Solving Protein Folding Problem Using Hybrid Genetic Clonal Selection Algorithm

Adel Omar Mohamed[†] *and Abdelfatah A. Hegazy*[††], **Amr Badr**[†††]

College of Computing & Information Technology, Arab Academy

Abstract:
Alzheimer's disease, Cystic fibrosis, Mad Cow disease, an inherited form of emphysema, even many cancers. Recent discoveries show that all these apparently unrelated diseases result from protein folding gone wrong. As though that weren't enough, many of the unexpected difficulties biotechnology companies encounter when trying to produce human proteins in bacteria also result from something amiss when proteins fold.
Protein folding problem is the process of predicting the optimal 3D molecular structure of a protein, or tertiary structure, which is an indication of its proper function.
Approach, An enhancement over persistent clonal selection algorithm was made to minimize the energy of proteins by adding crossover function from Genetic algorithm (GA). Energy was calculated using the Empirical Conformational Energy Program for Peptides (ECEPP) package.
Results: Experiments were performed on the Met-Enkephalin protein. The enhanced algorithm reached energy of -20.919 in 10 generations surpassing the Clonal Selection Algorithm which reached the same energy in 30 generations. A comparison was also made with the Genetic Algorithm (GA) which reaches this energy in 1000 generations. Results show that the enhanced algorithm is superior to Clonal Selection algorithm and GA.

Keywords:
Protein Folding, genetic algorithms, artificial immune system, Clonal Selection Algorithm, Met-Enkephalin.

1. INTRODUCTION

Proteins are fundamental components of all living cells. The bacteria that infect us, the plants and animals we eat, the hemoglobin that carries oxygen to our tissues, the insulin that signals our bodies to store excess sugar, the antibodies that fight infection, the actin and myosin that allow our muscles to contract, and the collagen that makes up our
Tendons and ligaments (and even much of our bones) are all examples of proteins. To make proteins, ribosomes string together amino acids into long, linear chains. Like shoelaces, these chains loop about each other in a variety of ways (i.e., they fold). But, as with a shoelace, only one of these many ways allows the protein to function properly. Yet lack of function is not always the worst scenario. Recent discoveries have shown that some diseases (Alzheimer's disease, Cystic fibrosis, Mad Cow disease, and many cancer types) are the result of misfolded proteins. Also, protein misfolding is behind many of the unexpected difficulties biotechnology companies encounter when trying to produce human proteins in bacteria. A misfolded protein can actually poison the cells around it, so misfolded protein could be worse than a normally folded one.
The prediction of molecular structure (polypeptide's native conformation) of a protein given only its amino acid sequence is not an easy task, but has numerous potential applications [1]. This structure prediction problem is commonly referred to as the protein folding problem. Efforts to solve it nearly always assume that the native conformation corresponds to the global minimum free energy state of the system. Given this assumption, a necessary step in solving the problem is the development of efficient global energy minimization techniques. This is a difficult optimization problem because of the non-linear and multi-modal nature of the energy function.
In this paper, we combine two methodologies which are Genetic Algorithms and Artificial Immune System (AIS), so as to automatically produce a system for protein folding problem.
The paper is organized as follows: in the next two sections we provide an overview of the clonal selection algorithm and the genetic algorithm. In section [4] we present our proposed hybrid algorithm between GA and AIS that will be tested on protein folding problem described in section [5] which speaks about the hybrid algorithm. The testing and delineated, followed by concluding remarks in Section [6].

2. The genetic algorithm (GA)

The standard genetic algorithm [2], [3] can be summarized as follows.
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 [4]. Genetic algorithm is a kind of local search method in essentially, Its main idea is, the algorithm generated a number of feasible solutions of the problem

Manuscript received December 5, 2010
Manuscript revised December 20, 2010

randomly (ie. chromosomes) in the solution space of a problem, The algorithm starts from these initial feasible solutions, and calculates the fitness level of each chromosome according to the objective function, through the crossover ,mutation, selection and other operations so that chromosome populations evolving from generation to generation and eventually converge to an "optimal solution." Genetic algorithm is a general-purpose optimization algorithm; its coding and genetic manipulations of the technologies are relatively simple, suitable for solving combinatorial optimization problems. It has two significant characteristics: First, the global solution space of the search capability; second is the implicit parallelism in the search [1]. Although the genetic algorithm has been shown be able to converge to the global optimum under certain conditions, but these conditions are not met yet in the real world. And the traditional genetic algorithm has some drawbacks like its weak local search ability and convergence of the short comings too fast or too slow, but the algorithm is still an excellent algorithm, and with the development of computer technology, genetic algorithms become a research hotspot. The shortcomings of the traditional genetic algorithm have been improved by many excellent scholars and experts, and make it success in the area of machine learning, pattern recognition, image processing, optimizing control and so on.

The traditional Genetic algorithms are often overlooked a critical part, which can impact the speed and quality of whole evolution, that is, the initial solutions (chromosomes)is generated randomly, the quality of these randomly generated initial solutions is often not high or too focused on the solution space of some area, even through the latter part of the continuous improvement of genetic operators, it fall into local minima easily, leading to the global optimal solution cannot be found. For this shortcoming of Algorithm, this article introduced species similarity and immunological principles of the algorithm in the initial stages ,so it can produce some high-quality solutions in initial phase, to improve the convergence speed and global search capabilities.

The genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

In a genetic algorithm, a population of strings (called chromosomes or the genotype of the genome), which encode candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem, evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached.

The below is standard Genetic algorithm.

Begin

Initialize the initial population p(t) randomly;

For I =1 to MaxGenerations do;

Begin

Evaluate Fitness of individuals;

Select new parents for reproduction;

Crossover;

Mutation;

Replace old populations with new populations;

End

End

3. The clonal selection algorithm

The standard clonal selection algorithm CLONALG [5], [6], [7], [8], [9] can be summarized as follows.

The clonal selection algorithm is used by the natural immune system to define the basic features of an immune response to an antigenic stimulus. It establishes the idea that only those cells that recognize thee antigens are selected to proliferate. The selected cells are subject to an affinity maturation process, which improves their affinity to the selective antigens [8].

In Artificial immune systems, Clonal selection algorithms are a class of algorithms inspired by the clonal selection theory of acquired immunity that explains how B and T lymphocytes improve their response to antigens over time called affinity maturation. These algorithms focus on the Darwinian attributes of the theory where selection is inspired by the affinity of antigen-antibody interactions, reproduction is inspired by cell division, and variation is inspired by somatic

hypermutation. Clonal selection algorithms are most commonly applied to optimization and pattern recognition domains, some of which resemble parallel hill climbing and the genetic algorithm without the recombination operator. Artificial immune system solving complex machine learning tasks, like pattern recognition and multimodal optimization

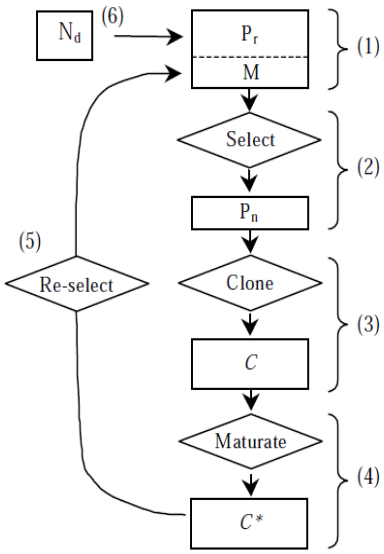

Figure 1: Block diagram of the clonal selection algorithm

After discussing the clonal selection principle and the affinity maturation process, the development of the clonal selection algorithm (CSA) is straightforward. The main immune aspects taken into account were: maintenance of the memory cells functionally disconnected from the repertoire, selection and cloning of the most stimulated cells, death of non-stimulated cells, affinity maturation and re-selection of the clones with higher affinity, generation and maintenance of diversity, hypermutation proportional to the cell affinity. Below is the standard clonal selection algorithm.

Begin

t=0;

Initialize the initial population P (t) randomly;

For I=1 to MaxGenerations do;

Begin

Evaluate affinity of individuals;

Generate clones of a subset of the antibodies in N with the highest affinity; (The number of clones for an antibody is proportional to its affinity)

HyperMutation;

Metadynamics; (randomize rest of populations other than selected)

Replace mutated antibodies with their parents;

End

End

4. The proposed hybrid algorithm

The proposed hyper genetic clonal selection algorithm can be summarized as follows.
The new proposed algorithm (Hybrid genetic clonal selection algorithm) modified standard clonal selection algorithm by import crossover function from Genetic algorithm [1]. We have imported the crossover operator from the genetic algorithms in order to increase the exploration of the landscape and to add a recombination operator in the clonal selection algorithm.

Begin

t=0;

Initialize the initial population P (t) randomly;

For I=1 to MaxGenerations do;

Begin

Evaluate affinity of individuals;

Generate clones of a subset of the antibodies in N with the highest affinity; (The number of clones for an antibody is proportional to its affinity)

HyperMutation;

Crossover;

Metadynamics; (randomize rest of populations other than selected)

Replace mutated antibodies with their parents;

End

End

5. The Protein folding problem

In this section, we test our new hyper clonal selection algorithm on met-Enkephalin protein
The below figure describe the amino-acid of Met-Enkephalin protein.

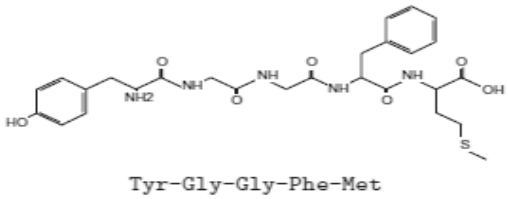

Figure 2: amino-acid of Met- Enkephalin protein

Enhancement over clonal selection algorithm: In this study, the enhanced algorithm proposed in solving protein folding problem with the addition of two modifications: Crossover, and keeping the best solution so far. The first modification is the addition of crossover. Crossover will be performed by adding a secondary tournament to each cycle that compares the performance of the current champion string with a new version of itself. If the new version wins, it replaces the old champion as the elite string for the next tournament. Note that our implementation of crossover function. This allows periodic sampling of individuals around the champion independent of the current state of the genome probability vector. We can more formally describe this operation by modifying the standard clonal selection algorithm to contain crossover function.

We also found that considering the probability vector as the final solution is not the optimal solution, so the second and final modification is to consider the final solution as the best individual (in our case, the one with minimum energy) in all generations. The complete algorithm became as in Fig. 4. The algorithm was implemented using Microsoft Visual C# 3.5. Empirical Conformational Energy Program for Peptides (ECEPP) package was used to evaluate energy of proteins.

A comparative study is made with two other algorithms (clonal selection algorithm, Genetic algorithm) that solve the same problem Met-Enkephalin protein using "ECEPPAK" tool for evaluate the energy of proteins. The comparison is based on the best fitness (minimum protein energy) that each algorithm has reached with respect to the overhead (number of energy evaluations) needed to reach this result.

In the hyper clonal selection algorithm, the selected individuals are randomly recombined and their offspring are crossover and mutated, so as to generate a new population of N-1 elements.

The best individual of the old population is then added to the new population, and the cycle of life continues. By doing so, the best individuals are treated as super individuals and mated together, hoping that this can lead to a fitter population. In the three algorithms we tested the energy of the immunes by using ECEPPAK tool

5.1- Testing Results:

5.1.1 Genetic algorithm

By using the genetic algorithm we could reach the energy of -20.525 after 1099 generations as the below screenshot.

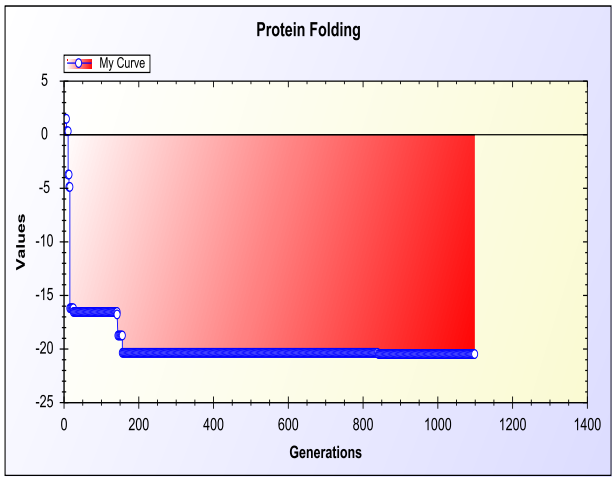

Figure 3: the protein folding problem using Genetic algorithm

5.1.2 Clonal selection algorithm

By using the clonal selection algorithm we could reach the energy of -21.043 after 30 generations as the below screenshot.

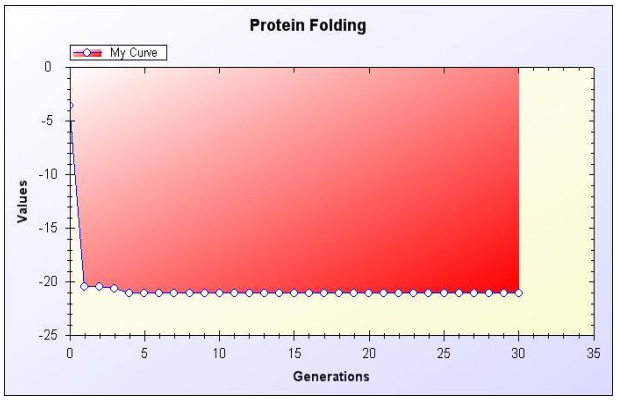

Figure 4: the protein folding problem using Clonal selection algorithm

5.1.3 Hyper clonal selection algorithm

By using our hyper clonal selection algorithm we could reach the energy of -20.919 after 10 generations as the below screenshot.

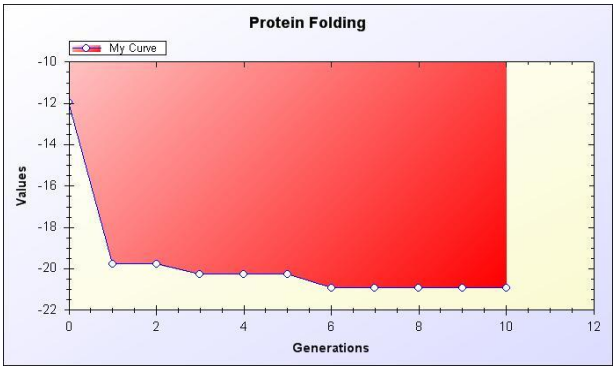

Figure 5: the protein folding problem using hyper clonal selection algorithm

6. Conclusions

In this research, artificial immune system is combined with genetic algorithms in one hybrid algorithm. A modification is proposed to the clonal selection algorithm which is inspired from the clonal selection principle and affinity maturation of the human immune responses. The adaptability of the mutation rate is introduced by simple degrading function. Also, the crossover is merged into the clonal selection algorithm, two-point crossover applied after the mutation process, to increase the exploration of the landscape. We claim that our evolved system exhibits two important characteristics; first, it attains high classification performance, with the possibility of attributing a confidence measure to the output diagnosis. The hybrid algorithm overcomes both the genetic algorithm and the artificial immune system and reached the minimum energy.

REFERENCES

[1] [1]. Hue, S.C. and K.A. Dill, 1993.The protein folding problem. Phys. Today, 46: 24-32. DOI: 10.1063/1.881371.
[2] D.A. Coley, An introduction to genetic algorithms for scientists and engineers, world Scientific Publishing Co.,inc., 2001.
[3] T. Back, The Interaction of Mutation Rate, Selection & Self-Adaptation within a Genetic Algorithm, In Proc. 2nd Int. Conf. on Parallel Problem Solving From Nature, North-Holland, Amsterdam, pp. 85-94, 1992.
[4] Grefenstette, J.J. Optimization of Control Parameters for Genetic Algorithms[J].. Systems, Man and Cybernetics, IEEE Transactions on,Jan. 1996,16(1): 122-128
[5] D. Dasgupta , Artificial Immune systems and their applications, Springer-Verlag, inc., 1999.
[6] D. Dasgupta, N. Attoh-Okine, Immunity-Based Systems, IEEE International Conference on Systems, Man, and Cybernetics, Orlando, Florida, pp 363-374, October 12-15,1997.
[7] L.N. De Castro, F.J. Zuben, Learning and optimization using the clonal selection principle ,IEEE transactions on evolutionary computation , vol.:6, num.:3, pp 239-251, Jun, 2002.
[8] L.N. De Castro, F.J. Zuben, The Clonal Selection Algorithm with Engineering Applications, Artificial Immune System Workshop, Genetic and Evolutionary Computation Conference , A. S. Wu(Ed.), pp. 36-37, 2000.
[9] L.N. De Castro, J. Timmis, Artificial Immune Systems (A new computational Approach) , Springer- Verlag, 2002.

Adel Omar Mohamed received the B.S. degrees in Computer Science and information system from Helwan university in 2004.

Abd El Fatah Hegazy Prof. Dr. ,Dean Assistant for Post Graduate Studies, Collage of Computing and Information Technology, Arab Academy for Science, Technology & Maritime
Transport, Cairo, Egypt.

Amr Badr Prof. Dr. of Artificial Intelligence, Faculty of Computers & Information ,Computer Science Department, Cairo University, Cairo, Egypt