IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. III (May-Jun. 2016), PP 26-32
www.iosrjournals.org

Ab-Initio Protein Tertiary Structure Prediction Using Genetic Algorithm

Basem Ameen Moghram[1,*], Emad Nabil[1], Amr Badr[1]
[1]*(Department of Computer Science, Faculty of Computers and Information, Cairo University, Cairo, Egypt)*

Abstract : *Proteins are vital components of all living cells and play a critical role in almost all biological processes. Protein structure identification is a significant challenging problem in computational biology. The Protein's three-dimensional (i.e., tertiary) molecular structure reflects its proper function. Therefore, the identification of protein structure is a significant step towards understanding the protein's function which is an important role to synthesize new drugs and vaccine design; since many diseases are shown to result from malfunctioning of proteins. In this paper, we propose a new technique that is called Protein Tertiary Structure Prediction using Genetic Algorithm (PTSPGA) to predict the proteins structures based on their primary structure. We have developed a simple Elitist-based genetic algorithm technique for predicting the protein's tertiary structure based on Ab-Initio Empirical Conformational Energy Program for Peptides (ECEPP/3) Force Field Model. The proposed GA was applied to find the lowest free energy conformation of the protein's sequence. The experimental results indicate that the PTSPGA is reliable and very accurate, and it is a promising new technique that can help researchers in predicting the protein tertiary structure and can be used in the intelligent design of new vaccines.*

Keywords: *Ab-initio Protein Tertiary Structure Prediction, ECEPP Force Field, Energy Function, Genetic Algorithm, Protein Folding, PTSPGA, Vaccine Design.*

I. Introduction

The Proteins are essential components of all living cells. Proteins play a key role in all biological processes and perform a vast array of functions within living organisms. Proteins are made of amino acid building blocks that are arranged in a linear chain and joined by amide bonds, frequently known as peptide bonds; therefore, proteins are otherwise called *polypeptides.*

There are twenty different standard amino acids, which are divided into various classes on the basis of its size and the other physical and chemical properties. This particular folded shape enables proteins to perform a specific biological function provided that the folding is in its native structure, which is the correct three-dimensional structure of the protein[1].

The primary structure of protein is a chain of amino acids(Fig. 1a). The protein secondary structure is regularly repeating local structures stabilized by hydrogen bonds. The most common secondary structure examples are the alpha helix (Fig. 1b), beta sheet and turns. The way that secondary structure elements pack together to form the entire fold of the protein atoms is called the protein's tertiary structure (Fig. 1c). Some proteins have two or more separate polypeptide chains, which may be identical or different. The organization of these proteins subunits in three-dimensional complexes constitutes the quaternary structure (Fig. 1d).

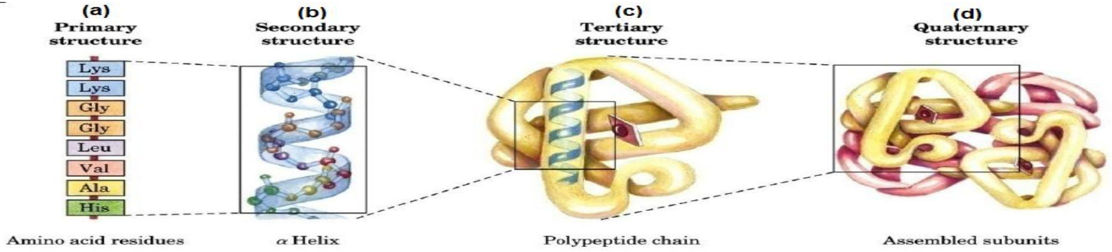


Figure1: Hierarchy of protein structure example

The understanding of the proteins functions is an important role to treat diseases, and synthesize new drugs and vaccines; since some diseases are shown because of proteins folding gone wrong. Whereas proteins functions cannot be understood without determining their three-dimensional structure.

Protein structure prediction problem is one of the major challenges in Bioinformatics and Molecular Biology[2]. The classical techniques for structure prediction of proteins are X-ray crystallography[3], [4] and Nuclear Magnetic Resonance (NMR) [5], [6]. But, these methods are expensive, time-consuming, and laborious, while computational methods are capable of reducing the cost, time, and saving the development resources. In recent years, the expanding scope of the practical knowledge and the improvement of intelligent methodologies have resulted in the expansion of computational techniques.

The computational methods for protein structure prediction can be generally classified into three approaches: homology, threading, and *ab-initio* methods [2]. Homology modeling, also known as comparative modeling of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure within the Protein Data Bank (PDB) [7] of a related homologous protein (the "template") [8]. Protein threading, also known as fold recognition, is a method of protein modeling, which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure. It differs from the homology modeling method of structure prediction as it (protein threading) is used for proteins which do not have their homologous protein structures deposited in the Protein Data Bank (PDB), whereas homology modeling is used for those proteins which do [9]. Ab-initio- or de-novo- protein modeling strategies do not rely on known structures in the PDB. Instead, they predict the three-dimensional structure of proteins given only their primary sequences. Homology and threading approaches are limited to predict the protein' structure which belong to protein families with known structures [2]. Where the ab-initio modeling can predict the structure of any unknown protein's sequence.

Ab-initio protein structure prediction has two types of modeling, which are Hydrophobic-Polar (HP) modeling and force field modeling. The HP model simplifies the protein by assigning each amino acid to be a point in a 2D representation (H,P) that is either hydrophobic (H) or polar (P). According to this model, the most stable structure is the one with the hydrophobic amino acids lying in its core. [10]. Ab-initio force field models use an energy objective function that evaluates the structure of a protein. This objective function attempts to represent the actual physical forces and chemical reactions occurring in a protein. Protein atoms are modeled as points in 3D representation. The bonds among atoms in proteins are modeled as Newtonian springs. The protein's tertiary structure is the conformation with the lowest free energy, according to the laws of physics and Anfinsen Hypothesis [11]. The energy function is usually based on molecular mechanics and force field components such as bond lengths, bond angles, dihedral angles, van der Waals interactions, electrostatic forces.

In this paper, we propose a new technique for predicting the protein tertiary structure based on its sequence. This technique is called Protein Tertiary Structure Prediction using Genetic Algorithm (PTSPGA). We use the Empirical Conformational Energy Program for Peptides (ECEPP/3) [12] as an objective function, while the protein's energy was evaluated using the Simple Molecular Mechanics for Proteins (SMMP) package [13]. Experiments were performed on the Met-enkephalin protein and some other proteins. The experimental results indicate that the PTSPGA is reliable and very accurate in predicting the protein tertiary structure.

II. Materials And Methods

i. Protein Conformation Representation

The amino acids of a protein chain are covalently joined by amide bonds, often called peptide bonds: for this reason, proteins are also known as polypeptides. Chemically, the peptide bond is a covalent bond that is formed between a carboxylic acid and an amino group by the loss of a water molecule [14]. The end of a polypeptide with the free amino group is known as the amino terminus (N terminus), that with the free carboxyl group as the carboxyl terminus (C terminus).

The peptide bond has a partial double-bond character, which means that the three non-hydrogen atoms that make up the bond (the carbonyl oxygen O, the carbonyl carbon C and the amide nitrogen N) are coplanar, and that free rotation about the bond is limited [14] (Figure. 2). The other two bonds in the basic repeating unit of the polypeptide backbone (i.e., main chain), the N–Ca and Ca–C bonds (where Ca is the carbon atom to which the side chain is attached), are single bonds and free rotation is permitted about them provided that there is no steric interference from, for example, the side chains. The angle of the N–Ca bond to the adjacent peptide bond is known as the **phi torsion angle (**□), and the angle of the C–Ca bond to the adjacent peptide bond is known as **the psi torsion angle (**□**)**, and the backbone torsion angle of the C'-N bond to the relative residue is named the **Omega torsion angle (**□**)**[14] (Figure. 2).

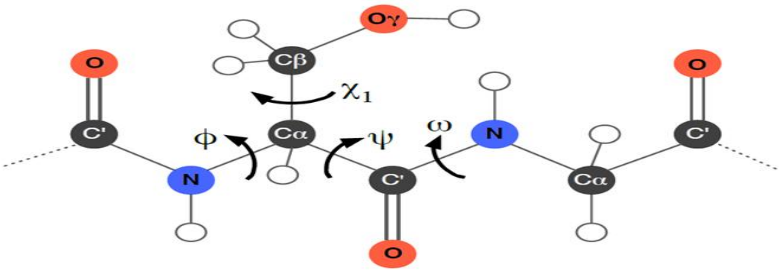


Figure 2: Overview of important torsion angles (φ, ψ, ω, and χ1) in proteins.

The side chain of amino acid branch out of the backbone chain from the Cα atom and have additional degrees of freedom, called the **Chi torsion angles (□$_i$).** The side chain torsion angles are denoted as χ_i for each successive bond along that chain (i.e., *i* is a number between 1 and 6, and it represents the number of χ angles for each amino acid). The overall protein structure can be described by its backbone and side-chain torsion angles. Therefore, the conformation of a protein is represented as a sequence of the torsion angles.

ii. Protein Tertiary structure prediction

In the next subsections, the proposed genetic algorithm of protein tertiary structure prediction will be described.

i) The Proposed Genetic Algorithm

Genetic Algorithms (GA) primarily motivated by the biological theory of evolution, and it was originally developed by J. Holland in 1975 [15]. The GA is a search heuristic that mimics the process of natural selection. GA is based on the operations of population reproduction and selection to achieve optimal results. Through artificial evolution, successive generations search for fitter adaptations in order to solve a problem. Each generation consists of a population of chromosomes, representing a series of candidate solutions (called individuals) to an optimization problem, generally evolves toward better solutions by applying genetic operations of recombination and mutation to create a new offspring population from the current population. The process evolution usually starts from a population of randomly generated individuals or, using domain background knowledge. The generation process repeats for many generations with the aim of maximizing the objective Function (also called fitness) of the chromosomes that are evaluated each generation. Usually, the GA termination condition is set to the maximum number of generations.

ii) Genetic Representation

In a GA, a population of chromosomes, representing a series of candidate solutions (called individuals) to an optimization problem, generally evolves toward better solutions. In this proposed algorithm, the chromosome is a representation of the protein torsion angles used to build the tertiary (3D) structure of a protein. A chromosome consists of several genes (see Figure 3). The chromosome length represents the protein sequence length. Each gene is represented by an array of real values. These real values are between (-180.0, 180.0) degree, which represents the amino acid torsion angles. The gene's array length represents the number of the gene's torsion angles (i.e., amino acid residue). Generating the proteins' conformations is initially done by randomly selecting the values of the torsion angles based on the domain knowledge of torsion angles.

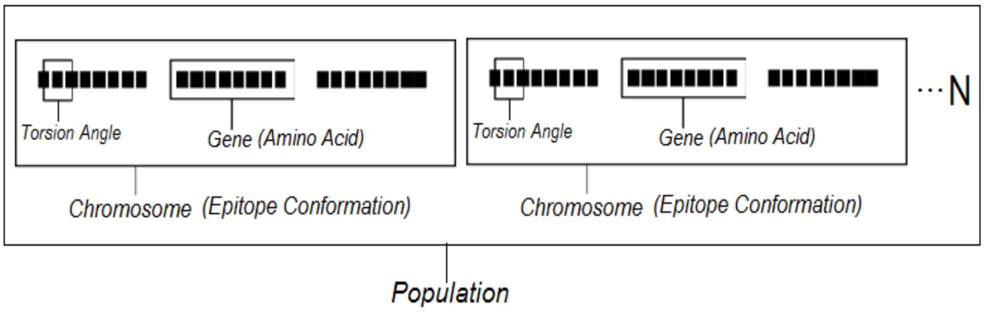


Figure 3: Protein Genetic Representation

iii) The Objective Function

In this proposed GA, the objective fitness function used is the protein conformation energy function. The energy function evaluates the protein structure based on the force fields and the molecular mechanics of the torsional angles. The energy of the protein is calculated using a protein's energy evaluator called the Empirical Conformational Energy Program for Peptides (ECEPP/3) [12], which is implemented as a part of the Simple Molecular Mechanics for Proteins (SMMP) package [13].

iv) The proposed Algorithm

In the following flow-chart, the proposed Elitism-based genetic algorithm for predicting protein's tertiary structure is summarized:

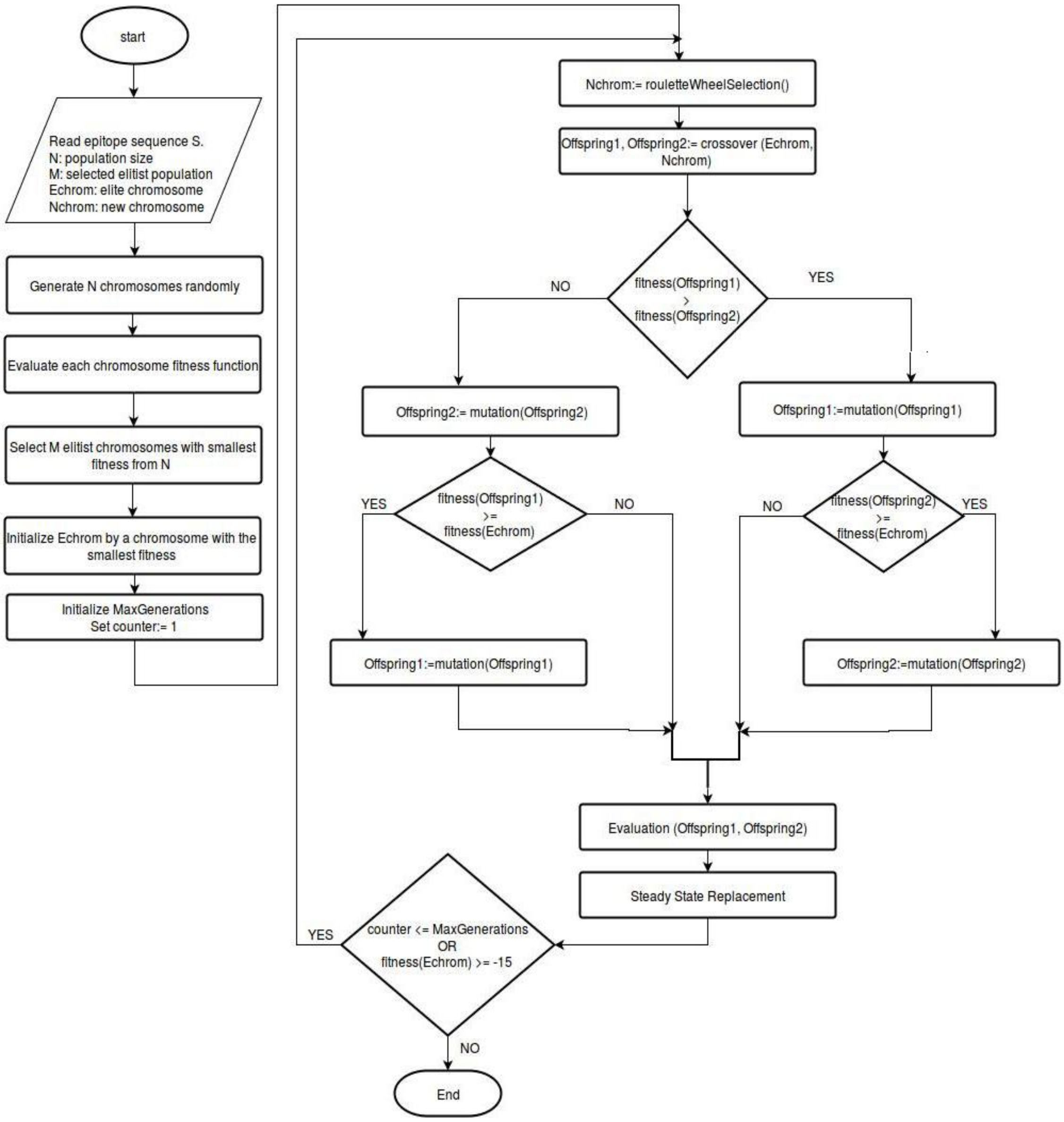

start
Read epitope sequence S.
N: population size
M: selected elitist population
Echrom: elite chromosome
Nchrom: new chromosome
Generate N chromosomes randomly
Evaluate each chromosome fitness function
Select M elitist chromosomes with smallest fitness from N
Initialize Echrom by a chromosome with the smallest fitness
Initialize MaxGenerations
Set counter:= 1
Nchrom:= rouletteWheelSelection()
Offspring1, Offspring2:= crossover (Echrom, Nchrom)
fitness(Offspring1) > fitness(Offspring2)
NO
YES
Offspring2:= mutation(Offspring2)
Offspring1:=mutation(Offspring1)
fitness(Offspring1) >= fitness(Echrom)
YES
NO
fitness(Offspring2) >= fitness(Echrom)
NO
YES
Offspring1:=mutation(Offspring1)
Offspring2:=mutation(Offspring2)
Evaluation (Offspring1, Offspring2)
Steady State Replacement
counter <= MaxGenerations OR fitness(Echrom) >= -15
YES
NO
End

Figure 4: The proposed Elitism-based GA for predicting protein's tertiary structure.

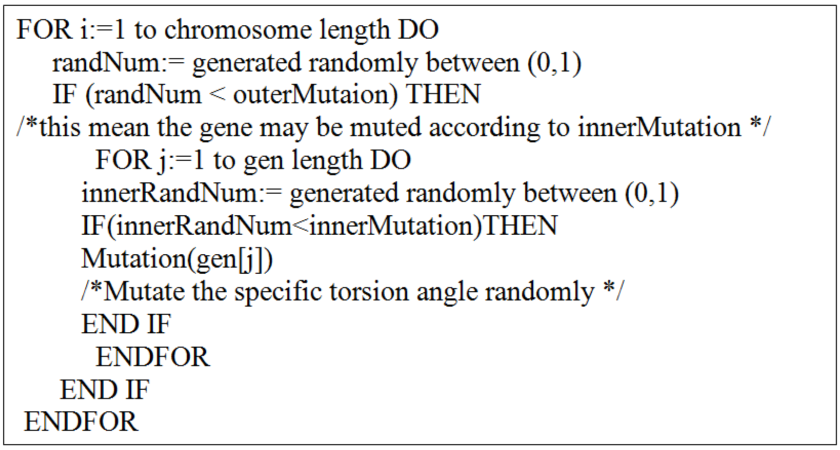
iii. Elitism & Genetic Operators

Elitist based strategy in GA implies that the best or the most fit solutions found are used to produce the next generations; and this is done by the competition between the offspring and their parents to choose the fitter one in the next generations[16]. Elitism steps are applied after evaluating the fitness function for the whole population and selecting the most fitter chromosomes in the selection process. Elitist GA performs better than the standard GA, since elitism keeps the best-found chromosomes for the next generations and it is applied only to "strong" chromosomes.

In the proposed GA, the genetic operators that were employed are: selection, crossover, mutation, and replacement. The selection operation uses roulette-wheel selection strategy that depends on selecting at random an individual from the elitist population. One-Point Crossover operation is applied to a pair of chromosomes, one of them is provided by roulette-wheel selection, and the other one is the Elite chromosome with the best fitness function gained during the successive generations. The crossover point along the chromosome (k) is selected at random between 2 and N-1, where all the genes between k and N are exchanged to produce the new two offsprings.

The operation of mutation is applied at random to the selected chromosome' genes base on the mutation probability. The mutation process is applied on two stages. The first stage "outer-mutation" is done on the level of the genes; by selecting the gene that may be mutated. While the second stage "inner-mutation" is done on the level of the genes' torsion angles; by mutating the randomly selected torsion. The two-level mutation speeds up the processing time, because it doesn't lose the existing good torsion angles by not mutating all the angles in the gene and mutating only the randomly selected torsion angles based on the inner mutation probability.

The replacement process uses the elitist steady state strategy to keep the best chromosomes if they either parents or offsprings for the next generations. The following pseudo code summarizes the mutation operation:



```
FOR i:=1 to chromosome length DO
    randNum:= generated randomly between (0,1)
    IF (randNum < outerMutaion) THEN
/*this mean the gene may be muted according to innerMutation */
          FOR j:=1 to gen length DO
        innerRandNum:= generated randomly between (0,1)
        IF(innerRandNum<innerMutation)THEN
        Mutation(gen[j])
        /*Mutate the specific torsion angle randomly */
        END IF
          ENDFOR
     END IF
 ENDFOR
```

Figure 4: A pseudo code summarizes the mutation operation

III. Results And Discussion

The proposed technique is implemented and tested using Java platform with JDK1.8.0_20 on a computer with Intel Core 2 duo 2.2 GHz processor and a memory of 3 GB, running on Microsoft Windows 7 operating system.

i. GA Performance and Parameter Settings

The proposed genetic algorithm is applied to minimize the free energy conformation of the protein's sequence. The GA applied is an elitist based strategy with the roulette-wheel selection operation and elitist steady state replacement. One-point crossover was applied and two levels of mutation were performed.

In our experiments, we have tested the proposed GA on different proteins, specially the small proteins with a length less than 33 amino acid, because of the limitation of the SMMP tool that was used to evaluate the energy of the proteins. The best genetic parameters setting are differ from protein to another, so we have set them as an average values that got good results with all the test examples; according to their lengths. In the standard genetic algorithm, there are many genetic parameters needed such as: population size, crossover probability, and mutation rate.

In the proposed GA for predicting the protein tertiary structure, the parameters employed in the experiments on the antigenic proteins (epitopes) are: population size N=150, the selected elitist population M=20 to apply the GA operations, and the crossover probability is equal to 1. The multistage strategy of mutation and the difference in protein's sequence length resulted in different mutation probabilities. The outer mutation probability for protein's nonamers subsequence(9 amino acid residues) was set to 0.18, for proteins with a length of 12 to 33 amino acid was set to 0.15. The inner mutation probability for nonamers was set to 0.25, other proteins with a length of 12 to 33 amino acid, the probability of mutation was set to 0.2. Larger proteins need to adjust the GA parameters in according to their length and to the ECEPP force-field tools limitations, since proteins with more than 100 residues in length couldn't be handled by some ECEPP tools.

The number of max generations also differs according to the length of the protein's sequence in the proposed GA. From the experiments, we found that the best max generations for the nonamer subsequence of the protein is 700 generations and 1500 generations for the proteins with ranging length between 12 and 33 amino acid. Some proteins has reached their local solution earlier after 200 generations or less, as well some nonamers reach their solution after only 50 generations, but the others proteins need more generations than those proteins. Table 1 clarifies some random examples.

Table 1: Some examples at random shows the performance of the proposed GA.

S.No.	Protein's Sequence or Nonamer Subsequence(9-mer)	Number Of Iterations	Conformation Energy kcal/mol
1	EPGPGPGFR	49	-19.644
2	LNGPGPGSP	99	-22.984
3	NFLQSRPEP	172	-22.32
4	KPGQPPRLL	262	-27.914
5	PSQQQPQEQ	368	-48.175

6	EGPEFFDQE	467	-28.28
7	NANPDCKTI	590	-20.914
8	RPTLAFLQD	662	-30.518
9	EHDLERGPPGPRRPP	250	- 53.518
10	DMTPADALDDFDL	440	-17.711
11	FSQILPDPLKPTKRS	682	-18.711
12	MRSPVFTDNSSPPVV	713	-20.837
13	RPFFHPVGEADYFEYHQEGGPDGEPD	802	-41.817
14	PLGFFPDHQLDPAFGANSNNPDWDFNP	945	-58.989
15	DPHLPTLLLGSSGSGGDDDDPHGPVQLSYYD	1187	-31.455

The inner mutation operation in the proposed GA plays a very important performance role. We test the proposed algorithm with and without the inner mutation operation. We found that the inner mutation operation sped up the processing time of the proposed algorithm rapidly from four to ten times and even higher than the proposed GA without the inner mutation operation. E.g., the conformation of the protein sequence "EHDLERGPPGPRRPP" reached the energy (-91.957 kcal/mol) after only 500 iterations. But, without the inner mutation operation, the conformation energy reached (-55.787 kcal/mol) after 4000 iterations. While some proteins reach more than 5000 iterations, but we could not reach negative energies for the proteins conformation without the inner mutation; whereas we reached their optimal negative energies after only 500 iterations when applying the proposed algorithm with the inner mutation.

Determining the peptide' end groups, N- and C-terminal residues in the protein tertiary structure prediction using ab-initio ECEPP force-field models is one of the most important parameters. ECEPP always assumes that the peptide chain has two end groups, so that, the two end groups were set as 'COOH' for the C-terminal group and 'NH2'for the N-terminal group in the SMMP package that used to evaluate the peptide conformation energy. As well as, determining the amino acids torsion backbone angles and its side chain angles clearly is another important role in predicting the tertiary structure, because torsion angles vary in number and form from one amino acid residue to another, especially the side-chain torsion angles that varied an amino acid from the others.

ii. Comparison with Benchmark Methods

As far as we know, most of the previous published work used the Met-enkephalin protein as an evaluation protein for their techniques. Therefore, the proposed algorithm was tested on Met-enkephalin protein; which is a small protein with 5 residues (YGGFM). We reached the conformation energy of (-9.5533 kcal/mol) after 5336 iterations and it is a very good result compared to the previous results, since most of the previous algorithms could not reach this result with the same iterations (Table 2).

The algorithm genetic parameters need to be configured specially for the Met-enkephalin protein , because it is very small protein and its parameters are different. The proposed GA parameters was configured as follows: the outer mutation probability was set to 0.6 ; the inner mutation probability was set to 0.06; the population size was set to 200 ; the *M* elitist chromosomes was set to 100; and crossover was set to 1 with a variable omega torsion angles. Table 2 below compares our result with the previous work depending on the Met-enkephalin protein .

According to the experimental results, we found that the Met-enkephalin protein is not a suitable protein to test the prediction algorithm, since we reached the solution after thousands of iterations while we reached the solution for larger proteins in hundreds iteration only as reported in Table 1.

Table 2: Performance comparison of PTSPGA with other GA variants

Algorithm	**Selection**	**Crossover**	**Population Size**	**Number of Trials**	**Function Evaluations**	**MinimumEnergy (kcal/mol)**
PTSPGA	roulette-wheel	One point	100	10	5336	-9.5533
SaDE [17]	-NA-	Binomial	120	50	17697	-12.91
RGA [17]	Tournament	Discrete	120	50	28771	-12.91
Dual DGA[18]	Tournament	One point	6400	-NR-	50000	-11
ECGA[19]	Pair-wise	uniform	-NR- `	-NR-	140000	-7.372
	Tournament	Crossover				
PSA GA[20]	Pair-wise Random	One point	16	50	96000	-10.1
			64	50	96000	-9.75
Hybrid GA[21]	Proportionate	-NR-	50	50000	50000	-10.951
	Tournament					-11.118

*NA–NotApplicable *NR – Not Reported

IV. Conclusion

We proposed an elitist based genetic algorithm technique for predicting the protein tertiary structure based on the ab-initio force- field prediction model. The proposed GA applied to find the lowest free energy conformation of the protein's sequence. The conformation energy is calculated using an energy evaluator called the Empirical Conformational Energy Program for Peptides (ECEPP/3).

The experimental results show that the proposed technique (PTSPGA) is reliable and very accurate. It is our hope that the proposed new prediction technique will be an addition to the rest of the techniques available for researchers and scientists for protein structure prediction and the rational design of intelligence vaccines.

Acknowledgment

This work was supported as a part of a Master's Degree Scholarship from the Ministry of Higher Education and Scientific Research of Yemen to Basem Ameen Moghram.

References

[1] H.-J. Bockenhauer and D. Bongartz, *Algorithmic Aspects of Bioinformatics*. 2007.

[2] C. A. Floudas, “Computational methods in protein structure prediction,” *Biotechnology and Bioengineering*, vol. 97, pp. 207–213, 2007.

[3] L. Bragg, *The Development of X-Ray Analysis*, 1st Editio. London, U.K: G. Bell, 1975.

[4] T. L. Blundell and L. H. Johnson, *Protein Crystallography*. New York: Academic Press, 1976.

[5] K. Wüthrich, *NMR of Proteins and Nucleic Acids*. New York: Wiley, 1986.

[6] E. T. Baldwin, I. T. Weber, R. St Charles, J. C. Xuan, E. Appella, M. Yamada, K. Matsushima, B. F. Edwards, G. M. Clore, and A. M. Gronenborn, “Crystal structure of interleukin 8: symbiosis of NMR and crystallography.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 88, no. 2, pp. 502–6, Jan. 1991.

[7] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank.,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[8] Y. Zhang and J. Skolnick, “The protein structure prediction problem could be solved using the current PDB library.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 4, pp. 1029–1034, 2005.

[9] J. U. Bowie, R. Lüthy, and D. Eisenberg, “A method to identify protein sequences that fold into a known three-dimensional structure.,” *Science*, vol. 253, no. 5016, pp. 164–170, 1991.

[10] N. Mansour, F. Kanj, and H. Khachfe, “Enhanced genetic algorithm for protein structure prediction based on the HP model,” *Search Algorithms Appl*, 2011.

[11] C. B. Anfinsen, “Principles that govern the folding of protein chains.,” *Science*, vol. 181, no. 96, pp. 223–230, 1973.

[12] G. Nemethy and K. Gibson, “Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-,” *J. …*, pp. 6472–6484, 1992.

[13] J. H. Meinke, S. Mohanty, F. Eisenmenger, and U. H. E. Hansmann, “SMMP v. 3.0-Simulating proteins and protein interactions in Python and Fortran,” *Comput. Phys. Commun.*, vol. 178, no. 6, pp. 459–470, 2008.

[14] G. a Petsko and D. Ringe, “From Sequence to Structure,” *Protein Struct. Funct.*, pp. 2–48, 2004.

[15] J. H. Holland, *Adaptation in Natural and Artificial Systems*. 1975.

[16] D. Thierens, “Selection schemes, elitist recombination, and selection intensity,” *ICGA*, pp. 152–159, 1997.

[17] S. Sudha, S. Baskar, and S. Krishnaswamy, “PROTEIN TERTIARY STRUCTURE PREDICTION USING EVOLUTIONARY ALGORITHMS,” *Int. J. Emerg. Technol. Comput. Appl. Sci. (IJETCAS)*, pp. 338–348, 2013.

[18] T. Hiroyasu, M. Miki, T. Iwahashi, and Y. Okamoto, “Dual individual distributed genetic algorithm for minimizing the energy of protein tertiary structure.,” *SICE Annu. Conf. Japan*, 2003.

[19] A. Badr, I. M. Aref, B. M. Hussien, and Y. Eman, “Solving protein folding problem using elitism-based compact genetic algorithm,” *J. Comput. Sci.*, vol. 4, no. 7, pp. 527–531, 2008.

[20] Tomasz D. Gwiazda, *Genetic Algorithms reference: Volume I Crossover for single-objective numerical optimization problems*. TOMASZGWIAZDA E-BOOKS, 2006.

[21] L. D. Merkle, G. B. Lamont, G. H. Gates, and R. Pachter, “Hybrid genetic algorithms for minimization of a polypeptide specific energy model,” in *Proceedings of IEEE International Conference on Evolutionary Computation*, pp. 396–400.