Research Article

Construction of Computer Algorithms in Bioinformatics of the Fusion Genetic Algorithm

Haimeng Sun

Jimei University Chengyi College, Xiamen, Fujian 361021, China

Correspondence should be addressed to Haimeng Sun; haimes@jmu.edu.cn

Received 8 June 2022; Revised 27 July 2022; Accepted 27 August 2022; Published 22 September 2022

Academic Editor: Xiantao Jiang

Copyright © 2022 Haimeng Sun. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the continuous in-depth exploration of life sciences, bioinformatics based on life sciences, computer algorithms, and statistics have gradually developed. The research of bioinformatics mainly focuses on the study of genes, and the structural characteristics of genes lead to a large amount of extremely complex data in the study of bioinformatics. Analyzing data in bioinformatics research requires accurate calculation by computer algorithms. However, common computer algorithms such as the dynamic programming algorithm and the genetic algorithm have the disadvantages of large memory or inaccurate optimization. Combining the ant colony algorithm (ACA) and GA can give the advantages of the two methods that should be fully utilized to efficiently analyze the biological information data. In this paper, the ant colony fusion genetic algorithm (ACA-GA), GA, and dynamic programming algorithm are used to compare and analyze the sensitivity, convergence speed, sequence alignment accuracy, and required memory space of gene sequences. The experimental results show that compared with ACA-GA, the dynamic programming algorithm has the advantage of finding the optimal alignment of 100%, but the memory required is too large, the memory required is more than 10 times that of the ant colony fusion genetic algorithm and the sensitivity is not as good as ACA-GA. The convergence speed of ACA-GA is faster than the gene comparison speed of GA and the accuracy is 2.6% better than that of GA on average. ACA-GA has the advantages of GA and ACA, which can improve the computational efficiency of biological data in bioinformatics.

1. Introduction

Bioinformatics, as an interdisciplinary field, is crucial to the advancement of biological sciences. Computer algorithms and biological science foundations are two fundamental technologies in bioinformatics. The job of computer algorithms is to collect, process, and organize data from biological research into useful biological information for researchers to evaluate and use. The properties of biological experimental data include massive data and complex data relationships. Therefore, sufficient computer algorithms are needed to accurately evaluate biological data. Common computer algorithms include the dynamic programming algorithm and GA, but both algorithms have different shortcomings. ACA is the shortest optimization method abstracted with reference to ant colony optimization and has the characteristics of positive feedback optimization. Fusion genetic algorithm is the fusion of the genetic algorithm and other algorithms, combining ACA and GA, ACA-GA has the characteristics of global convergence and shortest optimization, which is very suitable for dealing with complex biological data problems in bioinformatics. Therefore, this paper has research significance.

The research of biological science needs to deal with complex and changeable biological data, and relevant researchers use scientific computing methods to apply computer algorithms to bioinformatics. Among them, Bei used the fireworks algorithm to solve the problem of insufficient general optimization in biological science [1]. Geiger introduced a computer method to solve complex data and difficult resource scheduling, which can improve the calculation speed of biological data. And it can achieve remarkable results in biological data research [2]. Salihu’s research found that the use of the matrix solution method...
can improve the accuracy of biological scientific computing [3]. Courneya’s research pointed out that computer science algorithms can effectively improve the efficiency of data computing in bioinformatics [4]. Namasudra cited a computer algorithm to improve the accuracy of gene sequence alignment [5]. Although the efficiency of processing biological information data can be improved through computer algorithms and lack of research on fusion algorithms.

GA has the characteristics of global convergence optimization, and fusion GA that integrates other optimization methods with GA can play the advantages of different algorithms. Among them, Qi proposed a nuclear fusion method, which can effectively detect the sequence and structure of genes [6]. Lakshmi pointed out that fusion GA has the characteristics of GA and has excellent global optimization ability, which can effectively compare and analyze biological gene sequences [7]. Daniel integrated other computer algorithms and GA to improve the sensitivity of gene sequence detection [8]. Sulistyo can improve the analysis accuracy of bioinformatics data by using the fusion of the neural network and GA [9]. Arif proposed a curvelet transform fusion GA, which can reduce the required memory space while ensuring the accuracy of data analysis [10]. The fusion of GA and the advantages of different algorithms can improve the processing efficiency of GA, but the fusion method is not optimal.

ACA and GA are integrated, so that the integrated algorithm has the characteristics of GA and ACA at the same time. The algorithm advantage is shown as global convergence and shortest optimization [11]. Compare the most common gene sequence problems in bioinformatics by fusing GA with GA and dynamic programming algorithms. The innovation: the shortest optimization feature of ACA is combined with GA, and the comparison of biological data with the dynamic programming algorithm and the genetic algorithm is analyzed.

2. Method of Fusion GA

Bioinformatics is an interdisciplinary field that successfully integrates biology, mathematics, and computing. It mainly uses scientific methods such as mathematics and information science to analyze, process, and solve complex biological data in order to decipher the content of large amounts of bioinformatic data [12]. In the process of research, bioinformatics is mainly used to effectively compare and analyze biological gene sequences and biological protein structures. The traditional computer algorithm uses GA to process biological experimental data, but the data optimization efficiency is low. In bioinformatics, complex bioinformatic data can be processed jointly by fusing the ant colony method and GA. Figure 1 depicts the bioinformatics model.

It can be seen from Figure 1 that the processing of data in bioinformatics relies on computer algorithms. In this paper, the biological data is processed by the combination of ACA and GA [13].

2.1. GA. The reproduction, survival of the fittest, birth, old age, sickness, and death of living things in the natural environment are all a kind of laws of natural selection, and in essence, they are determined by genes and are a kind of genetic characteristics of living things. Through the study of biological heredity, GA that follows the laws of biological heredity is abstracted. GA follows the law of biological inheritance, so it has the characteristics of optimization, genetic algorithms are applied to complex data systems, and it can optimize the data structure of the system so that the system can show better characteristics. The discovery of GA provides a simple solution to complex data problems. It has superior ability to deal with complex functions, programs, system structures, and other issues [14].

GA is an abstraction of the evolutionary laws of organisms in nature. It is a natural law. Compared with other optimization algorithms, GA has the following advantages:

- GA is to analyze and process biological individuals or data after genetic action, not to directly operate the parameters of the system. Therefore, GA has a wide range of applications and is applicable to many fields.

- GA has a strong nonlinear processing capability due to the abstract biological genetic characteristics, but there are few linear problems in practical problem processing. Therefore, the problem that GA can handle is more nonlinear.

- GA adopts the method of intergroup search and intergroup optimization and abandons the exhaustive search method, which can effectively search the data efficiently, and GA can realize the global search.

- GA can analyze the data according to the changing laws of the data in the system without external assistance and is not affected by the data definition domain.

- GA can process complex data problems in parallel and improve data processing efficiency [15].

- The basic GA is composed of eight parts, which are the individual setting rule A, the individual adaptation method B, the biological initial state C, the biological population size D, the genetic rule E, the crossover rule F, the mutation rule G, and the termination condition H. The formula is expressed as follows:

\[
SGA = (A, B, C, D, E, F, G, H). \tag{1}
\]

SGA represents GA. \( H \) represents the termination condition.

The principle of GA is shown in Figure 2.

Due to the single evolution principle and the need to encode the problem to be solved and then solve the problem through the decoding step, GA also has many defects. The disadvantages of the GA are as follows:

- GA are prone to irregular and inaccurate problems when coding.
- When faced with complex and multilogical problems, the single genetic rule of GA cannot effectively solve the data problem and wastes too much time.
- GA is computationally slower than other algorithms.
- GA are prone to premature convergence problems.
ACA is a process of simulating the ant colony tasked with determining the quickest path between two places. Each time the ants pass by, they will leave some information. The more ants pass by, the more pheromones they leave behind, and each choice of ants is to move towards a path with more pheromones [16].

TSP problem: in a given $n$ cities, if the traveling salesman starts from one of the cities and then visits the remaining cities in turn, until returning to the original departure city, it is required that each city can only be visited once on the way to find the shortest access path [17].

In mathematical language, it can be expressed as follows: let $A = (a_1, a_2, \ldots, a_n)$ represent the set of $n$ locations, $R = (r_{ij}|a_i, a_j \in A)$ represent the connection between two places, and $D_{ij}(i, j = 1, 2, \ldots, n)$ represent the connection distance in $r_{ij}$, namely:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \quad (2)$$

In formula (2), $x$ and $y$ represent the abscissa and ordinate of the location, respectively.

Therefore, the problem of ACA is to determine the quickest route from the set $G = (A, R)$, that is, the shortest path that has to go through each place once.

Let $n$ be the number of cities, $m$ is the number of ants, $d_{ij}$ is the distance between two cities, and $\mu_{ij}(t)$ is the pheromone between cities $i$ and $j$ at time $t$, then the process of ants moving for optimization is as follows.

Let the initial pheromone be 0, that is,

$$\Delta \mu_{ij}(0) = 0. \quad (3)$$

The probability formula of ant transfer is
2.3. ACA-GA. ACA-GA is based on the global search ability of GA and the information feedback mechanism of ACA. It makes the solution rate and solution accuracy of the fusion algorithm more obvious than the two single algorithms [18].

The detailed steps of the ant colony fusion algorithm are as follows.

The initial pheromone size is set to $\mu_{ij}(0) = \text{const}$, the initial increment of pheromone is $\Delta \mu_{ij}(0) = 0$, the ant colony’s maximum number of iterations is set to $N_{\text{max}}$, and the maximum number of inheritance iterations is set to $G_{\text{max}}$.

Start the genetic iteration, which is $G = 0$.

Carry out ant colony iteration, set $N = 0$, when $m_k \leq n$, calculate the probability of ant transfer and update the local pheromone.

After the ants complete a city-wide search, move the ants to the starting point, record the taboo areas of the ants, and update the size and location of the pheromone left by the ant colony, so that the ants know where the shortest route is.

$N = N + 1$, when $N < N_{\text{max}}$, continue to start the ant colony iteration.

Record the fitness function of each ant: $S = 1/R_k$.

Loop $G = G + 1$, when $G < G_{\text{max}}$ is followed by genetic iteration, until the optimal path is output, and finally the optimal path map is found by combining the ant’s promotion area and the optimal route of the ant colony.

2.4. Experimental Design of ACA-GA. Experimental design: the experimental environment of this experiment is carried out on the MATLAB platform. BLAST is a matrix structure. The scores in the matrix correspond to the alignment results of two nucleotides. The aligned nucleotides are G, C, T, and A. The same expression is “+6”, and the difference is “−3”.

The advantages and disadvantages of computer algorithms in biological information data processing can be judged by comparing and analyzing four aspects of gene sequence sensitivity, convergence speed, sequence alignment accuracy, and required memory space [19]. The BLAST scoring matrix is shown in Table 1.

The function to judge the gene sequence is as follows:

$$F(H) = \sum_{i=1}^{L} \text{Cost}(a_i, b_i).$$

In formula (6), $L$ represents the length of the gene sequence.

Experimental data: There are three main types of biological sequences in biological data, that is, DNA sequences, RNA sequences, and protein sequences. In this paper, DNA sequence strings are taken as an example to test the DNA sequence alignment performance of ACA-GA, GA, and the dynamic programming algorithm. The conclusion is also applicable to other types of sequence alignments [20].

The experimental data came from the GenBank database, from which 10 DNA sequences were randomly selected, all of which were $\beta$-protein-like gene sequences of primates. Among them, the shortest sequence contains 590 nucleotides, the longest sequence contains 2256 nucleotides. Because it is a double sequence alignment, the number of sequences in each group is 2, which are all sequence comparisons between humans and other primates, that is, they are divided into nine groups [21]. The sequences of primates selected for the experiment are shown in Table 2.

In Table 2, the data names of the primate DNA sequence set are named after the names of other primates that are compared with the human $\beta$-protein-like genes. The last line of human $\beta$-protein gene information is particularly useful for the study of primate DNA sequence [22].

3. Result Analysis

3.1. Sensitivity to Gene Sequence Alignment. The sensitivity of gene sequence alignment represents the sensitivity of computer algorithms to detect gene sequences. The sensitivity comparison of the 10 gene sequences in Table 2 was carried out through ACA-GA and the dynamic programming algorithm, and the average of the 10 comparison results was taken in this comparison [23]. The sensitivity
results of the dynamic programming algorithm (A algorithm) and ACA-GA (B algorithm) to the gene sequence alignment are shown in Table 3.

In Table 3, the sensitivity of three groups of data in ACA-GA is lower than that of the dynamic programming algorithm, namely: the first group, the sixth group, and the ninth group. ACA-GA and the dynamic programming algorithm have two sets of data that are the same, namely: the third group and the eighth group. In ACA-GA, the sensitivity of four groups of data is higher than that of the dynamic programming algorithm, namely: the second group, the fourth group, the fifth group, and the seventh group. Therefore, ACA-GA is more sensitive than the dynamic programming algorithm in the overall sensitivity of the above 10 gene sequences [24].

### Table 3: Sensitivity results for gene sequence alignment.

<table>
<thead>
<tr>
<th>No.</th>
<th>Sequence of primates</th>
<th>Number of sequences</th>
<th>Length</th>
<th>Algorithm A</th>
<th>Algorithm B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lemur</td>
<td>2</td>
<td>590</td>
<td>334</td>
<td>267</td>
</tr>
<tr>
<td>2</td>
<td>Gorilla</td>
<td>2</td>
<td>1450</td>
<td>389</td>
<td>467</td>
</tr>
<tr>
<td>3</td>
<td>Loris</td>
<td>2</td>
<td>1100</td>
<td>567</td>
<td>567</td>
</tr>
<tr>
<td>4</td>
<td>Tarsier</td>
<td>2</td>
<td>1290</td>
<td>520</td>
<td>540</td>
</tr>
<tr>
<td>5</td>
<td>Prosimii</td>
<td>2</td>
<td>1100</td>
<td>502</td>
<td>520</td>
</tr>
<tr>
<td>6</td>
<td>Lemuridae</td>
<td>2</td>
<td>1678</td>
<td>670</td>
<td>640</td>
</tr>
<tr>
<td>7</td>
<td>Macaque</td>
<td>2</td>
<td>1087</td>
<td>489</td>
<td>521</td>
</tr>
<tr>
<td>8</td>
<td>Lagothrix</td>
<td>2</td>
<td>1890</td>
<td>698</td>
<td>698</td>
</tr>
<tr>
<td>9</td>
<td>Baboon</td>
<td>2</td>
<td>2256</td>
<td>702</td>
<td>667</td>
</tr>
<tr>
<td>10</td>
<td>Hominidae</td>
<td>—</td>
<td>1080</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

3.3. **Accuracy of Alignment of Gene Sequences.** The alignment accuracy of gene sequences is obtained by observing the best and worst results of gene sequence alignment, taking the average of the two results and comparing them with the known optimal solution. ACA-GA was used to update the characteristics of the pheromone and the traditional GA was used to compare the 10 groups of gene sequences in Table 2 [27]. Figure 4 shows the results of the gene sequence alignment accuracy of ACA-GA and the traditional GA.

In Figure 4, the accuracy of ACA-GA in the alignment of these 9 groups of gene sequences is better than GA. The
average accuracy of ACA-GA is 97.2%, while the average accuracy of the traditional algorithm is 94.6%. Therefore, ACA-GA is more accurate than the traditional GA in the accuracy of gene sequence alignment.

3.4. Memory Space Required to Align Gene Sequences. The memory space required for computing is also the main feature to measure the pros and cons of computer algorithms. The dynamic programming algorithm and ACA-GA are used to compare the accuracy of the first four groups of gene sequences, and the memory space occupied by the two algorithms during runtime is tested. The results of the gene sequence ratio accuracy and the memory space used by the dynamic programming algorithm and ACA-GA are shown in Figure 5.

From the data analysis shown in Figure 5, the dynamic programming algorithm can always maintain an optimal search plan, the accuracy of optimization can reach 100%, while the average accuracy of ACA-GA is 99.2%, and the gene comparison accuracy of the two algorithms is not much different. However, in Figure 5(b), the memory required by the dynamic programming algorithm for gene alignment is far greater than that of ACA-GA [28]. Therefore, ACA-GA can still have a very high gene alignment accuracy while maintaining a low memory space.

4. Conclusion

This paper compares and analyzes the sensitivity, convergence speed, accuracy, and required memory space of ACA-GA, traditional GA, and dynamic programming algorithm for gene sequence alignment. The experimental results show that ACA-GA is significantly faster than GA in the convergence speed of gene sequence alignment. And in the accuracy of gene alignment, the average accuracy of the two algorithms are 97.2% and 94.6%, respectively. In the comparison and analysis of gene sequences between ACA-GA and the dynamic programming algorithm, ACA-GA is more sensitive to the gene sequence alignment than the dynamic...
programming algorithm, and the accuracy is not much different. But the memory space occupied by the dynamic programming algorithm is far more than the memory space occupied by ACA-GA. By combining the advantages of ACA and GA, ACA-GA greatly improves the alignment accuracy of gene sequences and improves the efficiency of biological information calculation. Genetic algorithm has high scalability and realizes efficient processing of biological information data through the integration with other algorithms. Therefore, expanding more fusion GA will be the direction of future research.

Data Availability

The data used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References


