP has been proven to be an NP-hard problem[2]; the number of conformations grows exponentially with the number of residues. Thus the non-deterministic search techniques have dominated attempts, of which ample approaches such as, Monte Carlo simulation [3], Simulated Annealing [4], and Ant Colony Optimization [5], though because of their simplicity and search effectiveness, Genetic Algorithm (GA)[6],[7] is one of the most attractive algorithm and hybrids between deterministic and stochastic methodologies using nonlinear optimization techniques and maximum likelihood [8] approaches.

In Reference[6], a pioneer work using GA for PSP, torsion angle representation was used and the GA reached even lower energy levels than the protein in its native state; because of imprecise energy function. In Reference[7], the same author extended the work with hybrid representation and used three more operators to manipulate the individuals and concluded that none of the conformations generated with a non-biased fitness function were similar to the native conformation but all of them showed much better overall fitness than the native structure. Other variants of GA like Parallel and hybridization of GA with other techniques[9],[10], Improving diversity in population[11] were applied to this problem.

Calculation of protein free energy for all atoms can be obtained using various energy functions like, AMBER[12], CHARMM[13], ECEPP[14] and other energy function such as the residue-residue effective potential are also in use today.

Met-enkephalin, a natural endogenous opioid consisting of five residues, was considered in the gas phase with the N- and C-terms capped by acetyl and methyl-amide groups respectively. [Met]-enkephalin polypeptide is a paradigmatic example of multiple minima problem. It is estimated to have more than 10^{11} locally optimal conformations. It is for this reason that the peptide has been widely studied in the protein structure prediction context[11],[15]-[17]

Since the energy landscape of the problem is non-linear and massively multimodal, robust search techniques are required to locate the global free energy minimum. Though there were lots of researches on this protein, there is no global optimum for this problem. So, it needs new search algorithms or problem-specific operators to explore the search space.

Self-adaptive Differential Evolution (SaDE) algorithm[18], has been proposed as efficient and effective global optimizer in the continuous search domain. The algorithm makes use of pool of trial vector generation strategies and three crucial control parameters. Since these strategies and control parameters are gradually self-adapted by learning from their previous experiences in generating promising solutions, it can exhaustively search the continuous search space domain. The performance of the SaDE is extensively evaluated on a suite of 26 bound-constrained numerical optimization problems and is concluded that SaDE is more effective in obtaining better quality solutions, which are more stable with the relatively smaller standard deviation, and has higher success rates.

Recently, a number of crossover and mutation operators such as Discrete Crossover, Mid-point Crossover, Heuristic/Intermediate Crossover[19] Simulated Binary Crossover[20], Boundary mutation[21] and Polynomial mutation[22] are proposed for real-coded GA. Hence, in order to demonstrate the efficiency of SaDE and RGA on Protein Tertiary Structure Prediction, simulations are performed and reported in this paper.

The organization of this paper is as follows: In section II, Self-Adaptive Differential Evolution (SaDE) and Real-coded Genetic Algorithm (RGA) with recent version of crossovers and mutation are discussed. In section III, PSP and methodologies used for minimizing energy are explained. In Section IV results and performance of these algorithms are discussed.

II. Evolutionary Algorithm

Evolutionary algorithms (EAs) are search methods that take their inspiration from natural selection and survival of the fittest in the biological world. EAs operate with the population of solutions. The initial population is generated randomly. A basis of an evolutionary algorithm represents an evolutionary search in which the selected solutions undergo an operation of reproduction, i.e., a crossover and a mutation. As a result, new candidate solutions (offspring) are produced that compete, according to their fitness, with old ones for a place in the next generation. The fitness is evaluated by an evaluation function that defines requirements of the optimization (minimization or maximization)[23]. Recent popular versions of EAs such as RGA and SaDE algorithm as used in this paper are discussed in the following subsections.

A. Self adaptive Differential Evolution(SaDE)

The differential evolution (DE) algorithm, proposed by Storn and Price[24], is a powerful population-based stochastic search technique, which is an efficient and effective global optimizer in the continuous search domain. In DE, there exists much trial vector generation strategies out of which a few may be suitable for solving a particular problem. Moreover, three crucial control parameters involved in DE, i.e., population size, scaling factor, and crossover rate, may significantly influence the optimization performance of the DE. Therefore, to successfully solve a specific optimization problem, it is generally required to perform a time-consuming trial-and-error search for the most appropriate strategy and to tune its associated parameter values, i.e., the values of CR, F and NP. It is observed that trial-and-error searching process requires high computational costs and adapting appropriate strategies, when the population moves to different regions of search space during generations is very difficult. A self-adaptive DE (SaDE) algorithm is proposed by A. K. Qin et.al[18] to avoid the expensive computational costs spent on searching for the most appropriate trial vector generation strategy as well as its associated parameter values by a trial-and-error procedure. Here, both strategies and their associated parameters are gradually self-adapted by learning from their previous experiences in generating promising solutions.

A.a. Trial Vector Generation Strategy Adaptation

Instead of employing the computationally expensive trial-and-error search for the most suitable strategy and its associated parameter values, they maintain a strategy candidate pool including several effective trial vector generation strategies with effective yet diverse characteristics. Pool consists of the following strategies defined in Eqn.s (1)-(4)

1. DE/rand/1/bin:

$$u_{i,j} = \begin{cases} x_{r_{1,j}} + F \cdot (x_{r_{2,j}} - x_{r_{3,j}}), & \text{if } \text{rand}[0,1] < CR \\ & \text{or } j = j_{rand} \\ x_{i,j}, & \text{otherwise} \end{cases} \quad \dots (1)$$

2. DE/rand-to-best/2/bin:

$$u_{i,j} = \begin{cases} x_{i,j} + F \cdot (x_{best,j} - x_{i,j}) + F \cdot (X_{r_{1,j}} - X_{r_{2,j}}) \\ + F \cdot (X_{r_{3,j}} - X_{r_{4,j}}), & \text{if } \text{rand}[0,1] < CR \\ & \text{or } j = j_{rand} \\ x_{i,j}, & \text{otherwise} \end{cases} \quad \dots (2)$$

3. DE/rand/2/bin:

$$u_{i,j} = \begin{cases} x_{r_{1,j}} + F \cdot (x_{r_{2,j}} - x_{r_{3,j}}) + F \cdot (x_{r_{4,j}} - x_{r_{5,j}}), \\ & \text{if } (\text{rand}[0,1] < CR \text{ or } j = j_{rand}) \\ x_{i,j}, & \text{otherwise} \end{cases} \quad \dots (3)$$

4. DE/current-to-rand/1:

$$U_{i,G} = X_{i,G} + K \cdot (X_{r_{1,G}} - X_{i,G}) + F \cdot (X_{r_{2,G}} - X_{r_{3,G}}) \quad \dots (4)$$

Stochastic universal selection method is used to select one trial vector generation strategy for each target vector in the current population. The probabilities of the strategies are updated only after an initial learning period (LP) generation which is set by the user. The probabilities are initialized to $1/k$, i.e., all strategies have the equal probability to be chosen. After the initial LP generations, the probabilities of choosing different strategies will be updated at each subsequent generation by,

$$p_{k,G} = \frac{S_{k,G}}{\sum_{k=1}^K S_{k,G}} \quad \dots (5)$$

where

$$S_{k,G} = \frac{\sum_{g=G-LP}^{G-1} n_{S_{k,G}}}{\sum_{g=G-LP}^{G-1} n_{S_{k,G}} + \sum_{g=G-LP}^{G-1} n_{f_{k,G}}} + \varepsilon, \quad k = 1, 2, \dots, K; G > LP$$

where $S_{k,G}$ represents the success rate of the trial vectors generated by the k^{th} strategy and successfully entering the next generation within the previous LP generations with respect to generation G . The small constant value $\varepsilon = 0.01$ is used to avoid the possible null success rates.

A.b. Parameter Adaption

The scaling factor F is approximated by a normal distribution with mean value 0.5 and standard deviation 0.3. CR is approximated as normally distributed variable mean CRm_k with respect to the k^{th} strategy and standard deviation 0.1. Initially, CRm_k is set at 0.5 for all the strategies for an initial LP generation. After LP generations, the CRm_k value is adapted with median of the successful CR values (that have generated trial vectors successfully entering the next generation) over the past LP generations for every subsequent generation.

B. Real coded Genetic Algorithm

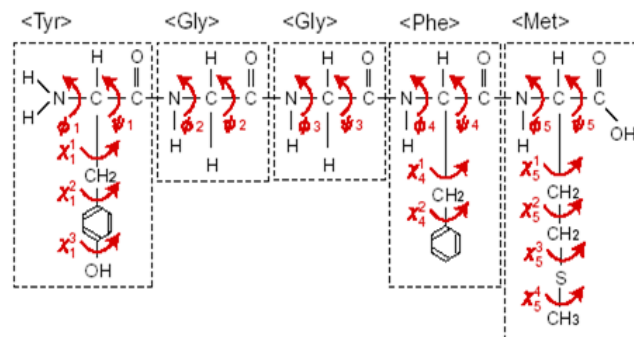
Genetic algorithms (GA), inspired by the biological theory of evolution is proposed by J. Holland in 1975, its self-organizing, adaptive, self-learning and group capacity to make it very suitable for the evolution in solving large-scale complex combinatorial optimization problem[25]. In GA, a population of chromosomes, representing a series of candidate solutions to an optimization problem (also called individuals), evolves toward better solutions. Selection and reproduction operators are used, inspired in mechanisms employed in the evolution of biological species[23].

In RGA, the representation of solutions is close to natural formulation of problems. Since there is no difference between genotype and phenotype, coding and decoding process is not needed as in Binary GA. The capacity of real parameters is to exploit the “graduality” of functions with continuous variables, where the “graduality” refers to the fact that slight changes in variables corresponds to slight change in function[26]. In RGA, the genes are represented as real world variables i.e. the chromosomes are vector of floating point numbers with required precision, in continuous domain. So, special real coded recombination operators are used. Various real-coded recombination operators like Discrete Crossover, Mid-point Crossover, Heuristic/Intermediate Crossover and Simulated Binary Crossover and the real-coded mutation operators like Boundary mutation and polynomial mutation are used.

III. PSP and its Implementation

The test-protein is Met-enkephalin, which consists of 5 residues (Tyr-Gly-Gly-Phe-Met amino acids) and it is made up of 24 torsion angles. The molecular structure of Met-enkephalin is given in Fig 1.

Figure 1 Molecular structure of Met-enkephalin.



The encoding method used is torsion angle representation. In this representation, protein is described by a set of torsion angles under the assumption of constant standard binding geometrics, bond length and angle constant. In this representation the conformations are created by changing the values of the torsion angles. The main chain torsion angles are: ϕ_i , ψ_i and ω_i and the side chain torsion angles are χ_i . As the overall structure of proteins can be described by their backbone and side chain torsion angles, the tertiary structure of a protein can be obtained by rotating the torsion angles around the rotating bonds. So, the protein conformation is represented as a sequence of the torsion angles.

Generating conformations is done by changing the values of the torsion angles randomly. Thus Met-enkephalin with 5 amino acids needs 24 torsion angles to represent a conformation. Of these 24 angles the ω is fixed to an ideal value of 180° . The remaining 19 angles can take a value between -180° and 180° .

A. Energy Function

Minimization of force fields ECEPP/2 & ECEPP/3[27] are used as objective function. The tool Simple Molecular Mechanics for Physics (SMMP)[28] is used to find the potential energy for each generated conformations. The objective function is defined in Eqn.(6):

$$E = \sum_{(i,j,k,l) \in D} \left(\frac{U_{\theta_{ijkl}}}{2} \right) (1 \pm \cos(n_{ijkl}\theta_{ijkl})) + \sum_{(i,j) \in N} \epsilon_{ij} \left[F_{ij} \left(\frac{r_0}{r_{ij}} \right)^{12} - 2.0 \left(\frac{r_0}{r_{ij}} \right)^6 \right] + \sum_{(ij) \in N} \left[332.0 \left(\frac{q_i q_j}{D_{rij}} \right) \right] + \sum_{(i,j) \in H} \epsilon_{ij} \left[\left(\frac{r_0}{r_{HX}} \right)^{12} - 2.0 \left(\frac{r_0}{r_{HX}} \right)^{10} \right] \quad \dots (6)$$

where the four terms represent the energy due to dihedral angle deformation non bonded interactions electrostatic interactions and hydrogen bond energy respectively. Specifically

- D the set of 4-tuples defining ω and χ dihedrals,
- N is the set of non-bonded atom pairs,
- H is the set of hydrogen bonding atom pairs,
- r_{HX} is the donor-acceptor distance,
- r_{ij} is the distance between atoms i and j
- θ_{ijkl} is the dihedral formed by atoms i, j, k , and l ,
- q_i is the partial atomic charges of atom i ,
- The $U_{\theta_{ijkl}}$'s, n_{ijkl} 's, F_{ij} 's, ϵ_{ij} 's, r_0 's and D are empirically determined constants

B. RMSD Calculation

The root-mean-square deviation (RMSD) is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins. Given two sets of n points v and w , the RMSD is defined in Eq.(7). For RMSD calculation, SUPERPOSE module from TINKER[29] tool is used.

$$RMSD(V, W) = \sqrt{\frac{1}{n} \sum_{i=1}^n \|v_i - w_i\|^2} \quad \dots (7)$$

$$= \sqrt{\frac{1}{n} \sum_{i=1}^n (v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2}$$

IV. Results and Discussion

SaDE and RGA are applied for Protein Structure Prediction with Met-enkephalin peptide. The algorithms are implemented using Java in Linux environment in PC with Intel Core2 operating at 2.00GHz with 3GB RAM. The SMMP package is used for ECEPP/2 and ECEPP/3 energy calculation. RGA with four combinations of genetic operators are used. In order to prove the consistency of SaDE and RGA, statistical results obtained in 50 independent trails are used. Convergence criterion based stopping is used. If there is no significant improvement in the best solution obtained from the last 100 iterations then algorithm is terminated. The quality of results is measured by finding RMSD for minimum energy conformation with native structure 1PLW and classical benchmark Scheraga. The results are compared with existing GA variants in terms of minimum energy and number of function evaluations.

A. SaDE Results

The simulation of PSP for Met-enkephalin using SaDE algorithm is tested for the four cases such as i) ECEPP/2 with varying omega value ($\omega = 175^\circ - 180^\circ$) ii) ECEPP/2 with fixed omega value ($\omega = 180^\circ$) iii)

ECEPP/3 with varying omega value ($\omega = 175^\circ - 180^\circ$) and iv) ECEPP/3 with fixed omega value ($\omega = 180^\circ$). The best energy obtained in 50 independent trails is tabulated in Table 1.

Table I Best Results of SaDE

Force field	Omega(ω)	Energy (kcal/mol)
ECEPP/2	Varying	-12.91
	Fixed	-10.72
ECEPP/3	Varying	-12.429
	Fixed	-10.903

Fig 2 shows the convergence characteristics of SaDE with ECEPP/2 (ω varying).

Figure 2 Convergence Characteristic Curve for ECEPP/2 conformation.

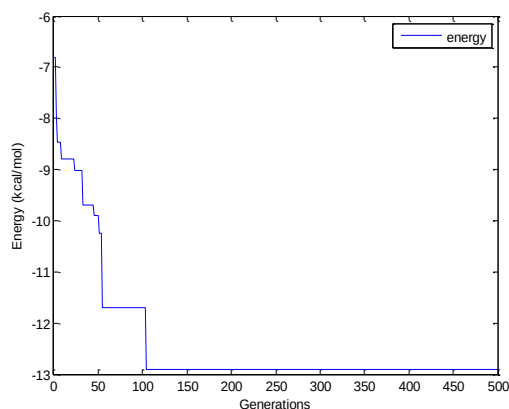
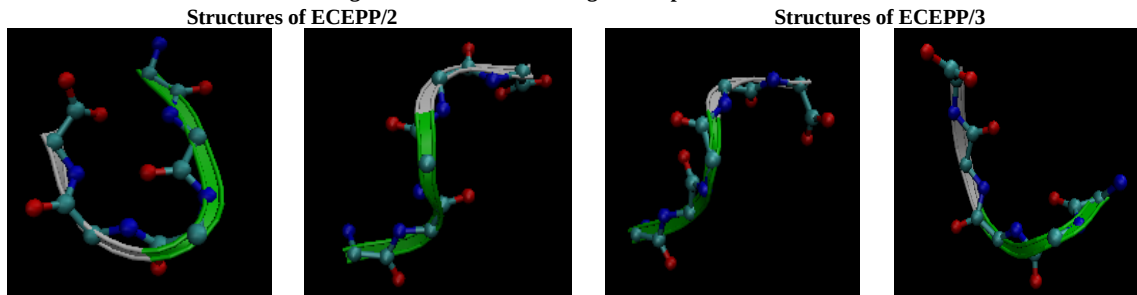


Fig 3 shows the backbone rendering of four major conformations of Met-enkephalin for the best results obtained using SaDE.

Figure 3 Backbone rendering of four predominant conformations



B. RGA Results

The best energy obtained for four genetic operator combinations in 50 independent trails are shown in Table 2.

Table III Best Results of RGA for four genetic operator combinations

S.No	Crossover	Mutation	Force Field	Energy (kcal/mol) (ω varying)	Energy (kcal/mol) (ω fixed)
1	Heuristic/Intermediate Crossover	Boundary Mutation	ECEPP/2	-7.75	-8.35
			ECEPP/3	-6.732	-7.69
2	Mid-Point Crossover	Boundary Mutation	ECEPP/2	-6.30	-8.27
			ECEPP/3	-7.18	-9.01
3	Simulated Binary Crossover	Polynomial Mutation	ECEPP/2	-8.63	-8.04
			ECEPP/3	-7.95	-6.45
4	Discrete Crossover	Boundary Mutation	ECEPP/2	-12.91	-10.72
			ECEPP/3	-12.429	-10.903

The results show better performance of Discrete Crossover and Boundary Mutation in finding the structure with lowest free energy.

Statistical results for ECEPP/2 with ω varying using SaDE and RGA with Discrete Crossover and Boundary Mutation are shown in Table 3, where Success Rate (SR) is defined as ratio of number of successful runs and the total runs. Successful run is the one, which gives energy that is better than or equal to the reported energy for the same problem.

Table IIIII Statistical Results for ECEPP/2 with ω varying

Algorithm	Best energy (kcal/mol)	Worst Energy (kcal/mol)	Average no. of function evaluations	Mean	Std. Deviation	SR (%)
SaDE	-12.91	-10.91	17697	-12.751	0.499	90
RGA	-12.91	-10.019	28771	-12.57	0.759	82

The best energy for both the algorithms is -12.91 kcal/mol, but the higher SR of SaDE proves the efficiency of the algorithm in producing consistent results. The mean energy obtained in 50 trails is very close to the best energy for SaDE, which concludes that the SaDE is more consistent than RGA. Also, reduced number of average function evaluations of SaDE shows that the computational time is 38% lesser as compared to RGA.

C. Performance Analysis

In Table 4, the results obtained using SaDE and RGA for ECEPP/3 are compared with MOLS[17] algorithm in terms of energy and RMSD_{all-atoms}. In MOLS, the best minimum and the next minimum energy conformations are considered and their RMSD_{all-atoms} with Scheraga conformation are tabulated. SaDE and RGA produce minimum energy conformations than MOLS and also their RMSD_{all-atoms} values are lesser than MOLS.

Table IV Performance Comparison with MOLS

Algorithm	Force Field	Energy (kcal/mol)	RMSD _{all-atoms} (Å°)
SaDE	ECEPP/3	-12.429	1.483
RGA	ECEPP/3	-12.429	1.485
MOLS	ECEPP/3	-10.1	1.8
		-9.0	1.5

Table 5 shows the performance comparison of SaDE and RGA with other GA variants reported (for ECEPP energy function) in terms of minimum energy, and number of function evaluations.

Table V Performance comparison of SaDE and, RGA with other GA variants

Algorithm	Selection	Crossover	Population Size	Number of Trials	Function Evaluations	Minimum Energy (kcal/mol)
SaDE	-NA-	Binomial	120	50	17697	-12.91
RGA	Tournament	Discrete	120	50	28771	-12.91
Dual DGA[30]	Tournament	One point	6400	-NR-	50000	-11
ECCA[31]	Pair-wise tournament	uniform crossover	-NR-	-NR-	140000	-7.372
PSA GA[9]	Pair-wise Random	One point	16	50	96000	-10.1
			64	50	96000	-9.75
Hybrid GA[32]	Proportionate	-NR-	50	50000	50000	-10.951
	Tournament					-11.118

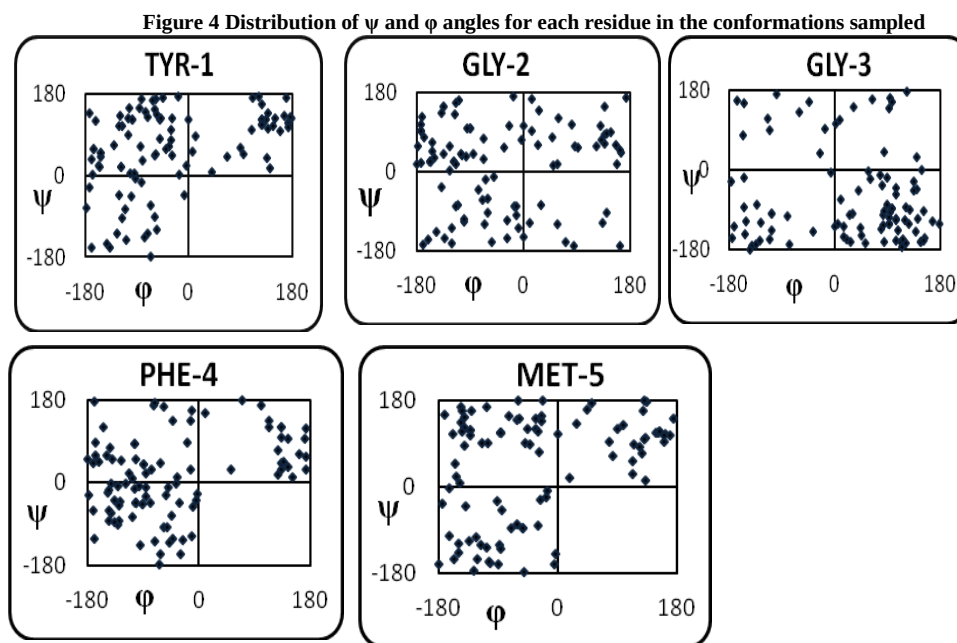
NA –Not Applicable

NR – Not Reported

The SaDE and RGA algorithms give the minimum energy conformations as compared to all other GA variants reported with minimum number of function evaluations.

D. Ramachandran Plot

A Ramachandran plot[33] (also known $[\phi, \psi]$ plot), is a way to visualize dihedral angles ψ against ϕ of amino acid residues in protein structure. It shows the possible conformations of ψ and ϕ angles for a polypeptide. Fig 4 shows the Ramachandran plot for the distribution of ψ and ϕ angles of each residue from the set of one hundred minimum conformations obtained using SaDE and RGA for ECEPP/2 with ω varying. The angles occur in the allowed or partially allowed regions of the Ramachandran diagram and also indicate the spread of conformations or the amount of conformational freedom present in the molecule of interest.



E. Clustering

Table 6 shows the clustering of hundred minimum energy conformations obtained using SaDE and RGA. Hierarchical cluster algorithm is used to cluster the conformations based on their RMSD values with Scheraga and 1PLW conformation. Several conformations are ranked to a same cluster, when the RMSD values of those conformations lie between two numerals.

Table VI Clustering Results

Scheraga				1PLW			
Cluster rank	Cluster Size	Minimum RMSD (Å°)	Average RMSD (Å°)	Cluster rank	Cluster Size	Minimum RMSD (Å°)	Average RMSD (Å°)
1	2	1.45	1.691	---	---	---	---
2	33	2.00	2.478	2	15	2.13	2.72
3	47	3.00	3.397	3	62	3.00	3.54
4	18	4.01	4.297	4	23	4.03	4.36

Fig 5(a) shows the main chain structure of Scheraga conformation. Figs 5(b)-(e) are the structure of conformations with minimum RMSD from each cluster. The conformations (d) and (e) tend to occur in the extended state.

Figure 5(a): Structure of Scheraga conformation. Figures. 5(b)-(e): Structure of conformations with minimum RMSD from each cluster

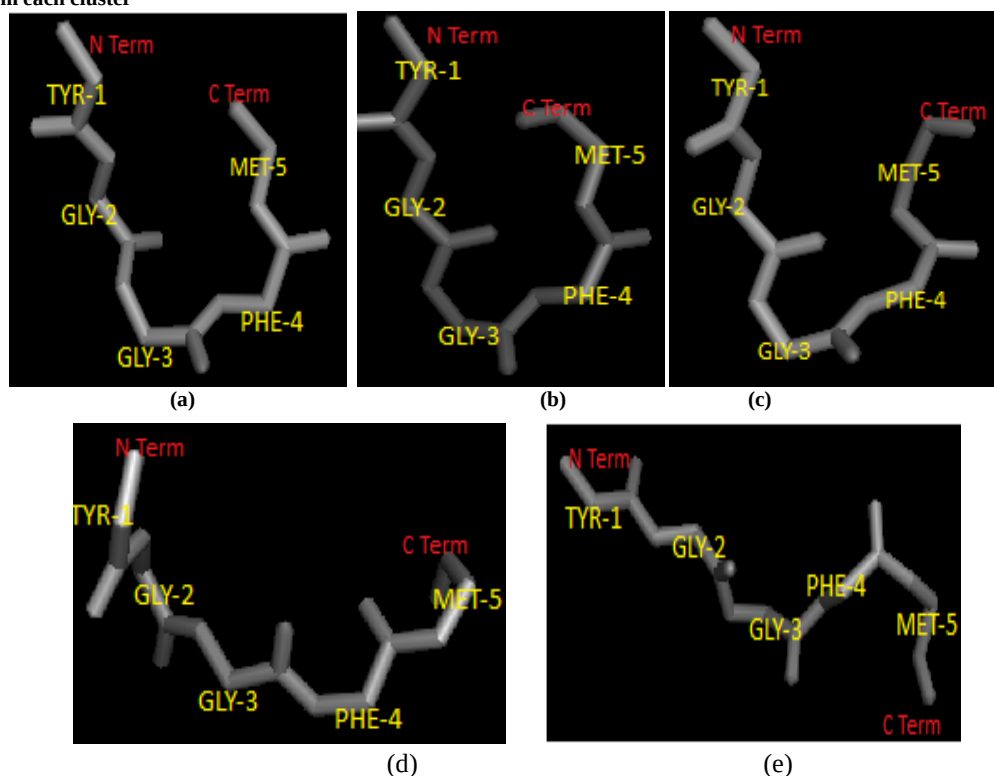


Fig 6(a) shows the main chain structure of 1PLW and Fig.s 6(b)-(d) are the structure of conformations with minimum RMSD from each cluster. The conformation 6(c) tends towards an extended one as in the earlier comparison.

Figure 6(a): Structure of 1PLW. Figures 6(b)-(d): Structure of conformations with minimum RMSD from each cluster.

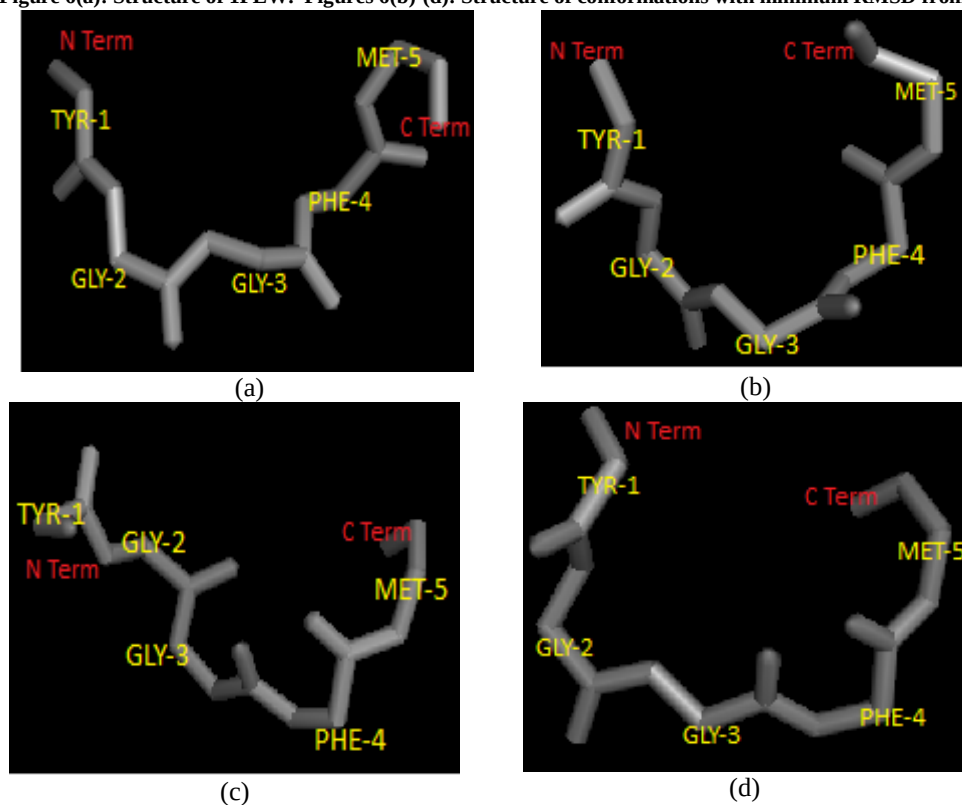
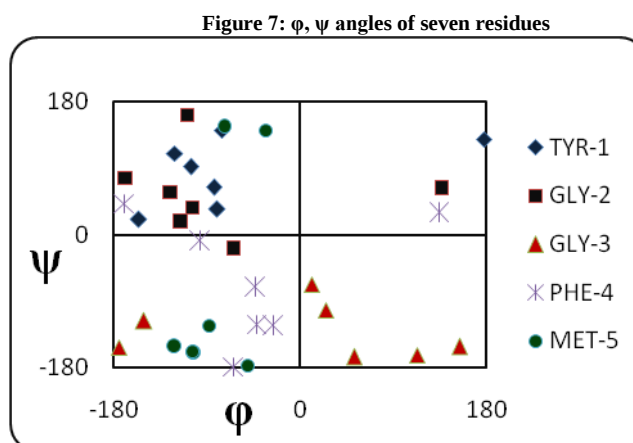


Fig 7 shows the distribution of phi-psi angles for minimum RMSD conformations from each cluster. The conformational angles show as in the molecular diagrams a tendency to occupy the extended state as well as the helical turn conformations.



V. Conclusion

In this paper, Protein Tertiary Structure Prediction using Evolutionary Algorithms (EAs) such as Self-Adaptive Differential Evolution (SaDE) and Real-coded Genetic Algorithm (RGA) with four recent versions of crossovers and mutation are reported. The algorithms are tested on a peptide Met-enkephalin for the energy functions ECEPP/2 and ECEPP/3 force fields. Among the four crossover and mutation combinations, RGA with discrete crossover and boundary mutation gives better performance. The best results obtained for ECEPP/2 with ω varying using SaDE and RGA with discrete crossover and boundary mutation are compared with native structure 1PLW and classical benchmark Scheraga conformation. Corresponding minimum RMSD values are 2.13 Å and 1.45 Å respectively. Statistical results of SaDE and RGA with discrete crossover and boundary mutation are compared in terms of best, mean, standard deviation of minimum energy and average number of function evaluations. Also, results of these algorithms show better performance with respect to minimum energy and number of function evaluations as compared with other GA variants reported. SaDE outperforms RGA with discrete crossover and boundary mutation in terms of number evaluations and success rate. Clustering a set of unique hundred best solutions obtained from SaDE and RGA with discrete crossover and boundary mutation gives seven independent clusters suggesting the robustness of these methodologies and the ability to explore the conformational space available and to populate the near native conformations. The angles of minimum RMSD conformations show that they have tendency to occupy the extended state as well as the helical turn conformations.

References

- [1] Hans-Joachim Böckenhauer, Dirk Bongartz, "Algorithmic Aspects of Bioinformatics," Natural Computing Series Springer-Verlag, 2007.
- [2] R. Unger and J. Moult, "Finding the Lowest Free Energy Conformation of a Protein is an NP-hard Problem: Proof and Implications", Bull., Math. Biol., 1993, 55, 1183.
- [3] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, W. Nadler, "Testing a new Monte Carlo Algorithm for Protein Folding", National Center for Biotechnology Information, 1998, 32, 52-66.
- [4] S. Kirkpatrick and C.D. Gelatt, "Optimization by Simulated Annealing", Science, 1983, 220, 671-680.
- [5] A. Shmygelska and H.H. Hoos, "An Improved Ant Colony Optimization Algorithm for the 2D HP Protein Folding Problem", Lect. Notes Comput. Sci., 2003, 2671, 400-412.
- [6] S. Schulze-Kremer, "Genetic Algorithms for Protein Tertiary Structure Prediction", Proc. European Conference on Machine Learning ECML-93, 1993, Vienna, Austria.
- [7] S. Schulze-Kremer, "Genetic Algorithms and Protein Folding Methods in Molecular Biology", Protein Structure Prediction: Methods and Protocols, 2000, 143, 175-222.
- [8] J. Ecker, M. Kupferschmid, C. Lawrence, A. Reilly, and A. Scott, "An application of nonlinear optimization in molecular biology", European Journal of Operational Research, 2002, 138, 452-458.

- [9] Tomoyuki Hiroyasu , Mitsunori Miki , Shinya Ogura , Keiko Aoi , Takeshi Yoshida , Yuko Okamoto, Jack Dongarra, “*Energy Minimization of Protein Tertiary Structure by Parallel Simulated Annealing using Genetic Algorithm*”, Proc. of Genetic and Evolutionary Computation Conference GECCO’02, 2002
- [10] Xiaolong Zhang, Ting Wang, Huiping Luo, Jack Y Yang, Youping Deng, Jinshan Tang, Mary Qu Yang, “*3D Protein structure prediction with genetic tabu search algorithm*”, BMC Systems Biology, 2010, 4 , S6.
- [11] Vinicius Tragante do O, Renato Tinos, “*Diversity Control in Genetic Algorithms for Protein Structure Prediction*”, J R Soc Interface, 2006, 3 ,139-151.
- [12] S. J. Weiner, P.A Kollman, D.T Nguyen, and D.A Case, “*An all atom force field for simulations of proteins and nucleic acids*”, Journal of Computational Chemistry, 1986,7 ,230–252.
- [13] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J States, S. Swaminathan, and M. Karplus, “*CHARMM: A program for macromolecular energy minimization and dynamics calculations*”, Journal of Computational Chemistry, 1983,4 ,187–217.
- [14] A. Momany, R.F. McGuire, A.W. Burgess and H.A. Scheraga, “*Energy Parameters in Polypeptides VII: Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions and Intrinsic Torsional Potentials for the Naturally occurring amino acids*”, J. Phys. Chem., 1975, 79, 2361-2381.
- [15] A. Nayeem, J. Vila, and H.A Scheraga, “*A comparative study of the Simulated Annealing and Monte-Carlo with minimization approaches to the minimum energy structures of polypeptides: [met]-enkephalin*”, Journal of Computational Chemistry, 1991, 12,594–605.
- [16] M. Shen, and K.F Freed, “*Long time dynamics of Met-enkephalin: Comparison of explicit and implicit solvent models*”, Biophysics Journal, 2002, 82 ,1791–1808.
- [17] K.Vengadesan, and N. Gautham, “*Energy landscape of met-enkephalin and leu-enkephalin drawn using mutually orthogonal latin squares sampling*”, Journal of Physical Chemistry B, 2004, 108, 11196–11205.
- [18] A.K. Qin, V. L. Huang, and P. N. Suganthan, “*Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization*”, IEEE transactions on Evolutionary Computation ,2009, 13 ,398-417.
- [19] Tomasz D. Gwiazda, “*Genetic Algorithms reference: Volume I Crossover for single-objective numerical optimization problems*”, TOMASZGWIAZDA E-BOOKS ,2006.
- [20] K. Deb, and R.B Agrawal, “*Simulated Binary Crossover for continuous search space*”, Complex Systems, 1995, 9,115-148.
- [21] Hsien-Chung Wu, “*Evolutionary Computation*”, Microsoft Academic Research ,1997.
- [22] K. Deb, and M. Goyal, “*A combined Genetic Adaptive Search (GeneAS) for engineering design*”, Computer Science and Informatics, 1996,26,30-45.
- [23] Eisuke Kita, “*Evolutionary Algorithms*”, InTech ,2011.
- [24] R. Storn and K. V. Price, “*Differential evolution- A simple and efficient heuristic for global Optimization over Continuous Spaces*”, J. Global Optim., 1997,11, 341–359.
- [25] J. J. Grefenstette, “*Optimization of Control Parameters for Genetic Algorithms*”, IEEE Transactions on Systems, Man and Cybernetics, 1996,16,122-128.
- [26] F. Herrera, M. Lozano and J.L. Verdegay, “*Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioral Analysis*”, Springer, Artificial Intelligence Review,1998,12, 265-319.
- [27] M. Jean Browman, Lucy M Carruthers, Karen L Kashuba, Frank A Momany, Marcia S Pottle, Susan P Rosen, and Shirley M Rumsey, “*ECEPP2: Empirical conformational energy program for peptides*”, Quantum Chemistry Program Exchange QCPE, Indiana University, 1983, pp: 855-4784.
- [28] F. Eisenmenger, U.H.E. Hansmann, S. Hayryan and C.K. Hu, “*[SMMP] A Modern Package for Protein Simulations*”, Comp. Phys. Comm. ,2001,138 ,192-212.
- [29] <http://dasher.wustl.edu/tinker/>
- [30] Tomoyuki HIROYASU, Mitsunori MIKI, Takashi IWAHASHI, Yuko OKAMOTO , “*Dual Individual Distributed Genetic Algorithm for Minimizing the Energy of Protein Tertiary Structure*”, SICE Annual Conference, 2003, Japan.
- [31] Amr Badr, Ibtehal M. Aref, Basma M. Hussien and Yosr Eman, “*Solving Protein Folding Problem using Elitism-Based Compact Genetic Algorithm*”, Journal of Computer Science, 2008,4,525-529 .
- [32] Laurence D. Merkle, Gary B. Lamont, George H. Gates, Jr. Ruth Pachter, “*Hybrid Genetic Algorithms for Minimization of a Polypeptide Specific Energy Model*” , Proceedings of the Third IEEE Conference on Evolutionary Computation, 1996, (Piscataway New Jersey: IEEE Service Center).
- [33] <http://dicsoft1.physics.iisc.ernet.in/rp/>

Acknowledgments

We thank the Dept of Biotechnology, Govt of India for the funding of the Centre of Excellence in Bioinformatics facility at MKU.