Universe
Scientific Publishing

# Research and implementation of gene sequence alignment based on Spark cloud computing and chaotic genetics

**Qingxue Liu**

**Jilin University of Architecture and Technology, Changchun 130012,China.**

***Abstract:*** aiming at the low speed and accuracy of the existing alignment methods, chaos genetic algorithm is used to search the optimal solution quickly, Spark cloud computing is used to parallelize the alignment, which greatly reduces the execution time and improves the alignment accuracy, it provides an effective tool for deciphering biological genetic code.

***Keywords:*** Spark Cloud Computing; Chaos Genetics; Gene Sequence Alignment Introduction

Sequence alignment is the basis of the research on biological postsequence, such as evolutionary tree, protein structure prediction, drug design and so on. In the research of sequence alignment, we trace the evolutionary relationship of sequences by finding the similar gene sequences, inferring the similarity and analyzing the evolutionary relationship. Biological sequence alignment has been studied extensively at home and abroad and many methods have been proposed. Through the analogy of natural selection process, genetic algorithm evolves a number of candidate solutions by designing coding methods, genetic and variable operators, designing objective functions. Although genetic algorithm is easy to parallelize and can reduce the time cost, it has some disadvantages such as local optimization and slow convergence. With the growth of sequencing data, the traditional parallel processing methods can not effectively store, analyze and process the data. In Spark cloud computing, the input data is cached in memory, and the data is loaded only once, which greatly saves the time of repeated reading and improves the efficiency of computation.

## 1. Principle analysis of gene sequence alignment and research on genetic algorithm and chaos theory

Genetic alignment is often used to compare the homology, or similarity, of two DNA sequences or biomolecular structure. Firstly, the classical dynamic programming is analyzed, which turns a big problem into a small problem and solves it step by step. Starting with the first character and assuming deletion, for every additional character, there are three possibilities: mismatch, match, deletion/insert. The corresponding score is calculated, and the highest score is the optimal solution.

The essence of double sequence alignment is to insert one or more gaps in any position of two sequences to be compared, so that the two sequences have the greatest similarity, and then infer their biological significance from the results of comparison.

Genetic algorithm (GA) is an evolutionary search algorithm for solving global optimal solution in computational mathematics. By using evolutionary programming, the solution space of the problem is coded to produce a certain number of individuals, which are then evolved by means of heredity, heredity, natural selection and hybridization. But the traditional genetic algorithm also has some shortcomings, such as premature convergence, random roaming and so on.

Chaos, an important branch of nonlinear science, is a description of the potential inherent law in the chaotic and disordered state of deterministic things. The basic idea of chaos can be understood as: firstly, the chaotic variables in the original chaotic space are linearly mapped to the solution space, and then, according to the characteristics of chaos, the target is searched in the solution space by Chaos, chaos optimization method has a strong sensitivity to the initial conditions, the initial conditions of a very small change, if the n times continuously magnified, will also cause a huge impact. It is this third property that makes the chaotic algorithm easily jump out of the local optimal point and find the global optimal solution.

In view of the characteristics of chaos optimization theory and genetic algorithm, a new method combining them is proposed, which can improve the convergence speed and get rid of the limitation of local optimization. Firstly, the initial individuals are generated by the chaotic optimization process, and then the optimal solution obtained by genetic algorithm is perturbed by Chaos to search and find the global

optimal point.

## 2. An improved algorithm of double sequence alignment based on chaotic genetic algorithm

There are some problems in classical genetic algorithm, such as uneven population distribution, slow convergence rate and local optimal solution. In order to solve this problem, a chaotic genetic algorithm based on Logistic map is proposed, in which chaotic sequence generated by chaotic map is used to replace the process of random individual generation to improve the genetic algorithm, the number of iterations is reduced to reach the global optimal solution. The main study of the following five aspects of the content.

① Genetic coding of gene sequences

Example: for the sequence S and sequence T shown in Figure 1, the Needleman-Wunsch algorithm is used to create a matrix. There are only three possibilities for moving from the current position of the Matrix to the next position, right, down along the diagonal, and down, and are represented by the characters R, B, and D respectively, then a result of the alignment of the two sequences can be represented by a line from the upper left corner of the matrix to the lower right corner, which can be represented by a string a, and the characters in the string belong to the set { R, B, D } . Moving down or to the right means inserting a space in a vertical or horizontal sequence, and moving diagonally means that the next character in the two sequences is a match. The result of the S-T alignment can be represented as a string, “BBBBBRBBBB,” which is the encoding of the chromosome.

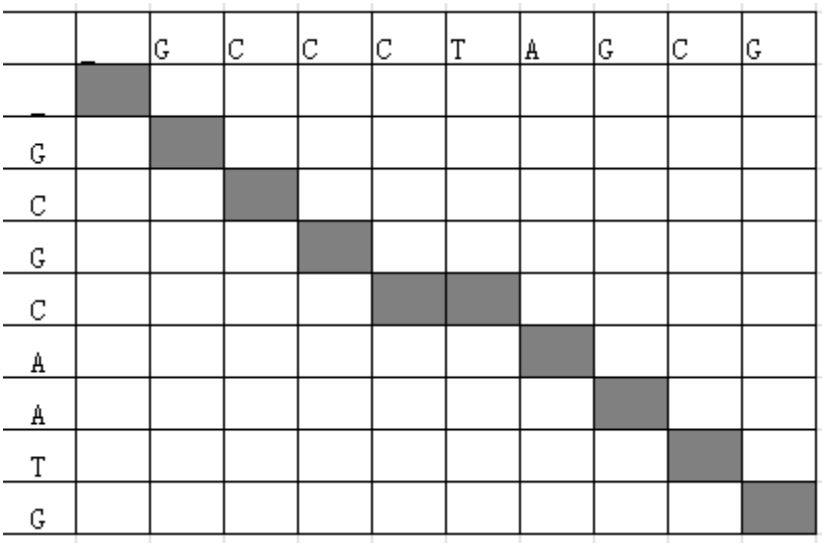


Fig. 1 an example of genetic coding for gene sequences

② Fitness calculation

For each individual, according to its corresponding sequence alignment string and space penalty formula to calculate the corresponding sequence alignment score, fitness value is the basis of whether the individual can iterate, in the process of its calculation to fitness value is not negative, if there is a negative value will need to do some conversion.

③ Penalty points for vacancies

In the process of evolution, organisms inevitably undergo genetic changes. Therefore, in actual sequence alignment, it is necessary to consider changes in base insertion or deletion, which may be one or multiple base deletions or insertions. Vacancy penalty is a method to reflect this variation. The penalty score for a single vacancy is a fixed value Xg. The first vacancy in a continuous vacancy is Xg, and each subsequent vacancy is represented by Yg. If the length of a continuous vacancy is K, the calculation method for the affine vacancy scoring method is Xg+k * Yg.

④ Design of chaotic genetic operators

Based on the main operations of sequence alignment and the evolutionary process of genetic algorithms, selection operations, chaotic crossover operations, and chaotic mutation operations were designed.

Selection operation: Each time two individuals are selected, only the individuals with higher fitness are retained for selection in the next generation population. The PopSize individuals in the generated initial population are evaluated for fitness, and the higher the fitness value, the greater the probability of individuals being inherited to the next generation. This algorithm adopts the Stochastic Tournament Model method, and the league size N takes a value of 2.

Chaotic crossover operation: With an L-position long chromosome, a random number $xn$ is taken as the initial value on the (0,1) interval, and a logistic model is used to iteratively generate a chaotic sequence $xn$+1 on the (0,1) interval. Using the formula C= $(int)xn+1L$ map the sequence $xn$+1 to the chromosome gene position space to determine the location of the crossover operation, and perform a single point crossover with probability Pc. In the process of forming new offspring, only a small number of genes need to be replaced.

Chaotic mutation operation: The method of using chaotic sequences to determine the position of chaos crossover is used to determine the mutation point position, and the mutation individuals are randomly selected from the parent generation with a probability of pm, and mutation operations are performed on a certain or several mutation positions.

⑤ End condition: Stop searching when the predetermined number of iterations is reached or the maximum value does not change.

## 3. Implementation of Gene Sequence Alignment Based on Spark Cloud Computing

Experimental environment

The Spark cluster in this experiment is composed of one master node (NameNode) and three slave nodes (DataNodes). Based on the virtualization big data platform of the college laboratory, the following configuration is adopted: 4 64 bit virtual machines, Centos6.5 operating system; The server is Hadoop-2.5.2 for Apache; The Spark version is Spark-3.0.6, and the JDK version is 1.8; The Scala version is 2.12.6 and adopts the integrated development environment Eclipse.

In order to improve the comparability of this experiment, the same operating system as the Spark cluster was installed on a virtual machine of the same configuration in the laboratory as a standalone experiment. And download alignment sequences of different sizes of datasets from NCBI (National Center for Biotechnology Information) for comparison of two experiments. The results are shown in the table below:

Comparison of experimental results

| name | SRR4253374 | SRR7359601 | SRR7368869 | SRR7369078 | SRR7368879 |
|---|---|---|---|---|---|
| file size(MB) | 9.3 | 156.7 | 271.9 | 468.5 | 1024 |
| Single machine running time(s) | 25 | 481 | 836 | 1681 | 3323 |
| Cluster runtime (s) | 65 | 176 | 354 | 455 | 811 |

## 4. Conclusion

This algorithm utilizes the complementarity between chaotic algorithm and genetic algorithm, which to some extent enhances the sensitivity of sequence alignment, improves alignment quality, and optimizes the local search ability of genetic algorithm. Due to the operating mechanism of Spark cloud computing, this algorithm can greatly improve computational efficiency when processing large-scale data through multi-step iterative calculations. As the file size increases, when processing the same dataset, the cluster method takes significantly more time than the single machine method, and the larger the file, the more obvious this advantage. But when analyzing small data, there are no more advantages.

## References

[1] Wang Fangfang, Ma Zhiqiang, Wang Suhua Sequence alignment method based on genetic algorithm [J] Journal of Jilin University (Information Science Edition), Issue 3, 2013

[2] Liu Zhenyu Parallelization Research and Comparison Platform Construction of Genomic Data Alignment Algorithm Based on Spark [J], Inner Mongolia Agricultural University, 2019

[3] Guo Jing, Wang Chao, Zhang Hongbin, Chen Lin. Overview of methods for constructing phylogenetic trees [J] Computer Application Research, Issue 3, 2013

[4] Lei Liang, Wang Tongqing, Peng Jun, Yang Bo. Research on the Application of Improved Adaptive Genetic Algorithm [J], Computer Science, 2009, vol36 (NO.6): 203

[5] Li Yan, Yuan Hongyu, Yu Jiaqiao, Zhang Gengwei, Liu Keping Overview of the Application of Genetic Algorithms in Optimization Problems [J] Shandong Industrial Technology, 2019, vol. 11 (NO.3): 416

Author Introduction: Liu Qingxue (1977-), female, Han nationality, master's degree, professor, engaged in research on intelligent algorithms and computer science and technology. Email: 406428957 @ qq.com

Supporting project number: Xiaokezi [2022] 009ZQKJ