Interdisciplinary Sciences: Computational Life Sciences (2021) 13:118–127
https://doi.org/10.1007/s12539-021-00419-6

ORIGINAL RESEARCH ARTICLE

An Algorithm for Gene Fragment Reconstruction

Ningyuan Fang[1] · Kaifa Wang[2] · Dali Tong[3]

Received: 5 July 2020 / Revised: 21 January 2021 / Accepted: 1 February 2021 / Published online: 20 February 2021

Abstract

Gene sequencing technology has been playing an important role in many aspects, such as life science, disease medicine and health medicine, particularly in the extremely tough process of fighting against 2019-novel coronavirus. Drawing DNA restriction map is a particularly important technology in genetic biology. The simplified partial digestion method (SPDP), a biological method, has been widely used to cut DNA molecules into DNA fragments and obtain the biological information of each fragment. In this work, we propose an algorithm based on 0–1 planning for the location of restriction sites on a DNA molecule, which is able to solve the problem of DNA fragment reconstruction just based on data of fragments' length. Two specific examples are presented in detail. Furthermore, based on 1000 groups of original DNA sequences randomly generated, we define the coincidence rate and unique coincidence rate between the reconstructed DNA sequence and the original DNA sequence, and then analyze separately the effect of the number of fragments and the maximum length of DNA fragments on the coincidence rate and unique coincidence rate as defined. The effectiveness of the algorithm is proved. Besides, based on the existing optimization solution obtained, we simulate and discuss the influence of the error by computation method. It turns out that the error of position of one restriction site does not affect other restriction sites and errors of most restriction sites may lead to the failure of sequence reconstruction. Matlab 7.1 program is used to solve feasible solutions of the location of restriction sites, derive DNA fragment sequence and carry out the statistical analysis and error analysis. This paper focuses on basic computer algorithm implementation of rearrangement and sequencing rather than biochemical technology. The innovative application of the mathematical idea of 0–1 planning to DNA sequence mapping construction, to a certain extent, greatly simplifies the difficulty and complexity of calculation and accelerates the process of 'jigsaw' of DNA fragments.

Keywords Shotgun · DNA sequence reconstruction · 0–1 planning · matelab7.1

Abbreviations

n	The number of restriction sites on a certain DNA molecule
A	The first set of data, consisting of all the length of fragments if the DNA molecule is cut into two fragments in each restriction site separately. The data size of A is $2n$
B	The second set of data, consisting of all the length of fragments if the DNA molecule is cut into fragments in each restriction site simultaneously. The data size of B is $n + 1$
M	The total length of the DNA molecule, equal to the sum of elements in B
I_i	The i restriction site in the DNA molecule
$x = [x_1, x_2, x_3, \cdots x_n]$	A permutation of the location of DNA restriction sites, where x_i can only be 0 or 1, while 1 means that I_i is closer to one end of the DNA molecule, and 0 means that I_i is closer to the other end
P and Q	The two ends of the DNA molecule, respectively

✉ Kaifa Wang
kfwang72@163.com

✉ Dali Tong
tongdali1985@163.com

[1] School of Sciences, Southwest Petroleum University, Chengdu 610500, Sichuan, People's Republic of China

[2] School of Mathematics and Statistics, Southwest University, Chongqing 400715, People's Republic of China

[3] Department of Urology, Daping Hospital, Army Medical University, Chongqing 400042, People's Republic of China

a_i The shorter distance between I_i and the two ends of the DNA molecule, while aa_i is the longer distance. Obviously, a_i is less than or equal to aa_i and a_i plus aa_i is equal to M

C The set of distance from each I_i to one end of the DNA molecule (for example, P end) in ascending order, and CC is the set of distance from each I_i to the other end (for example, Q end correspondingly) in ascending order. Here we let the distance take 0 if I_i is further away from the end

l_i The length of each fragment on the half of the DNA molecule near one end (for example, P end), and ll_i is the length of each fragment on the half of the DNA molecule near the other end (for example, Q end)

S A collection of the length of all the fragments obtained when the DNA molecule is cut by enzyme at all restriction sites simultaneously in the analysis. S is derived from A and $x_1, x_2, x_3, \cdots x_n$. The data size of S is $2n + 1$. Actually, there are n zeros in S

SS Consists of nonzero elements in the collection S. The data size of SS is $n + 1$

1 Introduction

It's commonly accepted that drawing DNA restriction map is an extremely significant method for genetic and biological analysis. In view of the high molecular weight of DNA and excessive number of base pairs, biochemical technology is used to cut DNA into small molecular fragments in scientific experiments [1, 2]. Specifically, the PDP method and the simplified PDP method (SPDP) based on the different enzyme cutting sites have been proposed in a series of classic studies. The basic biological information of each segment is analyzed to obtain the relevant information of the whole DNA molecule [3, 4]. This sequencing method is the "shotgun method" invented by Craig Venter, founder of selera genetic company in the United States [5].

Quickness, simplicity of implementation, and low cost are regarded as the advantages of the shotgun method. However, the workload is large. If the shotgun method is used, the rearrangement of DNA fragments is not easy for computation at all. In the determination of large genomes, such as the human genome and drosophila genome and so on, the improved whole-genome "shotgun method" has been extensively applied to complete the sequencing, which can sufficiently demonstrate its feasibility and effectiveness [6, 7].

Scientists, researchers and scholars have developed many algorithms and tools for predicting the precise results based on some features of target objects [8–10]. Recently, the extensive application of these algorithms has been witnessed in genetic analysis. Genfrag, a set of tools, was developed to generate benchmark data sets for testing DNA sequence assembly algorithms and to quest for the range of data and corresponding performance of assembly tools on "shot-gun" sequencing projects by Engle and Burks [11]. An open-source bioinformatic tool, called Grinder, was introduced by Angly et al., which could simulate amplicon and shotgun datasets from reference sequences [12]. In the detection of respiratory viruses in clinical specimens, four different bioinformatics algorithms were executed by Huang et al. to make the assessment of the performance of a metagenomic shot-gun sequencing method [13]. Based on Sanger methodology, a novel algorithm was applied by Shityakov et al. [14], which correctly predicted and stressed the performance of DNA sequencing techniques and confirmed the statistical significance of results. Although many algorithms have recently been proposed to obtain the DNA fragment sequence, these algorithms are considerably complex and require much additional information apart from the lengths of DNA fragments, which may limit their application scopes.

In this paper, based on the biological information of each segment by SPDP and the mathematical thought of 0–1 planning, we propose the general basic algorithm to solve the feasible solutions from all permutations of the location of DNA restriction sites, and further restore the possible DNA sequence. Besides, we evaluate the efficiency of this algorithm according to 1000 sets of DNA original sequences randomly generated. Moreover, the influence of measurement error of fragments' length on the algorithm is discussed. The proposed algorithm can be conducted just based on data of fragments' length, and thus this algorithm is relatively easy to be applied in practice.

2 Example Design

2.1 Example 1

The first set of data is 2, 3, 7, 8, 8, 9, 13, 14.

The second set of data is 2, 1, 4, 3, 6.

2.2 Example 2

The first set of data is 1, 2, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 12, 13, 14.

The second set of data is 1, 1, 2, 1, 2, 2, 1, 2, 3.

3 Problem Analysis and Tentative Ideas

We need to find the correct sequence of DNA fragments represented by the second set of data, so that when the DNA molecule is cut at each restriction site respectively, the data obtained is consistent with the first set of data. Because of the specific data given in example 1, there are four restriction enzyme cutting sites in this DNA molecule, and the first group of data (2, 14, 8, 8, 7, 9, 3, 13) is obtained when the DNA molecule is cut on a single site, while the second group of data is (1, 2, 3, 4, 6) after DNA molecule is cut on all the restriction enzyme cutting sites. In the first set of data, the minimum fragment length is 2, and only 2 in the second set of data can correspond to it, so the corresponding enzyme cutting site should be the closest site to the endpoint; while the number 3 in the first set of data can correspond to 1 plus 2 or 3 in the second set of data. Obviously, the larger number is in the first set of data, the corresponding combination in the second set of data will be more.

Further analysis, we can get each number in the first group of data from one end according to each enzyme cutting point, and then the rearrangement of the second set of data is correct. That is because numbers in pairs in the first set of data represent the same meaning, such as 2 and 14 or 3 and 13, which only represent different restriction sites. We divide the first set of data into two groups, namely (2, 3, 7, 8) and (14, 13, 9, 8). Only one set of data (2, 3, 7, 8) represents the shorter distance between each restriction site and two ends of the DNA molecule. Therefore, we only need to analyze a half of the first data to express each restriction site. If the data of the length of DNA fragments obtained after the DNA molecule is cut on each restriction site at the same time can be the same as the second data (1, 2, 3, 4, 6), then the sequence is meaningful. Based on the above analysis, we use 0–1 planning to calibrate the shear point position and finally get the result. Specific implementation of the algorithm is as follows [15].

4 0–1 Planning Method

4.1 The Establishment of 0–1 Equation Algorithm

Suppose: the data of fragments' length measured when the DNA molecule is cut at each restriction site separately is the first set of data:

$$A = \left[a_1,aa_1,a_2,aa_2,a_3,aa_3,\cdots,a_n,aa_n\right],$$

where a_i, aa_i are two data from the same cutting experiment and $a_i \le aa_i$ while n is the number of restriction sites on DNA;

The data of fragments' length measured when the DNA molecule is cut at each restriction site simultaneously is the second set of data $B = \left[b_1, b_2, b_3, \cdots b_{n+1}\right]$.

The total length of the sequenced DNA molecule is $M = a_i + aa_i = b_1 + b_2 + b_3 + \cdots + b_{n+1}$.

After processing, the first set of data becomes $A = \begin{Bmatrix} a_1\, a_2\, a_3\, \cdots\, a_n \\ aa_1\, aa_2\, aa_3\, \cdots\, aa_n \end{Bmatrix}$, where $a_i \le \frac{M}{2}$ and $a_i + aa_i = M, i = 1, 2, ..., n$. Therefore, just one between a_i and aa_i can convey the meaning of the first group of data, then A can be expressed as: $A = \left[a_1,a_2,a_3,\cdots,a_n\right]$.

Each of $a_1,a_2,a_3,\cdots,a_n$ is the distance from the corresponding restriction site to the nearest endpoint. As shown in Fig. 1, obviously, each restriction site I_i is either on the half segment of DNA near the P endpoint or the half segment of DNA near the Q endpoint.

We suppose that the value of $x_1, x_2, x_3, \cdots x_n$ should only be 0 or 1, and generate the sequences below:

$$C = \left[a_1 \cdot \mathrm{x}_1,a_2 \cdot x_2,a_3 \cdot \mathrm{x}_3,\cdots,a_n \cdot x_n\right],$$

$$CC = \left[a_1 \cdot (1 - x_1),\ a_2 \cdot (1 - x_2),\ a_3 \cdot (1 - x_3),\cdots,a_n \cdot (1 - x_n)\right].$$

If $x_i = 1$, it means that the restriction site I_i is on the half segment of DNA near the P endpoint. Otherwise, I_i is on the half segment of DNA near the Q endpoint.

Sort the numbers in the C and CC from small to large to get a new sequence.

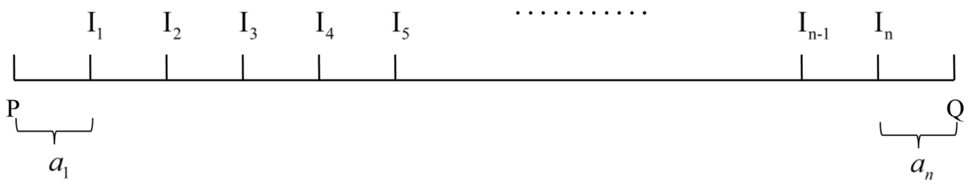

Fig. 1 Each restriction site I_i on the DNA molecule and corresponding a_i

$$C = \left[c_1, c_2, c_3, \cdots, c_n\right],$$
$$CC = \left[cc_1,\ cc_2,\ cc_3, \cdots, cc_n\right].$$

where C is the ascending order of $a_1 \cdot \mathrm{x}_1, a_2 \cdot x_2, a_3 \cdot \mathrm{x}_3, \cdots, a_n \cdot x_n$, and CC is the ascending order of $a_1 \cdot (1 - x_1),\ a_2 \cdot (1 - x_2),\ a_3 \cdot (1 - x_3), \cdots, a_n \cdot (1 - x_n)$.

The length of fragments between the adjacent restriction sites (including the two ends of DNA) on the DNA can be expressed as follows.

1. The length of each segment on the half segment of DNA near the P endpoint is:

$$\begin{cases} i = 1 \quad l_i = c_1 \\ 1 < i \le n \quad l_i = c_i - c_{i-1} \end{cases}$$

2. The length of each segment on the half segment DNA near the Q endpoint is:

$$\begin{cases} i = 1 \quad ll_i = cc_1 \\ 1 < i \le n \quad ll_i = cc_i - c_{i-1} \end{cases}$$

3. The length of the middle segment is:

$$M - (c_n + cc_n)$$

A value of 0 indicates no fragment here.

Use the length of the fragments above to build a sequence:

$$S = \left[l_1,\ l_2,\ l_3, \cdots, l_n, ll_1,\ ll_2,\ ll_3, \cdots, ll_n, M - (c_n + cc_n)\right].$$

Then the elements in the sequence are sorted from small to large.

$$S = \left[s_1,\ s_2,\ s_3, \cdots, s_{2n+1}\right]$$

Because the point I_i is either on the half of DNA near the P endpoint or on the half of DNA near the Q endpoint (that is, x_i is equal to 0 or 1), there are n zeros in $2n$ numbers in C and CC. The n zeros in the sequence S are removed to get the sequence SS representing the length of the segments between the restriction sites (including the two ends of DNA) on DNA.

$$SS = \left[s_{n+1},\ s_{n+2},\ s_{n+3}, \cdots, s_{2n+1}\right].$$

Sort the elements in the sequence B from small to large, and get:

$$B = \left[bb_1, bb_2, bb_3, \cdots bb_{n+1}\right]$$

$bb_1, bb_2, bb_3, \cdots bb_{n+1}$ is ascending order for $b_1, b_2, b_3, \cdots b_{n+1}$.

Therefore, assuming that SS and B are exactly the same sequences (that is, $SS = B$), and we can establish equations as follows:

$$\begin{cases} s_{n+1} = bb_1 \\ s_{n+2} = bb_2 \\ s_{n+3} = bb_3 \\ \cdots\cdots \\ s_{2n+1} = bb_{n+1} \end{cases}$$

By programming on Matlab7.1, we can get the position of each restriction site on the original DNA molecule by solving $x_1, x_2, x_3, \cdots x_n$, and then calculate the sequence of

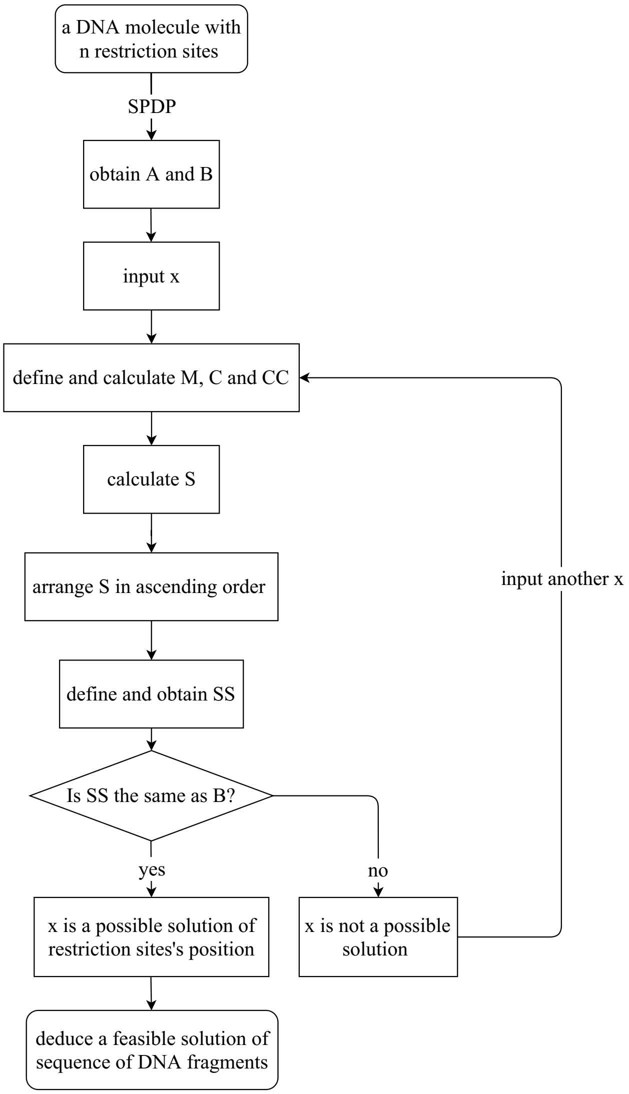
a DNA molecule with n restriction sites
SPDP
obtain A and B
input x
define and calculate M, C and CC
calculate S
arrange S in ascending order
define and obtain SS
Is SS the same as B?
yes
no
x is a possible solution of restriction sites's position
x is not a possible solution
input another x
deduce a feasible solution of sequence of DNA fragments

Fig. 2 Flow chart of the proposed algorithm

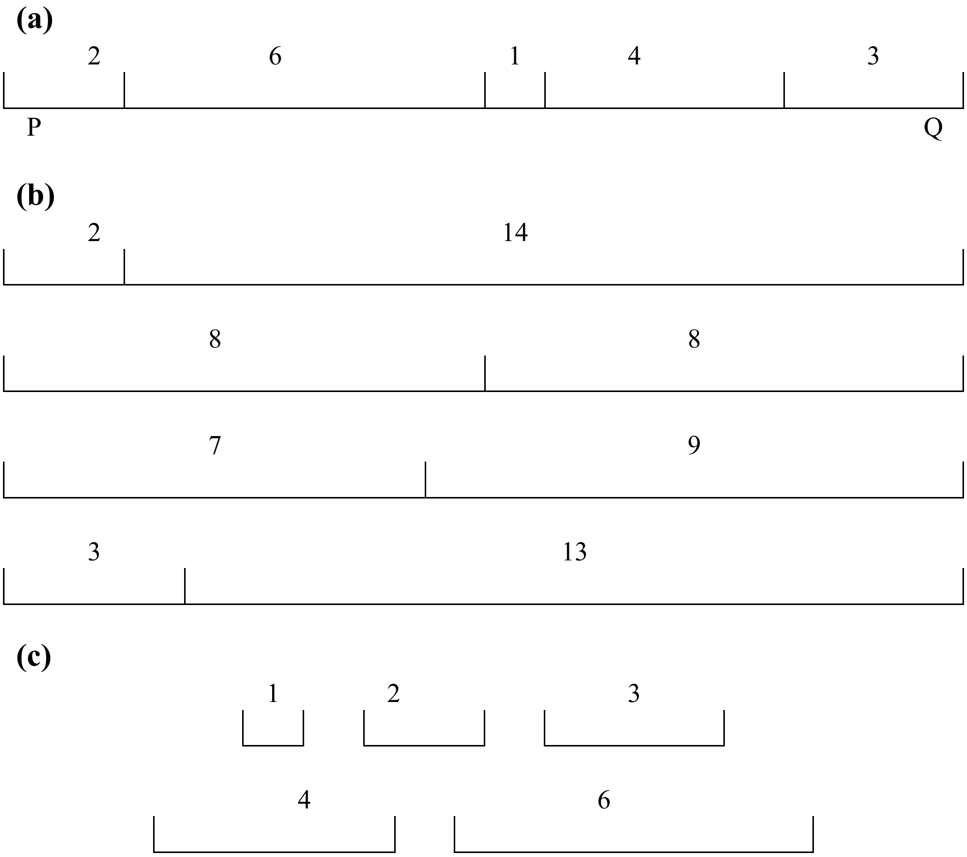

Fig. 3 **a** An instance of the original DNA sequence (that is, the true sequence that we try to reconstruct). **b** The first set of data $A = [2,3,7,8]$. **c** The second set of data $B = [1,2,3,4,6]$

the original DNA molecule. The flow chart of the proposed algorithm is shown in Fig. 2.

To clarify the algorithm, part of the solution process of example 1 is shown as in Fig. 3.

For example, we input $x = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}$ here, and then the process of calculation is as follows.

$$M = 1 + 2 + 3 + 4 + 6 = 16$$

$$C = \begin{bmatrix} 2 & 0 & 0 & 0 \end{bmatrix} \xrightarrow{\text{sort ascending}} C = \begin{bmatrix} 0 & 0 & 0 & 2 \end{bmatrix}$$

$$CC = \begin{bmatrix} 0 & 3 & 7 & 8 \end{bmatrix} \xrightarrow{\text{sort ascending}} CC = \begin{bmatrix} 0 & 3 & 7 & 8 \end{bmatrix}$$

$$\begin{cases} i = 1 \quad l_i = c_1 \\ 1 < i \le n \quad l_i = c_i - c_{i-1} \end{cases}$$

$l_1 = 0, l_2 = 0, l_3 = 0, l_4 = 2.$

$$\begin{cases} i = 1 \quad ll_i = cc_1 \\ 1 < i \le n \quad ll_i = cc_i - c_{i-1} \end{cases}$$

$ll_1 = 0, ll_2 = 3, ll_3 = 4, ll_4 = 1.$

$M - (c_n + cc_n) = 16 - 2 - 8 = 6$

$$S = \left[l_1,\ l_2,\ l_3, \cdots, l_n, ll_1,\ ll_2,\ ll_3, \cdots, ll_n, M - (c_n + cc_n)\right] \xrightarrow{\text{sort ascending}}$$
$$S = \left[s_1,\ s_2,\ s_3, \cdots, s_{2n+1}\right] = [0, 0, 0, 0, 1, 2, 3, 4, 6]$$

$$SS = \left[s_{n+1},\ s_{n+2},\ s_{n+3}, \cdots, s_{2n+1}\right] = [1, 2, 3, 4, 6]$$

Because SS is the same as B, $x_1, x_2, x_3, \cdots x_n$ is the possible permutation of the location of DNA restriction sites and a feasible solution of DNA fragments' sequence [2,6, 1, 4, 3] can be restored. The reconstructed sequence is exactly the same as Fig. 3a.

5 Implementation

First, we build a general 0–1 algorithm for the general situation, and then apply this algorithm to solve example 1 and example 2.

5.1 Example 1

The first set of data is 2, 3, 7, 8, 8, 9, 13, 14, then: $A = [2,3,7,8]$.

The second set of data is 2, 1, 4, 3, 6, then: $B = [2, 1, 4, 3, 6][2, 1, 4, 3, 6]$.

Total DNA length: $M = 2 + 1 + 4 + 3 + 6 = 16$.

By Matlab7.1 (see Appendix procedure 1), we solve out

$$x = \begin{bmatrix} 1000 \\ 1001 \\ 0110 \\ 0111 \end{bmatrix}.$$

The possible sequences of fragments represented by the second set of data are as follows:

2 6 1 4 3
2 6 1 4 3
3 4 1 6 2
3 4 1 6 2

If there is no difference between the two ends of P and Q, there is only one solution:

2 6 1 4 3

(or 3 4 1 6 2).

If P and Q are different in order, there are two solutions:

2 6 1 4 3
3 4 1 6 2

5.2 Example 2

The first set of data is 1, 2, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 12, 13, 14, then:

$A = [1,2,3,3,4,5,6,7]$

The second set of data is 1, 1, 1, 1, 2, 2, 2, 2, 3, then: $B = [1, 1, 1, 1, 2, 2, 2, 2, 3]$.

Total DNA length: $M = 1 + 1 + 1 + 1 + 2 + 2 + 2 + 2 + 3 = 15$.

By Matlab7.1 (see Appendix procedure 2), we obtain that:

$$x = \begin{bmatrix}
0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 & 0 & 1
\end{bmatrix}$$

The possible sequences of fragments represented by the second set of data are as follows:

```
3 1 2 2 2 2 1 1 1
3 2 2 2 2 1 1 1 1
3 1 2 2 2 2 1 1 1
3 2 2 2 2 1 1 1 1
1 2 2 3 1 2 1 1 2
1 2 3 2 2 1 1 1 2
1 2 1 3 2 1 2 1 2
1 2 2 1 2 3 1 1 2
1 2 1 1 2 2 3 1 2
1 2 1 2 1 3 2 1 2
1 2 2 3 1 2 1 1 2
1 2 3 2 2 1 1 1 2
1 2 1 3 2 1 2 1 2
1 2 2 1 2 3 1 1 2
1 2 1 1 2 2 3 1 2
1 2 1 2 1 3 2 1 2
2 1 2 3 1 2 1 2 1
2 1 3 2 2 1 1 2 1
2 1 1 3 2 1 2 2 1
2 1 2 1 2 3 1 2 1
2 1 1 1 2 2 3 2 1
2 1 1 2 1 3 2 2 1
2 1 2 3 1 2 1 2 1
2 1 3 2 2 1 1 2 1
2 1 1 3 2 1 2 2 1
2 1 2 1 2 3 1 2 1
2 1 1 1 2 2 3 2 1
2 1 1 2 1 3 2 2 1
1 1 1 1 2 2 2 2 3
1 1 1 2 2 2 2 1 3
1 1 1 1 2 2 2 2 3
```

If there is no difference between P and Q, there are 8 groups of solutions:

```
3     1 2 2 2 2 1 1 1
(or1  1 1 2 2 2 2 1 3)
3     2 2 2 2 1 1 1 1
(or1  1 1 1 2 2 2 2 3)
1     2 2 3 1 2 1 1 2
(or2  1 1 2 1 3 2 2 1)
1     2 3 2 2 1 1 1 2
(or2  1 1 1 2 2 3 2 1)
1     2 1 3 2 1 2 1 2
(or2  1 2 1 2 3 1 2 1)
1     2 2 1 2 3 1 1 2
(or2  1 1 3 2 1 2 2 1)
1     2 1 2 1 3 2 1 2
(or2  1 2 3 1 2 1 2 1)
1     2 1 1 2 2 3 1 2
(or2  1 3 2 2 1 1 2 1)
```

If P and Q are different in order, there are 16 groups of solutions:

```
3 1 2 2 2 2 1 1 1
3 2 2 2 2 1 1 1 1
1 2 2 3 1 2 1 1 2
1 2 3 2 2 1 1 1 2
1 2 1 3 2 1 2 1 2
1 2 2 1 2 3 1 1 2
1 2 1 2 1 3 2 1 2
1 2 1 1 2 2 3 1 2
1 1 1 2 2 2 2 1 3
1 1 1 1 2 2 2 2 3
2 1 2 3 1 2 1 2 1
2 1 3 2 2 1 1 2 1
2 1 1 3 2 1 2 2 1
2 1 2 1 2 3 1 2 1
2 1 1 1 2 2 3 2 1
2 1 1 2 1 3 2 2 1
```

To further illustrate the practical significance of our solution, a set of solution is extracted from the result of example 2 for instance.

1 2 2 3 1 2 1 1 2

The solution above reconstructs a DNA sequence as in Fig. 4.

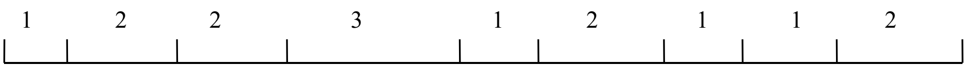

Fig. 4 A possible sequence of example 2

6 Error Analysis

Considering all kinds of factors, we think that the measurement of the length of the fragments is the main cause of the error. Assuming that there is no error in the total length of DNA, then the sum of the two data obtained when the DNA molecule is cut on each restriction site separately and the sum of all data obtained when the DNA molecule is cut on each restriction sites at the same time are the same and equal to the total length of DNA. According to the problem analysis, we briefly discuss the impact of the error on the results in two cases:

1. When the same error occurs in the measurement of fragments in the first set of data and corresponding fragments in the second set of data, the data change is equivalent to the data change caused by the change of the position of the corresponding restriction site in DNA. At this time, the result of the reconstruction of the restriction map will show the change of those restriction sites corresponding to the error data in DNA molecule. The determination of other restriction sites will not be affected. For example:

Suppose the real data of example 1 is:

The first set of data: 2, 14, 8, 8, 9, 7, 13, 3.

The second set of data: 2, 1, 4, 3, 6.

Assume that the data obtained due to the measurement error are:

The first data: 2, 14, 7, 9, 9, 7, 13, 3.

The second set of data: 2, 2, 4, 3, 5.

The x of the error data is solved by Matlab 7.1 program (see Appendix program 3):

$$x = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

The possible sequences of fragments represented by the second set of data are as follows:

2 5 2 4 3
2 5 2 4 3
3 4 2 5 2
3 4 2 5 2

If there is no difference between P and Q, the final result can be expressed as follows:

2 5 2 4 3

The results from real data are as follows:

2 6 1 4 3

It can be seen that the data error only leads to the change of the position of the restrictive sites that produce the error, and has no effect on the reconstruction of the position of other restrictive sites.

2. When the error only appears in the first set or the second set of data, or the first set and the second set of data produce unrelated errors at the same time, there may be a variety of results: most of the restrictive site changes, that is, the calculation results can be regarded as invalid, or reconstruction cannot be carried out. For example: Suppose the real data of example 1 is:

The first set of data: 2, 14, 8, 8, 9, 7, 13, 3.

The second set of data: 2, 1, 4, 3, 6.

Assume that the data obtained due to the measurement error are:

The first data: 2, 14, 7, 9, 9, 7, 13, 3.

The second set of data: 2, 1, 4, 3, 6.

The x of the error data is solved out by Matlab 7.1 program (see Appendix program 4): $x = \emptyset$.

The reconstruction cannot proceed due to the error.

7 Evaluation of the Model

The function (randi) of Matlab 7.1 software is used to randomly generate 1000 groups of original DNA sequences. We try to reconstruct the original DNA sequence based on data of fragments' length by SPDP using 0–1 algorithm.

The coincidence rate (including multiple solutions or unique solutions) and the unique coincidence rate (that is, reconstruction solution is unique and exact compared with the original DNA sequence) between the reconstructed DNA sequence and the original DNA sequence are defined.

First, 1000 sets of DNA sequences are randomly generated, and the second set of data consists of a set of random numbers (DNA fragments' length) between 1 and 30. The effect of the number of DNA fragments on the coincidence rate and the unique coincidence rate was studied.

It can be seen from Fig. 5 that the curve of coincidence rate is above 90%, and the curve of unique coincidence rate is above 80%. Especially when the number of fragments is greater than 6, the coincidence rate reaches 100%. With the

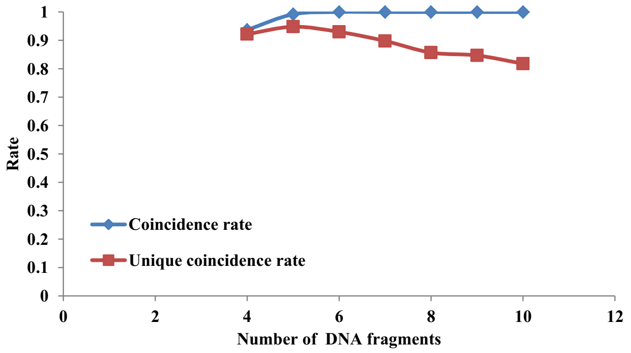

Fig. 5 Statistical analysis of the influence of different numbers of DNA fragments on the effectiveness of the algorithm

increase of the number of fragments, however, the unique coincidence rate will decrease (that is, multiple solutions will appear more).

Second, 1000 sets of DNA sequences are randomly generated, while the number of DNA fragments of the second set of data is 5, and the length of each fragment is a random number between 1 and M, where M is the maximum length of DNA fragments. We study the effect of the magnitude of M on the coincidence rate and unique coincidence rate.

It can be seen from Fig. 6 that the coincidence rate between the DNA sequence calculated by this algorithm and the original DNA sequence is above 98%, and the unique coincidence rate is above 80%. As the maximum length of DNA fragments becomes larger, the unique coincidence rate increases.

As shown in Figs. 5 and 6, the high coincidence rate and unique coincidence rate are observed, validating the effectiveness of the proposed algorithm.

8 Conclusions and Remarks

Our data and analysis support the advantages of the algorithm: (1) The algorithm makes full use of the search method and 0–1 planning knowledge, and optimizes the arrangement of different DNA fragments to find the most satisfactory solution; (2) In terms of operation, it simplifies the difficulty of artificial combination and pure mathematical reasoning, and provides a relatively fast and accurate method for high-throughput and large-scale DNA sequencing; 3. We try to simplify the variables of the data, gradually approach the length of each segment, arrange them in ascending order, and finally use different sorting results to set up equations with the data related.

However, this is an algorithm related to biological background, which means that there exists the uniqueness of objective facts. Due to the limitations of the conditions given, we are not able to determine which group of solutions is exactly the sequence of the original DNA in the face of multiple groups of solutions. For example, there are fragments with the same length in example 2. This algorithm starts from the length of the DNA fragments, but the possible situation where DNA fragments of the same length may represent different sequences is ignored inevitably, and thus the result of this solution is one-sided.

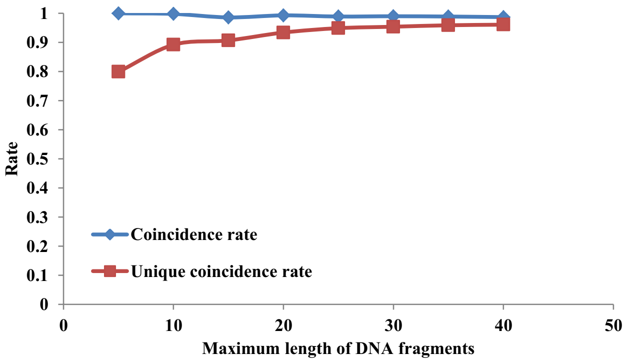

Fig. 6 Statistical analysis of the influence of the maximum length of DNA fragments on the effectiveness of the algorithm

Our algorithm can not only analyze genetic samples and DNA sequencing, integrate biological information of each segment but also be extended to other related life science fields like synthetic biology. In addition, if we can integrate biological knowledge and consider all kinds of variation factors, for instance, insertion, deletion and replacement of base pairs under experimental conditions, the algorithm will have a broader application prospect and solve more practical problems of biological genetic analysis.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12539-021-00419-6.

Acknowledgements This work is supported by the National Natural Science Fund of P. R. China (Grant No. 11771448). We thank Prof. Junjie Ren (School of Sciences, Southwest Petroleum University, Chengdu 610500, Sichuan, China) for his help in revising the English grammar. We thank Dr. Qiang Pan and Dr. Fan Jiang (Army Medical University) for their help in manuscript design.

Authors' contributions DT and KW designed the study. NF collected and processed the data. DT, KW and NF analyzed the results. DT, KW and NF wrote and revised the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflicts of interest The authors declare that there is no conflict of interest (financial and non-financial).

References

1. Viswanathan R, Cheruba E, Cheow LF (2019) DNA Analysis by Restriction Enzyme (DARE) enables concurrent genomic and epigenomic characterization of single cells. Nucleic Acids Res 47:e122. https://doi.org/10.1093/nar/gkz717
2. Cameron CJ, Dostie J, Blanchette M (2020) HIFI: estimating DNA-DNA interaction frequency from Hi-C data at restriction-fragment resolution. Genome Biol 21:11. https://doi.org/10.1186/s13059-019-1913-y
3. Alza L, Lavretsky P, Peters JL, Ceron G, Smith M, Kopuchian C, Astie A, McCracken KG (2019) Old divergence and restricted gene flow between torrent duck (*Merganetta armata*) subspecies in the Central and Southern Andes. Ecol Evol 9:9961–9976. https://doi.org/10.1002/ece3.5538
4. Maschmann A, Masters C, Davison M, Lallman J, Thompson D, Kounovsky-Shafer KL (2018) Determining if DNA stained with a cyanine dye can be digested with restriction enzymes. J Vis Exp. https://doi.org/10.3791/57141

5. Staden R (1982) Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Res 10:4731–4751. https://doi.org/10.1093/nar/10.15.4731
6. Venter JC, Adams MD, Myers EW (2001) The sequence of the human genome. Science 291(5507):1304. https://doi.org/10.1126/science.1058040
7. Myers EW, Sutton GG, Delcher AL et al (2000) A whole-genome assembly of Drosophila. Science 287(5461):2196–2204. https://doi.org/10.1126/science.287.5461.2196
8. Abualigah LM (2019) Feature selection and enhanced krill herd algorithm for text document clustering. Springer, Berlin, pp 1–165. https://doi.org/https://doi.org/10.1007/978-3-030-10674-4
9. Abualigah LM, Khader AT, Hanandeh ES (2018) Hybrid clustering analysis using improved krill herd algorithm. Appl Intell 48(11):4047–4071. https://doi.org/10.1007/s10489-018-1190-6
10. Abualigah LM, Khader AT, Hanandeh ES (2018) A new feature selection method to improve the document clustering using particle swarm optimization algorithm. J Comput Sci 25:456–466. https://doi.org/10.1016/j.jocs.2017.07.018
11. Engle ML, Burks C (1993) Artificially generated data sets for testing dna sequence assembly algorithms. Genomics 16(1):288. https://doi.org/10.1006/geno.1993.1180
12. Angly FE, Dana W, Forest R, Philip H, Tyson GW (2012) Grinder: a versatile amplicon and shotgun sequence simulator. Nuclc Acids Research 40(12):e94. https://doi.org/10.1093/nar/gks251
13. Huang W, Wang G, Lin H, Zhuge J, Nolan SM, Vail E, Dimitrova N, Fallon JT (2016) Assessing next-generation sequencing and 4 bioinformatics tools for detection of enterovirus d68 and other respiratory viruses in clinical samples. Diagn Microbiol Infect Dis 85(1):26–29. https://doi.org/10.1016/j.diagmicrobio.2016.01.013
14. Shityakov S, Bencurova E, Frster C, Dandekar T (2020) Modeling of shotgun sequencing of dna plasmids using experimental and theoretical approaches. BMC Bioinformatics. https://doi.org/10.1186/s12859-020-3461-6
15. Guo JY, Lu WX, Yang QC, Miao TS (2019) The application of 0–1 mixed integer nonlinear programming optimization model based on a surrogate model to identify the groundwater pollution source. J Contam Hydrol 220:18–25. https://doi.org/10.1016/j.jconhyd.2018.11.005