Modified Crow Search Algorithm for Protein Structure Prediction

Hesham Awadh Bahamish1, Naziha Mohammed Al-Aidroos2, Aziz Nasser Boraik3

1Assistant professor, Computer Science Department, College of Computers and Information Technology, Hadhramout University, Yemen
2Associate professor, Computer Science Department, College of Computers and Information Technology, Hadhramout University, Yemen
3Assistant professor, Information Technology Department, Faculty of Computing and Information Technology, University of Science and Technology, Hadhramout Branch, Yemen

Abstract - This paper investigated the capability of the Crow Search Algorithm to solve the protein structure prediction problem (CSA-PSP) and proposed a modification to the algorithm (MCSA-PSP) to enhance its performance. Crow search algorithm is a nature-based optimization algorithm. It is inspired by the manner of the crow flocks in nature. Three modifications to the original algorithm were proposed which are: 1) Using random genetic crossover for generating new solutions, 2) using the Metropolis criteria in accepting the new solutions and 3) a new method for generating new solutions. The performance of the original and modified algorithm was tested on a well-known benchmark protein (Met-enkephalin). The results showed that the algorithm was able to find the reported lowest free energy configuration and that the introduced modifications enhanced the performance of the original algorithm.

Key Words: Crow Search Algorithm, Protein Structure Prediction, Optimization, Conformation Search

1. INTRODUCTION

Living organisms are made of small molecules called proteins. These proteins are built of different combinations of amino acids. The amino acids sequence or chain is known as the primary structure that is folded due to amino acid interactions into the secondary structure and then into the tertiary structure which is the active state or biological active state. Quaternary protein structure is formed of more than one protein chain.

Proteins perform the biological functions in the human body. The importance of the protein structure comes from the fact that the protein cannot perform its biological function unless it folds into the native state (i.e. tertiary structure) [1]. Therefore, in order to know the protein function, it is necessary to reveal its tertiary structure first.

The protein structure prediction (PSP) problem is one of the most complex problems in Structural Biology, Computational Biology, and Bioinformatics [2-5]. Protein structure can be determined experimentally and computationally. The current computational prediction methods are classified into three classes. The Homology Modelling, Threading, and Ab initio

In the Ab initio method, the structure is predicted based on amino acids sequence. This is based on the thermodynamic hypothesis stated by Anfinsen [1]. In this hypothesis, the protein structure or native state is the structure, which has the lowest free energy. So to predict the protein structure computationally, the problem is formulated into an optimization problem where the goal is to search the problem search space to locate that lowest free energy structure (conformation).

Exploring the protein search space is an NP hard problem due to the gigantic number of possible solutions. A huge number of optimization algorithms were applied to solve this problem such as Genetic Algorithms (GA) [6, 7], Simulated Annealing (SA) [8], Bat Algorithm (BA) [9], Marriage in Honey Bee algorithm (MBO) [10], Harmony Search (HS) [11]. Bees Algorithm [12] Artificial Bee Colony (ABC) [13, 14] to name some. Like other algorithms, which are inspired by nature, the Crow Search Algorithm (CSA) is a population-based optimization algorithm, which is inspired by the foraging behavior of crows. It is considered as is one of the most recently developed metaheuristic algorithms [15-17]. Askarzadeh proposed it in 2016 and applied it to solve continuous optimization problems, and since that time it was extensively applied to solve many optimization and real-world problems [18]. It showed an enhancement over other state-of-art algorithms in solving constrained engineering design problems [19].

Previous works reported the CSA advantages such as simplicity and ease of programming [16, 19], using only two variable parameters (flight length and awareness probability) [20, 21], the fast convergence rate [20], and good and acceptable results for solving optimization and real-world problems. Despite these advantages, other researchers reported some disadvantages of the algorithm. These disadvantages are the premature convergence [16, 19, 22], trapping in local minima [18], the poorly exploration capability of the algorithm for certain types of hard problems [16, 18] and in some circumstances it may fail to find the optimal solution [19]. Therefore, many modifications to the original algorithm had been proposed to enhance its performance.

Khalilpourazari and Pasandideh proposed a hybrid crow sine-cosine algorithm SCCSA [23] to enhance the exploration and exploitation of the CSA algorithm. The hybrid algorithm
was evaluated in unimodal, multimodal, fixed-dimensional multimodal, and composite benchmark functions.

Hassanien proposed a hybrid crow search algorithm (RCSA) by combining CSA and rough searching scheme (RSS) and dynamic flight length to improve the algorithm exploration process. By this hybridization, the convergence of the algorithm was improved and the algorithm was able to get out of the local minima trap. The algorithm was applied to solve engineering optimization problems. The advantage of the hybridized algorithm was the exploitation of the favorable areas in the search space that may contain the global optimal solutions [24].

CSA was hybrid with space transform search (STS) algorithm [18] to enhance the CSA diversification capability and was applied to solve IEEE CEC 2017 test functions [18].

Han et al. proposed ISCSA algorithm, in which the CSA was improved by adding three modifications. These modifications were: introducing the spiral search mechanism, a weight coefficient and Gaussian variation, and random perturbation approach [16].

To speed up the algorithm and enhance the value of the solutions, Allaoui et al. combined the CSA with a local search method to solve the DNA fragment assembly problem [25].

Díaz et al. modified the awareness probability and the random perturbation. The modified algorithm was applied to solve complex optimization energy problems [15].

Jain et al. added an experience factor, adaptive modification operator, and Levy flight to balance the CSA exploration and exploitation abilities and prevent premature convergence [19].

Fallah et al. proposed ICSA [26]. In this algorithm, the crow with the worst function value was replaced by a newly generated solution.

In this paper, the CSA algorithm was applied to solve the PSP and three modifications were proposed to enhance the performance of the CSA algorithm. These modifications are using random genetic crossover for generating new solutions, using the Metropolis criteria in accepting the new solutions, and a new method for generating new solutions. The remainder of the paper is organized as follows: Section 2 presents the crow in nature. Section 3 describes the standard crow search algorithm. Section 4 describes the crow search algorithm for protein structure prediction and the proposed modifications to CSA. Section 5 contains the experiments and algorithm implementation. Section 6 presents and discusses the results. Finally, the paper is concluded in Section 7.

2. Crow in Nature

The crow is an intelligent creature. It is considered one of the most intelligent birds [25-27]. It has a large brain size compared to its body size [22]. It is the same relative size as the chimpanzee brain [28]. Besides that, the crow has a brilliant memory [17] that enables it to:

- Remember its hidden food places, which can be from last season [29].
- Recognize the faces of humans.
- Modifies its migration route to get around dangerous areas where the crow had been attacked by humans or other animals.
- Retrieve other information that it uses in the future.

The intelligence of the crow is considered as a seven-year-old human child intelligence [30]. The intelligent behavior of the crow is clearly noticed in the daily activities that the crow performs in the nature environment. Scientists studied the crow behavior in its environment and revealed remarkable intelligent characters [27].

Crow plans to get its food [31]. It adapts to the changes in its environment. Hence it uses tricks, exploits other birds, and makes some kind of tools that are used later to ease the process of eating, getting food, or solving problems [27]. Crow communicates with other crows throw some complicated ways [26, 27, 32, 33] which could be considered some kind of language [29].

Crow is a greedy bird so it tries to find food by observing other crows; it does that to know their hidden food storages to acquire better food [24, 32]. This task is not as simple as it seems since the crow is always alert of other crows watching it. In consequence, it changes its food storages or pretends that it is keeping its food in other fake places [18, 32]. It uses previous knowledge to predict the behavior of other crows [32, 34, 35].

The food foraging behavior of the crow can be described as following [32]:

1. Crow is a social bird, which lives in groups called a flock.
2. Crow has the ability to remember where it hides its food.
3. Crow tries to reveal the other crows hiding places to steal their food.
4. In order to protect its food place, crow uses a probability.

3. Crow Search Algorithm

In this section, the CSA is described in detail. The algorithm is based on the natural intelligent behavior of the crow in getting and keeping its food away from pilferers. In this algorithm, N is the number of crows and the food sources are the possible solutions to the problem, which form the problem search space. Each solution is consists of PV problem variables or problem dimensions. AP is the awareness probability, FL is the flight length, and MAX ITER is the maximum number of iterations. In order to store the good solution in each iteration of the algorithm, each crow has its own memory $C_{i}$ and $MEM_{t}$ where $t$ is the current iteration and $i$ is the index of the crow.
The steps of the algorithm are outlined as the following phases:

**Phase one (initial phase):**
1. For 1 to N crow, an initial food source (solution) is generated (randomly or based on problem domain).
2. Each generated food source is evaluated using problem-specific evaluation function.
3. Each crow stores (memories) its initial food source in its memory ($C_{t, MEM}$).
4. Set the values of $AP$ and $FL$ parameters.

**Phase two (optimization process):**
In this, phase for $MAX_{ITER}$ iterations, each crow $i$ has to decide based on a randomly generated value $ar$ to be in one state of the following states:

State 1: Follow a random crow $j$ to steal its food source (solution) if $ar$ is greater than or equal $AP$ eq. 1.

$$C_{t+1,i} = C_{t,i} + ar_i * FL_i * (C_{t, MEM_i} - C_{t,i}), \text{ if } ar_i \geq AP \quad (1)$$

State 2: Generate a new food source randomly eq. 2.

$$C_{t+1,i} = \text{ random food source} \quad (2)$$

After that, the generated food sources are checked for visibility. Only the visible food sources are evaluated using the evaluation function and the crow's memory is updated as the following:

$$C_{t+1,i} \text{ if } C_{t+1,i} \text{ is better than } C_{t, MEM_i}$$

$$C_{t+1,i} = C_{t, MEM_i} \text{ otherwise} \quad (3)$$

This continues until the $MAX_{ITER}$ is reached and the food source with the best value is selected as the optimal solution of the problem. The following pseudo code describe the standard CSA.

---

**Phase one (initial phase):**
- Initialize the food sources (solutions).
- Evaluate the generated food sources.
- Memorize the initial food source in its memory ($C_{t, MEM}$).
- Set the values of $AP$ and $FL$ parameters.

**Phase two (optimization process):**
While $t < MAX_{ITER}$
for $i = 1 : N$ (all $N$ crows of the flock)
  Randomly choose one of the crows to follow e.g. $j$
  if $ar_i \geq AP$
    $C_{t+1,i} = C_{t,i} + ar_i \cdot FL_i \cdot (C_{t, MEM_i} - C_{t,i})$
  else
    $C_{t+1,i} = \text{ a random food source}$
end if
end for
- Check the feasibility of new solutions
- Evaluate the new position of the crows
- Update the memory of crows using (3)
End while

---

4. CROW SEARCH ALGORITHM FOR PROTEIN STRUCTURE PREDICTION

In this section, the CSA is applied to predict the tertiary structure of the protein, this algorithm is called (CSA-PSP), and then we introduced three modifications to the algorithm to enhance its performance. The modified algorithm is called (MCSA-PSP).

In the CSA-PSP, each crow's food source is considered as a protein conformation (solution to the problem). This conformation consists of a set of torsion angles $\mathbb{N}$ which is the dimension of the search space. Phase one begins by setting the values of the parameters: $FL$, $AP$, and generating random protein conformations by setting random values in the range $\{\pm\}$ to each torsion angles. After that, these conformations are evaluated by a fitness function (energy function) to determine the energy of the generated conformations.

These initialized conformations are saved in the memory of each crow. After initializing the conformations and crows memories, phase two starts. For $MAX_{ITER}$ iterations, each crow $i$ has to decide based on a randomly generated value $ar$ to be in one of the following states:

State 1: Follow a random crow $j$ to steal its food source (solution) if $ar$ is greater than or equal $AP$ eq. 4.

$$Conf_{t+1,i} = Conf_{t,i} + ar_i \cdot FL_i \cdot (Conf_{t, MEM_i} - Conf_{t,i}), \text{ if } ar_i \geq AP \quad (4)$$

State 2: Generate a new food source randomly eq. 5.

$$Conf_{t+1,i} = \text{ random food source} \quad (5)$$

After that, the generated conformations checked for visibility. Only the visible conformations are evaluated using the evaluation function and the crow's memory is updated as the following:

$$Conf_{t+1,i} \text{ if } \text{ (Conf}_{t+1,i} \text{ is better than } Conf_{t, MEM_i})$$

$$Conf_{t+1,i} = Conf_{t, MEM_i} \text{ otherwise} \quad (6)$$

This continues until the $MAX_{ITER}$ is reached and the conformation with the lowest free energy value is selected as the optimal solution of the problem.

4.1 Modified CSA-PSP

Three modifications to the CSA are introduced. The first is to generate the new conformations based on a random genetic cross over instead of using eq. 1, the second is accepting or rejecting the newly generated conformations based on Metropolis criteria, and the third is a new method for generating new random solutions. In the following subsections, these modifications are described.

4.1.1 Modification No. 1: Genetic Crossover:

In this modification instead of generating the new food source using eq. 1 for all problem variables, we modified it by performing a genetic crossover between the selected solution from the crow memory and the current solution of the crow.
The values of the variables in group one are taken from the chosen solution from the crow memory \((C_{MEM})\) and the values of the variables in group two are taken from the current solution \((C_o)\).

<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
<th>V8</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
</tr>
<tr>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
</tr>
<tr>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
</tr>
<tr>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
<td>G2</td>
<td>G1</td>
</tr>
</tbody>
</table>

Fig. 1: Generating Solution by Crossover

### 4.1.2 Modification No. 2: Metropolis Criteria:

In CSA and all its variant modifications, the new food source is saved in the crow memory if it is better than the previously saved value. This is a greedy selection criterion. Using the Metropolis criterion [36] unable the algorithm to escape from the local minima traps by accepting low-quality solutions, which allows the overcoming of barriers and finding better solutions.

In this modification, a form of Metropolis criterion is used to accept or reject the replacement of the crow's memory by the new food source when the food source is worse than the crow's memory value. If the value of the food source is better it is accepted and replaced the old value in the crow's memory. Otherwise, it accepted if:

\[
R = exp(-(Ct_i - Ct_{MEM}))
\]

Where \(R\) is a generated random number in \([0,1]\).

### 4.1.2 Modification No. 3: Solution Generation:

In this modification, the new solution is generated as following:

\[
\alpha = r1 * 0.5 \\
\text{if } r2 < 0.5 \\
\quad \text{new} = \max_{\text{dim}} + \alpha; \\
\text{else} \\
\quad \text{new} = \max_{\text{dim}} - \alpha;
\]

While \(r1\) and \(r2\) are randomly generated numbers and \(\max_{\text{dim}}\) is the maximum value of the problem dimension.

## 5. Experiments and Algorithm Implementation

Table I shows the specification of the machine, programming languages, and tools which were used in the implementation of the algorithm.

### Table I Specifications of Algorithm Implementation

<table>
<thead>
<tr>
<th>Processor</th>
<th>i7-3740QM CPU @ Intel® Core (TM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>2.70 GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>16 GB</td>
</tr>
<tr>
<td>Programming language</td>
<td>Visual C++</td>
</tr>
<tr>
<td>Operating System</td>
<td>Windows 10</td>
</tr>
</tbody>
</table>

The algorithm was tested using Met-enkephalin, a benchmark protein. Its amino acids sequence is \((\text{Tyr-Gly-Gly-Phe-Met})\) which contains 24 main and side-chain torsion angles. For the evaluation of the generated solutions, the ECEPP/3 [37, 38] function was used.

In all experiments, the values of parameters were set as following: \(AP=0.2, FL=2, N=5, \text{and the number of iterations } = 5000\). Table II shows the settings of the eight experiments. In experiment 1 the standard CSA was used. In experiment 2, modification No. 1 was used. In Experiments 3, modification No. 2 was used. In experiment 4, modification No. 3 was used. In experiment 5, modification No 1 and 2 were used. In experiment 6, modification No. 1 and 3 were used. In experiment 7, modification No 2 and 3 were used. In experiment 8, the three modifications were used.

### Table II Experiments Settings

<table>
<thead>
<tr>
<th>Experiment No.</th>
<th>Modification No. 1</th>
<th>Modification No. 2</th>
<th>Modification No. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>8</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

## 6. Results and Discussions

Table III shows the results of 30 independent runs of the eight experiments. The best, average, worst energies, standard deviation, and success rate are listed. In all experiments, the algorithm was able to find the reported lowest free energy solution (-12.43 kcal/mol).

### Table III: Results of the Eight Experiments

<table>
<thead>
<tr>
<th>No.</th>
<th>Best</th>
<th>worst</th>
<th>Average</th>
<th>STD</th>
<th>Success rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-12.43</td>
<td>-6.89</td>
<td>-11.54</td>
<td>0.78</td>
<td>27%</td>
</tr>
<tr>
<td>2</td>
<td>-12.43</td>
<td>-9.34</td>
<td>-11.59</td>
<td>0.73</td>
<td>30%</td>
</tr>
<tr>
<td>3</td>
<td>-12.43</td>
<td>-9.81</td>
<td>-11.66</td>
<td>0.71</td>
<td>33%</td>
</tr>
<tr>
<td>4</td>
<td>-12.43</td>
<td>-10.15</td>
<td>-11.50</td>
<td>0.71</td>
<td>27%</td>
</tr>
<tr>
<td>5</td>
<td>-12.43</td>
<td>-9.78</td>
<td>-11.51</td>
<td>0.75</td>
<td>30%</td>
</tr>
<tr>
<td>6</td>
<td>-12.43</td>
<td>-10.51</td>
<td>-12.03</td>
<td>0.61</td>
<td>66%</td>
</tr>
<tr>
<td>7</td>
<td>-12.43</td>
<td>-10.60</td>
<td>-12.05</td>
<td>0.50</td>
<td>60%</td>
</tr>
<tr>
<td>8</td>
<td>-12.43</td>
<td>-11.27</td>
<td>-12.27</td>
<td>0.37</td>
<td>83%</td>
</tr>
</tbody>
</table>
The performance of the standard CSA was slightly improved in terms of success rate from 27% to 30% when modification no 1 was used and to 33% when modification no 2 was used, while there was no improvement when modification no 3 was used.

The combination of modification no 1 and 2 improve the success rate of the standard CSA to 30% while a significant improvement was gained when the combination of modification no 1 and 3 (%66), and the combination of modification no 2 and 3 (%60) were used.

The effect of the combination of the three modifications is clearly shown in the results of experiment 8. Which gained the best results in terms of success rate, average energy, and standard deviation.

The convergence of the algorithm in the eight experiments is shown in chart 1. It has clearly shown the effects of the modifications on the performance of the algorithm.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Best energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSA</td>
<td>-12.43</td>
</tr>
<tr>
<td>BA [9]</td>
<td>-12.43</td>
</tr>
<tr>
<td>MOLS [39]</td>
<td>-10.10</td>
</tr>
<tr>
<td>Basin paving [40]</td>
<td>-12.43</td>
</tr>
<tr>
<td>TS [41]</td>
<td>-12.39</td>
</tr>
<tr>
<td>Conformation Space Annealing [8]</td>
<td>-11.71</td>
</tr>
<tr>
<td>MBO [10]</td>
<td>-12.43</td>
</tr>
</tbody>
</table>

7. CONCLUSIONS

In this paper, the capability of the CSA algorithm to solve the PSP problem was investigated. The algorithm was used as an optimization algorithm to solve the PSP problem. Three modifications to the original CSA algorithm were proposed. The original and the modified algorithms were able to find the reported lowest free energy conformation of the test protein. The proposed modifications enhanced the performance of the algorithm and the significant improvement was when the three modifications were used. Future work will be a study of the algorithm performance on longer protein size.

REFERENCES


