IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 13, 2020, accepted December 22, 2020, date of publication December 25, 2020,
date of current version January 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047469

Bacterial Foraging Algorithm Based on Activity of Bacteria for DNA Computing Sequence Design

YAO YAO, JIANKANG REN, (Member, IEEE), RAN BI, AND QIAN LIU, (Member, IEEE)
School of Computer Science and Technology, Dalian University of Technology, Dalian 116000, China

Corresponding author: Jiankang Ren (rjk@dlut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62072067, Grant 61602080, Grant 61761136019, Grant 61872436, Grant 61601080, Grant 61602084, and 61772112; in part by the National Key Research and Development Program of China under Grant 2017YFC0704200; in part by the China Postdoctoral Science Foundation under Grant 2018T110221 and Grant 2016M591431; in part by the Social Science Foundation of Liaoning Province under Grant L17CTQ002; in part by the Natural Science Foundation of Liaoning Province under Grant 20180550540; and in part by the Fundamental Research Funds for the Central Universities under Grant DUT18RC(4)008.

ABSTRACT Since the quantity and quality of DNA sequence directly affect the accuracy and efficiency of computation, the design of DNA sequence is essential for DNA computing. In order to improve the efficiency and reliability of DNA computing, there is a rich literature targeting at generating DNA sequences with lower similarity that can hybridize at a lower melting temperature. However, it is not trivial to improve both melting temperature and similarity for the DNA sequence, since DNA sequence design problem under the constraints of Hamming distance, secondary structure and molecular thermodynamic is known to be NP-hard. For the sake of achieving the lower melting temperature and similarity for the generated DNA sequence, we proposed an improved method for the bacterial foraging algorithm based on activity of bacteria (BFA-A). In particular, the effect of bacterial vitality on foraging ability is considered, and a competitive exclusion mechanism is introduced to improve the quality of the generated DNA sequences. In BFA-A, high-quality DNA strands are replicated to avoid the participation of inferior strands in the operation, and the active regulation mechanism and the competitive rejection mechanism are used to improve and accelerate the chemotaxis process. Experiments show that our proposed approach significantly outperforms existing methods in terms of melting temperature and similarity. In addition, the experimental results also show that our method can reduce the number of iterations, and has guiding significance to generate high-quality DNA sequences more efficiency.

INDEX TERMS DNA computing, DNA sequence design, optimization methods.

I. INTRODUCTION

DNA computing is one of novel computational models by combining computer science and biological science. Owing to biological characteristics, DNA computing can take advantage of the high parallelism, high storage density and low power consumption. Therefore, it has great potential in solving complex combinatorial optimization problems such as NP-complete problems, which making it is one of the important ways to develop non-traditional high-performance computing. A large number of studies have shown that DNA computing can be widely used in various fields, such as Hamiltonian loop problem solution [1], artificial neural network establishment [2], handwritten content recognition [3], circuit logic gate design [4]–[6], Nano robot design [7]–[9], DNA origami nanomaterial design [10], image encryption [11]–[13], and storage technology innovation [14]. With the development of biotechnology, it is conceivable that DNA computing will be used to solve more realistic problems. For instance, in the latest research of the synthetic molecular science, 4 new bases ('Z' and 'P' are complementary; 'S' and 'B' are complementary) have been artificially synthesized [15]. The number of DNA bases has been expanded from 4 to 8, which enables a richer coding format of DNA, and thus effectively reducing the similarity between DNA sequences.

Although researchers from different fields such as computer science, biology, and mathematics research on DNA computing from various aspects, there are still lots of theoretical challenges to truly realize DNA computers. In particular,

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaokang Wang.

DNA sequence coding is the major problem, since the quantity and quality of the generated DNA sequence have a strong influence on the reliability and accuracy of DNA computing [16]. Generally, low similarity and low melting temperature are required to reduce the probability of mismatch hybridization and to increase the reaction efficiency [17]. This can be explained by the fact that the DNA sequences involved in the reaction need to be simultaneously melted to ensure that each sequence reacts with each other. For instance, we consider 2 DNA strands X_1 and X_2, and they hybridize with each other. We suppose that the melting temperature (Tm) of X_1 is 59 °C and the Tm of X_2 is 67 °C. Therefore, when the environment temperature reaches 59 °C, X_1 will melt, but X_2 cannot. Due to the large difference between the Tm of DNA strands X_1 and X_2, the efficiency of the reaction is greatly reduced. Moreover, multiple constraints on the biological properties should be satisfied for the DNA sequences to decrease the error of DNA hybridization reaction. Generally, these constraints are defined based on molecular thermodynamic, secondary structure, Hamming distance, etc. Therefore, the problem of DNA sequence coding can be transformed into a multi-objective optimization problem with low reaction temperature and low mismatch hybridization as the main objective. In the previous work, researchers generally considered six constraints, and they are melting temperature (Tm), GC base content (GC), Continuity (Con), Hairpin structure ($Hair$), Similarity (Sim) and H-measure (Hm). To achieve the objective of low mismatch, false negative and false positive hybridization should be prevented in DNA sequences generation. Moreover, in order to improve the efficiency of reaction, all the DNA sequences should be melted at a similar lower temperature, and it is also one of the main objectives.

The problem of DNA sequences coding is a typical multi-objective and multi-constraint problem, which is known to be NP-Hard [18]. Among multiple optimization goals, to generate the DNA sequences with lower Tm is the main goal. However, due to multiple constraints and biochemical characteristics, it is not trivial to achieve this goal [19], [20]. In our previous work, we have effectively reduced the Tm of the generated DNA sequence by bacterial foraging algorithm (BFA) [21]. Different from the other existing methods, the half-replication characteristic of the BFA is used to eliminate inferior DNA sequences. With this characteristic, we can maintain the intrinsic similarity between different sequences, and thus the Tm variance of the multiple sequences can be reduced. A large number of experiments have shown that the coding method based on BFA can generate DNA strands with lower Tm. In addition, we also give an evaluation strategy to better quantify the quality of DNA sequences.

As an extension of this work, in this paper, we make the following further contributions:

- Active regulation mechanism and competition exclusion mechanism are considered in BFA to improve the generating efficiency and the quality of DNA sequences;
- A special relationship between Sim and Hm that is approximately equal to a fixed value is found and analyzed;
- A novel constraint on special DNA sequence pieces that can be recognized by endonucleases is proposed.

Extensive experiment demonstrates that the proposed method outperforms the existing methods in terms of Tm, Sim and Hm, especially Tm.

The remainder of the paper is structured as follows. In Section II, we discuss the related work. In Section III, we present the objectives and constraints of the problem. The detailed description of BFA-A based DNA computing sequence coding algorithm is given in Section IV. In Section V, we present the experimental results and finally conclude in Section VI.

II. RELATED WORKS

DNA sequence coding design is a complex problem, since it has different design goals for different experimental purposes, and the coding itself subject to the biological characteristics of DNA needs to meet various constraints. The initial theoretical basis for coding design is still not sufficient. Therefore, the early researchers often selected the expected sequences from a large number of generated codes. Hartemink *et al.* [22] used an exhaustive method to traverse all possible coding combinations. Frutos and Thiel [23], Arita and Kobayashi [24], and Liu *et al.* [25] etc. respectively used a specially designed template and mapping strategy to select dissimilar sequences from a large number of automatically generated sequences. This method of screening from a large number of codes is simple and straight forward, but the screening work often takes a long time due to the large number of potential DNA sequences. In the early research, there is a coding design based on graph theory. Feldkamp *et al.* [26] designed the DNA sequence based on the directed graph. The idea of this method is relatively complicated, and the obtained code cannot well meet the actual requirements. In the subsequent research, researchers often aimed at reducing melting temperature and similarity, and improving stability. At the same time, with the study of the biological characteristics of DNA, researchers began to consider the multiple constraints of coding design. The coding design problem is formed as a single-objective multi-constraint optimization problem, or multi-objective multi-constraint optimization problem. Usually the reduction of melting temperature and similarity and the optimization of stability and code acquisition efficiency are considered as the objective, and hairpin structure, repeatability, GC base content, Hamming distance, etc. are considered as constraints. After that, a large number of coding design methods based on heuristic algorithms are proposed by solving the formed multi-constraint optimization problem.

In view of Hamming distance constraints, there are some methods to reduce coding similarity. Marathe *et al.* [27] adopt the dynamic programming technology to design

DNA sequences based on Hamming distance. Tanaka and Nakatsugawa [28] proposed a method based on simulated annealing to generate more stable and reliable sequences. Zhang *et al.* [29] proposed an algorithm based on the minimum free energy criterion to generate DNA sequences with almost no false hybridization. Zhang *et al.* [30] also proposed a method based on the improved genetic algorithm to design the stable DNA sequences. Compared with the original genetic algorithm for the DNA coding, this method can effectively improve the sequence stability. Luo and Luo [31] proposed a method based on weed algorithm to generate DNA sequences with high Hamming distance. Yang *et al.* [16] proposed a niche-based weed algorithm to improve the lack of the original weed algorithm for DNA coding to reduce the similarity and improve the stability.

There are also several DNA sequence design methods considering the constraints on melting temperature, free energy and stability. For example, Shin *et al.* [32]. proposed a multi-objective evolutionary algorithm to solve the DNA sequence design problem using the nucleic acid computing simulation toolkit/sequence generator (NACST) system. Xu and Zhang [33], [34] proposed a genetic algorithm optimized by particle swarm to generate DNA sequences with low *Tm*. Guo *et al.* [35] proposed a coding design method based on Bloch's quantum chaos algorithm, which dynamically adjusts the quantum rotation angle to obtain DNA codes with better thermodynamic characteristics. Liu and Wang [36] combined the bat algorithm and particle swarm algorithm to design the DNA code with good stability and thermodynamic properties.

To improve the efficiency of DNA sequence acquisition, Yin and Ye [37], [38] proposed a particle swarm optimization algorithm based on cultural evolution, which combines cultural evolutionary algorithm with particle swarm optimization. By taking full advantage of evolutionary information, their methods can improve the search ability and the efficiency of DNA code acquisition. Xiao and Jiang [39] proposed a dynamic membrane evolution algorithm, which uses particle swarm optimization to update and improve the global search ability by using adaptive differential evolution algorithm. Choong and Lee [40] proposed a DNA sequence generation method based on convolutional neural networks, which can obtain DNA codes very quickly. Wang and Shen [41] proposed an improved non-dominated sorting genetic algorithm II (INSGA-II) for DNA coding design, which introduces constraints into the non-dominated sorting process, which has good convergence and good population diversity. Liu and Wang [42] modified the simulated annealing algorithm and the group search algorithm, and combined them into a hybrid iterative search algorithm to improve the efficiency of the DNA sequence generation.

All of the above methods can optimize the quality of the generated DNA sequences to a certain degree, but they are not capable of producing DNA sequences with sufficiently low *Tm* and *Sim*. In order to address this problem, we propose a DNA sequence design method based on the bacterial foraging algorithm (BFA). It should be noted that the bacterial foraging algorithm has been extensively used in image matching and job scheduling, etc. [43], [44] to avoid exhaustive search and improve search efficiency. It has the advantages of convenient and fast convergence owing to its parallel search of the swarm. In addition, it does not require gradient information of the object in the optimization process and has strong versatility. Although the BFA method for DNA sequence design can effectively reduce the variance of the *Tm*, it cannot generate the DNA sequences with sufficiently low *Sim* and *Hm*, since the replication mechanism can reduce bacterial diversity.

In order to improve the problem of insufficient optimization of *Sim* and *Hm* in the BFA method, we propose the bacterial foraging algorithm based on activity of bacterial (BFA-A) by introducing activity regulation mechanism and competitive exclusion mechanism. These two mechanisms play a critical role in the process of chemotaxis to effectively reduce the melting temperature and similarity. BFA-A is composed by the four behaviors (i.e., initialization, chemotaxis, replication, and dispersion), which is an abstraction of the natural growth of bacterial colonies [45], [46].

III. OBJECTIVES AND CONSTRAINTS

In order to obtain DNA sequences with error-free hybridization at lower melting temperature, we consider six constraints based on the basic biological properties of the double helix structure, and they are Similarity (*Sim*), H-measure (*Hm*), Continuity (*Con*), Hairpin structure (*Hair*), melting temperature (*Tm*), and GC base pair content (*GC*). With these constraints, reaction temperature, the occurrence of mismatch hybridization and the complexity of the synthesis can be decreased and the stability of the DNA strand will be improved. Note that, considering that biological reactions often require multiple enzymes, we also introduce the seventh constraint, i.e., specific sequence constraints. That specific DNA fragments can be recognized by several endonucleases for specific tailoring. The cleavage site forms a blunt end or a sticky end, and some special portions can be inserted into these blunt or sticky end, such as fluorophores, then the end can be combined by the DNA ligase. Different from the constraints based on biological characteristics, the seventh constraint is a constraint condition oriented to actual experimental requirements, which makes coding design move from theoretical design to practical application.

A. OPTIMIZATION OBJECTIVE

In order to improve the reaction efficiency and reduce the mismatch probability, Similarity is considered as one of optimization targets. Low *Tm* means low average and low variance of *Tm*. Low Similarity means low *Sim* and low *Hm*. The DNA strands which participate in reaction can reform a stable double helix, and thus the experiments result can be detected or separated easily. Low *Tm* and low *Sim (Hm)* could improve the stability of reformed double helix. In this

paper, we try to design DNA sequences with low *Tm* and low *Sim* (*Hm*).

B. CONSTRAINTS

1) MELTING TEMPERATURE

Melting temperature is defined as the temperature at which 50% of the base pairs in double-strand DNA are melted into a single strand. In the DNA computing experiment, all DNA strands should be placed in the same solution, and the temperature is slowly raised up to reach *Tm*. At the same time, all DNA strands are required to be simultaneously melted. At this temperature, the double helix structure starts to open to form a single strand, and then they recover to the double helix structure with the temperature dropping. Therefore, the lower *Tm* implies higher reaction efficiency. For this reason, low average temperature and especially low variance of temperature should be satisfied to ensure the reaction efficiency. In this paper, *Tm* is calculated based on the nearest-neighbor model [47].

2) SIMILARITY

Similarity is used to describe the degree of similarity between the base compositions of two DNA sequences. In order to ensure low mismatch hybridization, the isotropic sequence of two DNA strands should be as unique as possible, and the sequence should be not repeated after the sliding. The similarity is defined based on the sliding hamming distance [48], i.e.,

$$Sim = \sum_{j=1}^{n} min_{-m<k<m}^{j=1:n\&\&i\neq j}[Ham(X_i, shift(X_j^k))] \tag{1}$$

where n represents the total number of DNA strands, $Ham(X_i, shift(X_j^k))$ is the Hamming distance between X_i and X_j, and m is the total number of bases in each DNA strand. For sliding sequence X_j, $k > 0$ means that X_j slides to the right with $|k|$ step size, and $k < 0$ means that X_j slides to the left with $|k|$ step size. $min_{-m<k<m}^{j=1:n\&\&i\neq j}$ indicates the minimum Hamming distance between X_i and X_j, when X_j slides from the left end to the right end.

3) H-MEASURE

H-measure is used to describe the degree of similarity between the base compositions of DNA sequences X_i and its complementary strand X_j^c. It is defined as follows [48]:

$$Hm = \sum_{j=1}^{n} min_{-m<k<m}^{j=1:n}[Ham(X_i, shift(X_j^{ck}))] \tag{2}$$

where the meaning of n, m and k is same as (1), and the Hamming distance between X_i and sliding complementary sequence X_j^c is defined as $Ham(X_i, shift(X_j^{ck}))$. Here, sequence X_j^c is slidable. $min_{-m<k<m}^{j=1:n}$ indicates the minimum Hamming distance between X_i and X_j^c, where X_j^c slides from the left end to the right end.

H-measure and Similarity can influence the probability of a mismatch hybridization in the experiment. For instance, Fig. 1 shows the DNA sequences X_1 and X_2 with different Hamming distances at different relative positions, and the resulting incorrect hybridization. For X_1 and X_2 at the relative position in Fig. 1 (1), there is only one identical base. However, when X_2 slides to right until the relative position in Fig. 1 (2) is achieved, there are eight identical bases. Too much identical bases will cause DNA strand X_1 to make a partial hybridization with X_2^C (the complementary strand of X_2) as shown in Fig. 1 (3), and this kind of hybridization should be avoided. Therefore, the DNA sequences with too much identical bases should be avoided in the DNA sequence coding.

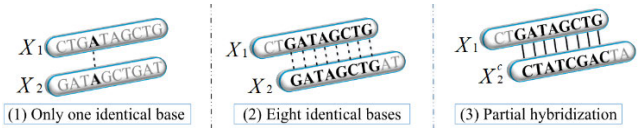

FIGURE 1. Reason for judging the sliding Hamming distance.

4) GC BASE PAIR CONTENT

GC content is a kind of molecular thermodynamic constraint. Because the content of GC base affects the stability and *Tm* of the DNA strand, it has to be considered as a constraint in the DNA sequence optimization. The GC content of the DNA sequence is defined as the relative percentage of bases 'G' and 'C', i.e.,

$$GC = \frac{sum(G) + sum(C)}{num} \tag{3}$$

where n represents the total number of DNA strands, $Ham(X_i, shift(X_j^k))$ is the Hamming distance between X_i and X_j, and m is the total number of bases in each DNA strand. For sliding sequence X_j, $k > 0$ means that X_j slides to the right with $|k|$ step size, and $k < 0$ means that X_j slides to the left with $|k|$ step size. $min_{-m<k<m}^{j=1:n\&\&i\neq j}$ indicates the minimum Hamming distance between X_i and X_j, when X_j slides from the left end to the right end.

5) CONTINUITY

Continuity refers to the continuous appearance of a certain base in a DNA sequence. This constraint is based on the secondary structure. It is necessary to avoid too much continuous same bases, since this kind of DNA strands is unstable, and it is hard to synthesize [16].

6) HAIRPIN STRUCTURE

The Hairpin structure is a typical secondary structure caused by single-stranded DNA molecules folded by themselves. The bases in Hairpin loop and stem cannot combine with other bases, and thus the single DNA strand with Hairpin structure cannot hybridize with other single strand. For this

reason, the Hairpin structure has to be avoided. Hairpin structure can be calculated as follows [32]:

$$Hair = \sum_{i=1}^{m} \sum_{P=P\min}^{n-R\min} \sum_{r=R\min}^{n-2p} T \times \left(\sum_{j=0}^{pinlen(p,r,i)-1} bp(X_{p+i-j}, X_{p+i+j+1}), \frac{pinlen(p,r,i)}{2} \right) \tag{4}$$

$$bp(X_1, X_2) = \begin{cases} 1 & X_1 = X_2^c \\ 0 & otherwise \end{cases} \tag{5}$$

$$pinlen(p,r,i) = min(p+i, n-p-i-r) \tag{6}$$

where X_2^c and X_2 are complementary strands, p is the stem length of the Hairpin, r is the loop length of the Hairpin, R_{min} is the minimum length forming the Hairpin loop, and P_{min} is the minimum length forming the Hairpin stem. As shown in Fig. 2, it is an example of a Hairpin structure.

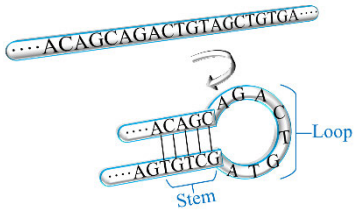

FIGURE 2. Hairpin structure.

7) SPECIAL SEQUENCE

In the DNA biochemical reaction experiments, multiple biological enzymes are usually used to process DNA strands. For instance, ligase can be linked to DNA molecules, and endonucleases can cleave DNA molecules. In this paper, four common endonucleases are selected, and they are AluI, HaeIII, EcoRI, and HindIII [50]–[52]. As showed in Fig. 3, the sequence "AGCT", "GGCC", "GAATTC" and "AAGCTT" can be recognized by AluI, HaeIII, EcoRI, and HindIII, respectively. The cutting method is shown with the red line in Fig. 3. Among them, the incision of AluI and HaeIII is called blunt end, and the incision of EcoRI and HindIII is called sticky end. In general, the sticky end is suitable for the attachment point of the DNA strand displacement reaction, and the blunt end is suitable for the work at the end of the reaction to form a stable structure without exposed bases. Therefore, the cut of DNA strands with endonucleases is also a way to ensure reaction efficiency.

IV. ALGORITHM DESIGN

The DNA sequence coding problem is a multi-constraint multi-objective optimization problem, which is an NP-hard

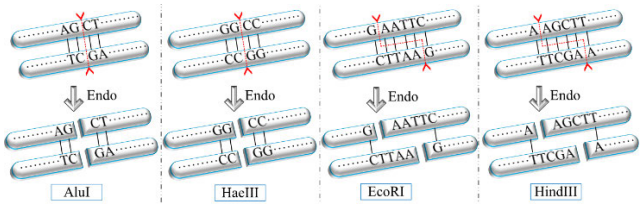

FIGURE 3. The special sequences recognized by four endonucleases.

problem. To this end, in this paper, we consider a heuristic algorithm to solve it. In consideration of the process of chemotaxis and replication characteristics, BFA-A is designed for the DNA sequence coding problem.

BFA-A is divided into four stages, i.e., initialization (Fig. 4 (1)), chemotaxis (Fig. 4 (2)), replication (Fig. 4 (3)), and dispersion (Fig. 4 (4)). Each DNA sequence is defined as a bacterium in the algorithm, and a base change in a DNA sequence is defined as a chemotaxis. For convenience of expression, in the following algorithm description, the DNA sequence is described as bacteria. In Fig. 4 (1), we initiate a population of bacteria, and represent them in format "number – Generation". The "number" is based on the total number of populations. "Generation" is the generation of this bacteria, and it represents whether it is younger or not. In Fig. 4 (2), chemotaxis is a foraging behavior that implements a type of optimization where bacteria try to move to the nutrient substances, to avoid poisonous substances. In Fig. 4 (3), the bacteria in the good environment (i.e., the blue zone shown in Fig. 4 (2)) will be replicated, since this part of bacterial has strong survivability. Once the replication is finished, the "Generation" of parental bacteria increases 1, the "Generation" of new offspring bacteria is set to 1, and the children will inherit the parent's "number". In Fig. 4 (4), a few bacterial in a specific area are eliminated with a small probability because of the dispersion event.

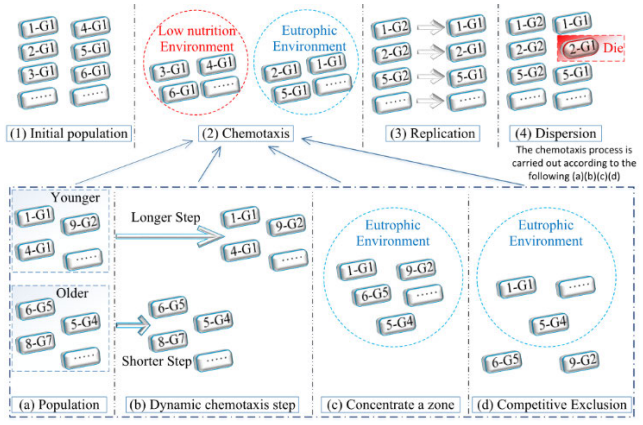

FIGURE 4. The main idea of BFA-A method.

Fig. 4 (a), (b), (c) and (d) give the main idea of chemotaxis, and it is the main improvement of BFA-A. Fig. 4 (a) and (b) are the activity regulation mechanism, (c) and (d) are competitive exclusion mechanism. We consider that the bacteria will

gradually get old with growing, and the aging bacteria mean that their activity will decrease. Therefore, the movement step size of the bacteria in different age periods should be dynamically adjusted. The younger ones have a bigger movement step size (see Fig. 4 (a) and (b)), which can facilitate the algorithm to expand the global search range and avoid early maturity [53]–[55]. While the older bacteria movement step size is smaller, which is beneficial to the convergence of the algorithm. The difference in chemotactic ability of bacteria of different ages is also reflected in the total number of chemotaxis in each chemotaxis cycle, so the total number of chemotaxis times of the older one also decreases with aging. Considering that bacteria concentrate on the eutrophic zone with chemotactic, excessive bacteria quickly consume nutrients and make the eutrophic zone to low-nutrient zone (Fig. 4 (c)). This can cause bacteria unable to survive in the natural area. In order to avoid this phenomenon, we introduce a competitive exclusion mechanism. When bacteria are too concentrated in the same area, the bacteria are forced to separate (Fig. 4 (d)).

The characteristics of BFA-A can be summarized as follows:

- The bacterial activity affects the chemotactic ability of bacteria, and dynamic chemotaxis increases global search ability and improves convergence;
- The bacteria which live in a favorable position will be replicated, so that the population can always move towards a favorable living environment;
- The bacteria are not overly concentrated in a certain area, due to competitive exclusion mechanism. Therefore, the rapid consumption of nutrients in the area can be avoided;
- A few of bacteria will be eliminated during the dispersion events to avoid trapping in local optimum.

A. BFA-A ALGORITHM OVERVIEW

Before giving a detailed description of the algorithm, we first give some relevant definitions and data structures used in the algorithm. In order to obtain a set of sequences with low melting temperature, low similarity, and high stability, we first define some key initial information. This initial information determines the execution of the algorithm in the stages of chemotaxis, replication, and dispersion, and it affects the effects of competitive exclusion mechanisms and activity regulation mechanisms. As shown in **Algorithm 1**, it is the pseudocode of BFA-A. In the first step, the DNA sequence set (DNA_{set}) and the other main parameters (N_{re}, N_c, DNA_{num}, P_{ed}, E_{de}, Lim) are initialized. Here, N_{re} represents the total replication steps, N_c means the total chemotaxis steps in one loop of replication, DNA_{num} is the scale of DNA_{set}, P_{ed} is the probability of dispersion event, E_{de} is the number of DNA sequences that have been eliminated in a dispersion event, and Lim is used to describe the similarity threshold between sequences. Note that N_c will be changed by the age of bacterial due to the activity regulation mechanism.

Algorithm 1 BFA-A

Input: Initial DNA_{set}, N_{re}, N_c, DNA_{num}, P_{ed}, E_{de}, Lim.
Output: Final DNA_{set} with low melting temperature, low similarity and high stability.

for $kr = 1$ **to** N_{re} **do**
  for $i = 1$ **to** DNA_{num} **do**
    for $kc = 1$ **to** N_c **do**
      $X_i = DNA_{set}[i]$;
      $E_c = ActivityBaseNum(X_i.Gen)$;
      $N_c = ActivityChemotaxisNum(X_i.Gen)$;
      $S_0 = Analysis\ (X_i)$;
      $X_i' = Chemotaxis\ (X_i, E_c)$;
      $S_1 = Analysis\ (X_i')$;
      if $S_1 \geq S_0$ **then**
        $X_i = X_i'$;
      end if
      if $X_i.SIMILARITY \geq Lim$ **then**
        $X_i = Competition(X_i)$;
      end if
    end for
    $DNA_{newset}\ [i] = X_i$;
  end for
$DNA_{newset} = Dispersion\ (DNA_{newset}, P_{de}, E_{de})$;
Rank DNA_{newset} **to get** $DNA_{betterhalfset}$;
$DNA_{newbetterhalfset} = DNA_{betterhalfset}$;
$DNA_{newbetterhalfset}.Gen = 1$;
$DNA_{betterhalfset}.Gen++$;
$DNA_{set} = [DNA_{newbetterhalfset}, DNA_{betterhalfset}]$;
end for

The algorithm has three layers of loops, the outermost layer is the number of replication N_{re}, the middle layer is the traversal of all sequences in the set of DNA sequences DNA_{set} (the number of sequences in the DNA_{set} is DNA_{num}), and the innermost layer is the number of chemotaxis N_c. Here, each DNA sequence simulates a bacterium, and the base conversion is bacteria foraging. This algorithm enables all DNA sequences to achieve base conversion by the middle layer loop and the innermost layer loop.

In the innermost layer, a sequence X_i is extracted from DNA_{set} as the first step. An attribute of X_i is generation (abbreviated as $X_i.Gen$), which represents the age of bacteria. Young bacteria have high activity, which means high foraging ability. Bacteria map to DNA sequences, and thus younger DNA sequences have a stronger ability to change bases.

According to the $X_i.Gen$ information, the function $ActivityBaseNum()$ is used to get the number of changed base E_c of sequence X_i during a loop of chemotaxis. Then, the number of chemotaxis N_c of sequence X_i is obtained by function $ActivityChemotaxisNum()$. E_c and N_c depend on $X_i.Gen$, and if $X_i.Gen$ is bigger, E_c and N_c are smaller. After that, function $Analysis()$ (this function is based on the equations in section ***IV.B.1***) is used to calculate the score of X_i, and the score is denoted as S_0. The score of X_i measures

the quality of the sequence. Then, sequence X_i changes its bases by function *Chemotaxis*(), the number of changed base is determined by E_c, and the new sequence is denoted as X_i'. Analogously, we get the score S_1 of X_i' by *Analysis*(). After we get the score, we can determine the difference before and after base changing. If S_1 is bigger than S_0, it means that the bases are indeed changing in the direction to which we want them to move. Therefore, the new sequence X_i' will replace the original sequence X_i. On the contrary, it means that after changing the base, it deviates from the desired target, and the original X_i remains unchanged. Then, we need to determine the similarity between sequences, SIMILARITY is another attribute of X_i. If $X_i.SIMILARITY$ is bigger than *Lim*, it indicates that the density of bacteria is too high. We use the function *Competition*() to realize competitive exclusion process. That is to keep the various base components in X_i unchanged, and rearrange the base positions. After that, the adjusted X_i is restored in the new set DNA_{newset}. At the same time, the chemotactic process of the DNA sequence ends.

Next, the dispersion process for the random E_{de} sequences with the probability of occurrence P_{de} for DNA_{newset} is implemented by *Dispersion*(). Then, we sort the sequences in DNA_{newset} from high to low, and take out the top half of the sequence to form a high-quality sequence set $DNA_{betterhalfset}$. After that, we replicate this set to obtain $DNA_{newbetterhalfset}$. Note that, these two sets are not exactly the same, and their attribute *Gen* is different. $DNA_{newbetterhalfset}.Gen$ is set to 1, while $DNA_{betterhalfset}.Gen$ increases by one for each replication. At last, we can obtain the final DNA_{set} by the union of these two sets.

B. DNA SEQUENCE DESIGN BASED ON BFA-A

In this subsection, five main strategies of the BFA-A method for the DNA sequence design are introduced, namely DNA sequence score strategy, initialization, chemotaxis, replication and dispersion.

1) DNA SEQUENCE SCORE STRATEGY

In order to make a judgment about the direction of chemotaxis and the choice of replication, the quality of each DNA sequence should be quantified for BFA-A. To this end, we propose a score strategy to quantify the quality of a DNA sequence. In order to make a fair comparison with our previous work [21] in the experiments, the same score strategy is used as the one used in BFA [21]. In this paper, *Tm* is considered as the essential constraint, since the low *Tm* is beneficial to the operation of the experiment. In addition, *Sim* and *Hm* are also the essential constraints, since Hamming distance directly affects the probability of the mismatch hybridization. In the evaluation function, we mainly consider six factors, and they are *Con*, *Hair*, *Hm*, *Sim*, *Tm* and *GC*. Based on the results of the simulation experiments, different weights are assigned to these factors, and *Tm*, *Sim*, *Hm*, *Con* account for 40 points, 20 points, 20 points and 20 points, respectively. The total score of each DNA sequence is denoted by S, and S_{Tm}, S_{Sim}, S_{Hm}, S_{Con} are the score of *Tm*, the score of *Sim*, the score of *Hm*, the score of *Con*, respectively. They are calculated as follows:

$$S = \begin{cases} S_{Tm} + S_{Sim} + S_{Hm} + S_{Con} & GC = 0.5 \&\& Hair = 0 \\ 0 & GC \neq 0.5 || Hair \neq 0 \end{cases} \tag{7}$$

$$S_{Tm} = \begin{cases} 40 & Tm \leq 57 \\ 40 - 4 \times (Tm - 57) & 57 < Tm < 67 \\ 0 & Tm \geq 67 \end{cases} \tag{8}$$

$$S_{Sim} = \begin{cases} 20 & Sim' \leq 5 \\ 20 - 2.5 \times (Sim' - 5) & 5 < Sim' < 13 \\ 0 & Sim' \geq 13 \end{cases} \tag{9}$$

where $Sim' = Sim/(NUM_{DNA} - 1)$, NUM_{DNA} is the total number of DNA sequences that generated by initialization.

$$S_{Hm} = \begin{cases} 20 & Hm' \leq 5 \\ 20 - 2.5 \times (Hm' - 5) & 5 < Hm' < 13 \\ 0 & Hm' \geq 13 \end{cases} \tag{10}$$

where $Hm' = Hm/NUM_{DNA}$.

$$S_{Con} = \begin{cases} 20 & Con = 0 \\ 10 & Con = 1 \\ 5 & Con = 2 \\ 0 & Con > 2 \end{cases} \tag{11}$$

According to the above evaluation criterion, for a DNA sequence with lower *Tm*, *Sim*, *Hm*, and *Con*, its score is higher; for a DNA sequence with $GC \neq 0.5$ or $Hair \neq 0$, its score is zero. Therefore, the idea of scoring strategy is keeping *GC* and *Hair* stable, while reducing *Tm*, *Sim*, *Hm*, and *Con* as low as possible.

2) INITIALIZATION

The initialization process mainly includes the initialization of N_{re}, N_c and P_{ed}. N_{re} is the number of the replication, N_c is the maximum number of chemotaxis and P_{ed} is the probability of dispersion. The initial DNA sequences are generated randomly. In addition, the population scale of DNA sequences is also initiated in a random way.

3) CHEMOTAXIS

In the natural environment where the bacteria live, some are eutrophic area and some are not, even poisonous area. Therefore, the bacteria will move from one place to another to find food, and this process is often referred to as chemotaxis [56].

Fig. 5 (1) indicates the initial state of bacteria. In the next moment, bacteria are moving to look for food for survive. Here, the pink square is eutrophic zone and the green square is poisonous zone. In Fig. 5 (2), bacteria start to move for a while, and bacteria gradually move closer to the eutrophic zone and away from the poisonous zone. For the score of a DNA sequence, most DNA sequences are transformed from a

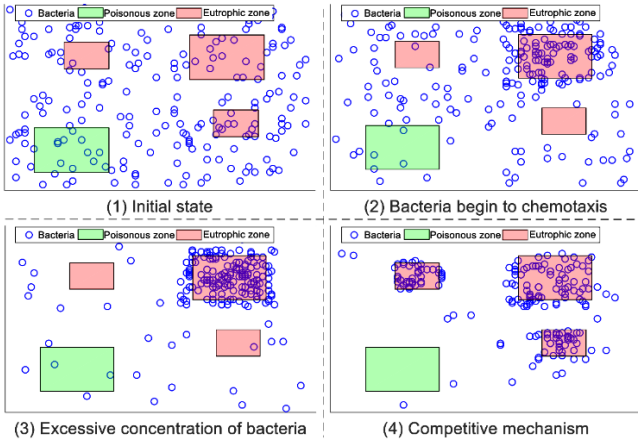

FIGURE 5. Chemotaxis process in BFA-A.

low score to a high score through adjusting the bases of each DNA sequence.

In order to improve the original BFA method, the BFA-A method makes certain adjustments. They are the bacterial activity regulation mechanism and competitive exclusion mechanism. For the BFA method, the characteristics of chemotaxis can be summarized as following:

- The length of chemotaxis step is 1, and it means that only one base will change for each chemotaxis;
- The total number of chemotaxis N_c is fixed;
- Due to chemotaxis and replication, bacteria will gradually concentrate in a certain area, and the diversity of DNA sequences will decrease, which leads to high *Sim* and *Hm*.

In the above process, the chemotaxis step size is fixed, and it limits the diversity of DNA sequences and the global search ability of the algorithm. Note that insufficient DNA sequence diversity will lead to excessive *Sim* and *Hm*. To this end, we try to improve the chemotaxis operation [57]–[59] in BFA-A, and its characteristics are summarized as following:

- Since the generation of each DNA sequence is known, the length of chemotaxis step (E_c) can be adaptively adjusted based on the generation of each DNA sequence according to the related rule given in Table 1. Owing to this adjustment rule, the algorithm has strong global search ability in the early stage, and has better convergence in the later stage. Note that this rule is based on the fact that the maximum chemotaxis time (N_c) is 20.
- N_c is dynamically adjusted based on the generation of each DNA sequence, and the corresponding rule is defined in Table 1. With this mechanism, the convergence can be ensured by reducing the chemotactic ability of bacteria getting old.
- If there are lots of bacteria concentrate to a certain zone, we will force them to separate away. The corresponding rule is that when the *Sim* or *Hm* of two DNA sequences is greater than a threshold (here, threshold is set to 10), the bases in the DNA sequence will be rearranged, while ensuring that the content of "A T G C" is unchanged.

TABLE 1. Dynamic length of chemotaxis step and chemotaxis times.

Gen	E_c	N_c
1-2	5	20
3-4	5	18-20
5-6	4	16-20
7-8	4	14-20
9-10	3	12-20
11-12	2	10-20
Over 12	1	0-20

By this way, *Sim* and *Hm* can be effectively reduced, and thus, the probability of mismatch is reduced.

From Fig. 5, we can find that there are two small pink eutrophic zones. However, they are not as good as the big zone. Small nutrient areas are not sufficiently attractive to bacteria, so the bacteria gather at large nutrient areas. From Fig. 5 (3), we can find that lots of bacteria gather at the big upper eutrophic zone, and the nutrition will be consumed rapidly. As shown in Fig. 5 (4), we introduce the competitive exclusion mechanism to avoid nutrition consumption quickly, and we can see that some individual will be separated to other zones. A large number of bacteria gather around the large nutrient areas, and thus it is more competitive for the large nutrient areas. In addition, some of the less competitive bacteria are squeezed into small nutrient areas. With this mechanism, the load balance can be achieved by the eutrophic zone, and bacteria can survive longer in eutrophic environment. Moreover, the diversity of bacteria that multiply in multiple areas is higher than that of bacteria that are propagated in a specific area. Therefore, the diversity of DNA sequence is also correspondingly improved with more eutrophic zones, and thus *Sim* and *Hm* can be reduced with this mechanism.

4) REPLICATION

After chemotaxis is finished, only DNA sequences with high scores will be replicated, while the ones with low scores will be not. Note that the total population is unchanged for each replication operation. Replication has the characteristics of retaining a better half of the DNA sequences and replacing the poorer half of the DNA sequences. DNA sequences with high scores can be retained to the next generation. This evolutionary process of the bacterial foraging algorithm advantageously guarantees that the base components of the DNA sequences with high scores are retained, and low scores sequences are eliminated. So it is very useful for the optimization of DNA coding design.

5) DISPERSION

Dispersion event will occur with a specific probability after the completion of each replication. The random elimination can facilitate avoiding local optimizations; however, the optimal solution may also be eliminated. In order to avoid dispelling the optimal solution, the dispersion event is only carried out for the DNA sequences with middle scores.

V. SIMULATION RESULT

In order to show that the algorithm in this paper has the ability to optimize DNA coding design, we designed 2 simulation experiments for the BFA-A method to evaluate the melting temperature and similarity of the DNA sequence. In the first part of this section, we analyze the performance of BFA-A method under 6 traditional constraints (*Sim*, *Hm*, *Tm*, *Con*, *GC*, *Hair*), and we compare BFA-A method with 5 other methods of DNA sequence coding. In the second part, we analyze the performance of BFA-A under the constraint of the special sequence which can be recognized by four common kinds of endonuclease.

A. THE PERFORMANCE OF BFA-A

We evaluate the performance of BFA-A against NCIWO [16], NACST [32], IWO [31], IGA [30] and BFA [21]. The setting of main parameters used in our experiment is given in Table 2. N_{re} represents the replication steps of each bacteria and it is set to 20. N_c is the maximum chemotaxis steps of each bacteria and it is set to 20 as the initial value. Note that N_c changes with the ''age'' of the bacteria. P_{ed} is the probability of dispersion and it is set to 0.1.

TABLE 2. Main parameters.

Parameter	Mean	Value
N_{re}	Replication steps	20
N_c	Chemotaxis steps	20
P_{ed}	Probability of dispersion	0.1

As shown in Table 3, there are 7 sequences with 20 bases for BFA-A, BFA, NCIWO, NACST, IWO and IGA in the experiment. Based on the definition of each constraint given in Section ***III***, we consider 6 constraints (*Sim*, *Hm*, *Tm*, *Con*, *GC*, *Hair*) of all DNA sequences generated by different algorithms. *Sim* (*Hm*) and *Tm* are the main optimization targets in this paper, so we first analyze their values. After that, other values such as *Con*, *GC*, and *Hair* that affects structural stability and secondary structural properties are analyzed. The comparison of *Sim* and *Hm* is illustrated in Fig. 6.

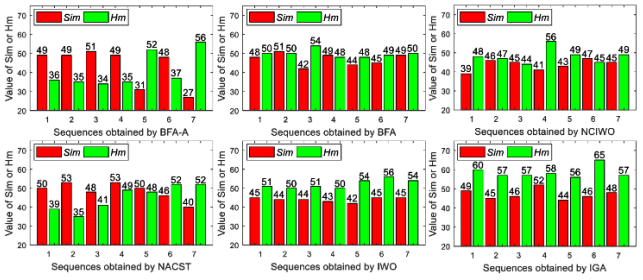

FIGURE 6. *Sim* and *Hm* comparison in different methods.

In Fig. 6, two adjacent bars in the histogram represent *Sim* and *Hm* of the same sequence. *Sim* and *Hm* together reflect the similarity of DNA sequences. We can clearly see that when one of *Sim* and *Hm* is larger, the other is smaller, which seems to indicate that *Sim* and *Hm* are complementary. To verify this relationship, we calculate the sum of *Sim* and *Hm* for each DNA sequence generated by different algorithms, and the result is illustrated in Fig. 7.

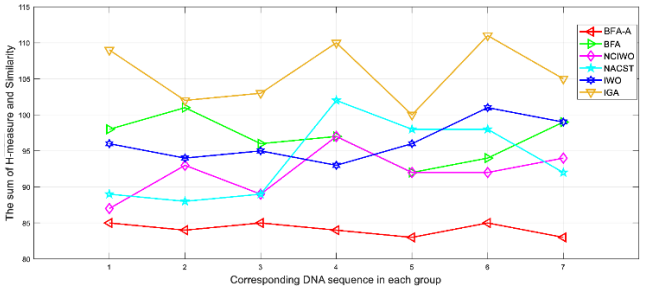

FIGURE 7. The sum of *Sim* and *Hm* for different methods.

We can clearly find that the sum of *Sim* and *Hm* for each DNA sequence is approximately equal to a constant value. Now, we discuss this complementary relationship through a simple mathematical reasoning.

We assume that $Sim + Hm$ approximately equals to a constant value (expressed as C), i.e.,

$$Sim + Hm \approx C \tag{12}$$

According to section III. B, we can get the specific mathematical expressions of *Sim* and *Hm*, and they are,

$$Sim = \sum_{j=1}^{n} min_{-m<k<m}^{j=1:n\&\&i\neq j}[Ham(X_i, shift(X_j^k))] \tag{13}$$

$$Hm = \sum_{j=1}^{n} min_{-m<k<m}^{j=1:n}[Ham(X_i, shift(X_j^{ck}))] \tag{14}$$

According to the known formulas and existing conditions, it can be clearly seen that, when $i \neq j$, $Sim + Hm \approx C$ can be transformed into,

$$\sum_{j=1}^{n} min_{-m<k<m}^{j=1:n}[Ham(X_i, shift(X_j^k)) + Ham(X_i, shift(X_j^{ck}))] \approx C \tag{15}$$

In (15), it is obvious that C will change with n. Now, we assume $n = 2$, since 2 DNA strands are required at least, and the formula (15) transforms into, (when $n = 2$, the result is C')

$$min_{-m<k<m}^{j=1:2}[Ham(X_i, shift(X_j^k)) + Ham(X_i, shift(X_j^{ck}))] \approx C' \tag{16}$$

where k is the sliding distance of DNA strand X_j. We assume that k and j are fixed values, then the formula can be transformed into,

$$[Ham(X_i, shift(X_j)) + Ham(X_i, shift(X_j^c))] \approx C' \tag{17}$$

For any DNA sequence in ideal condition, bases 'A', 'T', 'G', 'C' should account for 25%, respectively. For instance,

TABLE 3. DNA sequences generated by different algorithms.

Algorithm	Sequence from 5' to 3'	*Sim*	*Hm*	*Tm*	*GC*	*Con*	*Hair*
BFA-A	CTACTTCCTACCTACCTACC	49	36	59.54	50%	0	0
	CCTATCTCACTACCTCCTAC	49	35	59.62	50%	0	0
	CTCTCTCTCTCCTCTAACTC	51	34	59.55	50%	0	0
	CTCTCTCTTCTCTCTCTCAC	49	35	60.14	50%	0	0
	GAGTAGTAGGAGGAGAAGAG	31	52	59.55	50%	0	0
	CTACCTTCTCCTAACTCCTC	48	37	60.14	50%	0	0
	GAAGAGGAGGAAGAGTAGAG	27	56	60.14	50%	0	0
	Group result	**304**	**285**			0	0
BFA	GTGTAGTACTCGTCTAGCAC	48	50	61.42	50%	0	0
	CTGTCTGTACTCAGCGATAG	51	50	61.50	50%	0	0
	CTGATCTACGTATGCACGTC	42	54	62.46	50%	0	0
	CGTCTCTCTATCTAGCGTGA	49	48	62.44	50%	0	0
	TCATACTCTCTCGTGCTGAG	44	48	62.96	50%	0	0
	GCGTGTAGCGACTACTGATA	45	49	63.63	50%	0	0
	GTAGCGACTGTGTCATACAC	49	50	62.62	50%	0	0
	Group result	328	349			0	0
NCIWO	ACACCAGCACACCAGAAACA	39	48	66.99	50%	9	0
	GTTCAATCGCCTCTCGGTAT	46	47	64.26	50%	0	0
	GCTACCTCTTCCACCATTCT	45	44	63.55	50%	0	0
	GAATCAATGGCGGTCAGAAG	41	56	63.58	50%	0	0
	TTGGTCCGGTTATTCCTTCG	43	49	64.44	50%	0	0
	CCATCTTCCGTACTTCACTG	47	45	62.30	50%	0	0
	TTCGACTCGGTTCCTTGCTA	45	49	65.61	50%	0	0
	Group result	306	338			9	0
NACST	CTCTTCATCCACCTCTTCTC	50	39	61.38	50%	0	0
	CTCTCATCTCTCCGTTCTTC	53	35	61.44	50%	0	0
	TATCCTGTGGTGTCCTTCCT	48	41	64.46	50%	0	0
	ATTCTGTTCCGTTGCGTGTC	53	49	65.83	50%	0	0
	TCTCTTACGTTGGTTGGCTG	50	48	64.63	50%	0	0
	GTATTCCAAGCGTCCGTGTT	46	52	65.30	50%	0	0
	AAACCTCCACCAACACACCA	40	52	66.71	50%	9	0
	Group result	340	316			9	0
IWO	GATGGATTTACCTTGCACCT	45	51	62.29	45%	9	4
	CCTTCTCTCGTCTTCATACA	44	50	60.72	45%	0	0
	ACGATCGATTAATGGGAGTC	44	51	61.52	45%	9	3
	ATAAGTAGGGACTGCTCTAC	43	50	59.84	45%	9	0
	CCTAAGAACACAGGGCATAG	42	54	62.04	50%	9	4
	CTGGAAGCGTTTGCTAACTT	45	56	63.38	45%	9	6
	GCAGATTCCCGGATACTCAG	45	54	64.34	55%	9	7
	Group result	308	366			54	24
IGA	AGAGTACGTCAGATGACTGC	49	60	63.53	50%	0	0
	TGCTGTAGATCGTCGCATCA	45	57	66.05	50%	0	0
	CTACTACGAGTCACACACAG	46	57	61.63	50%	0	0
	GTGAGAGCTCAGTCGATGAT	52	58	63.61	50%	0	0
	TACGTCTCTGTCTGCTTTGC	44	56	64.65	50%	9	0
	ACACACACTCACTAGTGACG	46	65	63.95	50%	0	0
	CATACGTGAGTGCTGATACG	48	57	62.74	50%	0	0
	Group result	330	410			9	0

if the 4^{th} base of DNA sequence X_i is 'A', which is expressed as X_i [4] $='$ A', then in DNA sequence X_j (or X_j^c), the probability of X_j [4] $='$ A$'$ (or X_j^c [4] $='$ A$'$) should also be 25%. It means that for any DNA sequence set in ideal conditions, the sum of *Sim* and *Hm* for any two sequences in this set should be equal to 25% of the length of DNA sequence.

Now, we assume that k is not a fixed value, and this means that X_j (or X_j^c) can shift to left or right. According to the above analysis, the probability of X_j^k [4] $='$ A$'$ (or X_j^{ck} [4] $='$ A$'$) is still 25%. *Sim*+*Hm* is still equal to a constant value.

Then, we assume that j is not a fixed value and $n > 2$, and it means that DNA strand X_j (or X_j^c) can be changed. The above inference for the sum of *Sim* and *Hm* for each DNA sequence is approximately equal to a constant value is still true obviously. For more DNA sequences, even if some sequences are not ideal, the overall result is still ideal when n is large enough. Therefore, we can get $Sim + Hm \approx C$.

Note that, the actual value of C has a substantial relationship with n, and that inference is under the ideal condition (uniform distribution of four kinds of base). For DNA computing, we need to consider some constraints, *Tm*, *Hair* and so on, and these constraints will limit the base composition. Therefore, *Sim* or *Hm* will be larger than 25%. However, in Fig. 7, *Sim* and *Hm* are approximately equal to a constant value, and this means that even under these constraints, the inference is still correct.

In addition, from Fig. 7, we can find that the sum of *Sim* and *Hm* for BFA-A is smaller than the others. It demonstrates that the DNA sequences obtained by BFA-A have lower *Sim*

and *Hm*, which means that these DNA sequences have lower mismatch probability. We also can find that BFA, NCIWO, NACST and IWO perform similarly on *Sim* and *Hm*, and IGA is inferior.

Considering that *Tm* affects the efficiency and difficulty of experiment, as shown in Fig. 8. The DNA sequences obtained by BFA-A have the smallest maximum, minimum and average values of *Tm*. IWO is slightly worse than BFA-A, but better than other methods. We also can find that NCIWO, NACST, BFA and IGA perform worse on these 3 values of *Tm*.

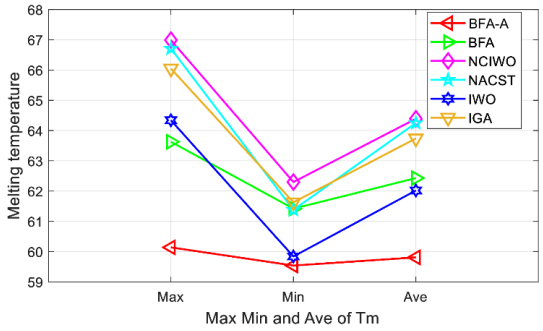

FIGURE 8. Max, Min and Ave of *Tm* comparison in different methods.

In order to comprehensively analyze *Tm* of DNA sequences obtained by different algorithms, we report the variance of *Tm* in Fig. 9. We can find that BFA-A outperforms all existing algorithms in terms of the *Tm* variance, and it means that BFA-A can generate DNA sequences with relatively lower and closer melting temperature. From the base composition given in Table 3, we can find that the reason of the best performance of *Tm* is that different adjacent bases have different stability. It can be clearly seen that there are lots of "T·C", "A·C" and "C·C" neighbor combinations in the DNA sequences generated by BFA-A. According to [49], the thermodynamic stability of adjacent base pairs conforms to inequality (18),

$$G \cdot C > A \cdot T > G \cdot G > G \cdot T \geq G \cdot A > T \cdot T \geq A \cdot A > T \cdot C \geq A \cdot C \geq C \cdot C \quad (18)$$

where the "T·C", "A·C" and "C·C" neighbor combination has a lower thermodynamic stability in (18), which means

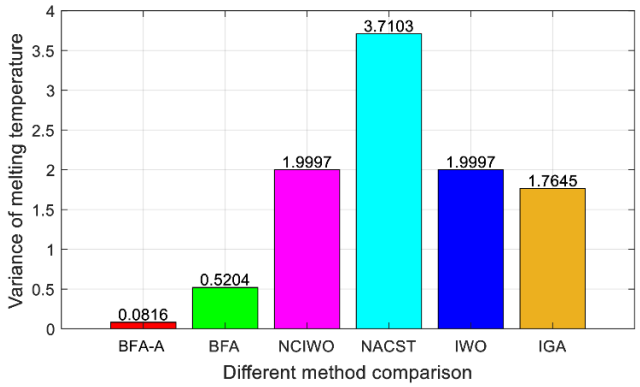

FIGURE 9. Variance of *Tm* for different methods.

that the DNA strand has a lower *Tm*. This explains that BFA-A has a good performance in terms of *Tm*.

The constraints *GC*, *Hair* and *Con* affect the stability of the DNA strand, the ability to fold itself, and the difficulty of synthesis, respectively. We compare these 6 methods in Table 4. For all methods except IWO, *GC* is set to 50%, and for IWO, *GC* is larger than 45% and less than 55%. BFA-A and BFA strictly control *Con* and *Hair* which are always 0. Note that, for NCIWO, NACST and IGA, their *Con* is worse than BFA-A and BFA, although they are as good as BFA-A and BFA in terms of *Hair*.

TABLE 4. *GC*, *Hair* and *Con* comparison in different methods.

Method	*GC*	*Hair*	*Con*
BFA-A	50%	0	0
BFA	50%	0	0
NCIWO	50%	0	9
NACST	50%	0	9
IWO	45%-55%	24	54
IGA	50%	0	9

For BFA, due to the replication progresses, bacteria will gradually gather around a certain zone. It means that the diversity of DNA sequences is reduced, and thus the *Sim* and *Hm* of DNA sequence will increase. Instead, in BFA-A, we fix out the excessive concentration problem by introducing the competitive exclusion mechanism. Therefore, *Sim* and *Hm* of DNA sequences obtained by BFA-A are greatly decreased compared with BFA. In addition, since the activity regulation mechanism, the dynamic chemotaxis step length makes the BFA-A method have greater global search ability and convergence ability. In the initial stage of the algorithm, the obtained DNA sequences are more diverse. In the later stages of the algorithm, better DNA sequences can be generated after the algorithm convergence.

The simulation results are summarized as follows:

- BFA-A is slightly better than the other algorithms on *Sim* and *Hm*;
- Compared with other algorithms, BFA-A has an obvious good performance in the *Tm*.
- In terms of *Con*, *Hair* and *GC*, BFA-A is as good as BFA.

To further illustrate the experimental result performance of the BFA-A method, we continue to make several groups of the experiment with larger bacterial scale as showed in Table 5.

In Table 5, *Gen* is the average generation of bacteria (DNA strands) in this group, and *Sim*, *Hm*, *Con*, *GC* and *Hair* are the average values of different constraints index in this group. We can find that the average generation is below 5, and it means that the bacteria are young. Young bacteria have high life activity and strong chemotaxis ability, and they are more likely to find a nutritious environment. The experimental results show that the method can reduce the number of iterations and has guided significance for improving the efficiency of DNA sequence acquisition.

TABLE 5. Another groups of experiment.

No.	*Gen*	*Sim*	*Hm*	*Con*	*GC*	*Hair*	*Tm*(ave/var/max/min)
1	3.87	7.92	7.12	0	50%	0	59.70/0.1793/60.73/58.96
2	3.10	8.40	6.96	0	50%	0	59.71/0.1798/60.51/58.72
3	4.17	7.98	7.03	0	50%	0	59.68/0.1420/60.21/58.66
4	3.77	8.34	7.00	0	50%	0	59.64/0.1550/60.29/58.72
5	4.78	7.88	7.14	0	50%	0	59.51/0.1916/60.73/58.66
6	3.99	8.10	7.12	0	50%	0	59.47/0.1607/60.21/58.73
7	4.47	8.00	7.24	0	50%	0	59.65/0.1797/60.73/58.65
8	4.24	7.97	7.11	0	50%	0	59.57/0.2415/60.73/58.65
9	4.20	8.14	7.16	0	50%	0	59.49/0.1978/60.73/58.65
10	4.47	7.96	6.96	0	50%	0	59.41/0.1727/60.21/58.65
11	3.38	8.29	7.90	0	50%	0	59.49/0.1565/60.72/58.72
12	4.55	7.86	6.89	0	50%	0	59.43/0.1874/60.73/58.66
13	4.10	7.69	6.93	0	50%	0	59.47/0.2468/61.03/58.65

From Table 5, we can see *Con* and *Hair* always equal to 0, and *GC* is 50%. In addition, we can find that the average and variance of *Tm* are low. From the results, we can confirm that DNA sequences generated by BFA-A have good performance in terms of *Tm*, *Sim* and *Hm*.

B. THE PERFORMANCE OF BFA-A UNDER ENDONUC-LEASE RECOGNITION SEQUENCE CONSTRAINT

Biochemical reactions require a variety of enzymes, and for instance, endonuclease and ligase are two commonly used enzymes. DNA endonuclease can specifically recognize a piece of sequence and cut it off to form a blunt end or a sticky end in the incision. DNA ligase will reconnect the incisions again. Different endonucleases can recognize different DNA sequences. Therefore, we should consider retaining some specific sequence pieces for the design of DNA sequences. In this section, we consider 4 kinds of common DNA endonuclease, AluI, HaeIII, EcoRI, and HindIII. As shown in Fig. 3, they can recognize special sequence "AGCT", "GGCC", "GAATTC" and "AAGCTT", respectively. Under this constraint, the experiments are performed with the parameters given in Table 2 to verify the effect of BFA-A on decreasing the melting temperature and the similarity.

Table 6 shows the DNA sequence sets under 4 kinds of endonuclease constraints, which are generated by BFA-A. In Table 6, identifiable special fragments are bolded. In the previous section, we obtained the inference that $Sim + Hm$ is approximately equal to a constant. The experiments in this section can further confirm the correctness of this inference. The result for *Sim* and *Hm* is given in Table 6, and the comparison result is shown in Fig. 10. When *Sim* is high, *Hm* will be low. This result is similar to Fig. 6.

To further illustrate the correctness of the inference, as shown in Fig. 11, we report the sum of *Sim* and *Hm* in each group. We can find that the sum of *Sim* and *Hm* is approximately equal to a constant. This phenomenon is similar to the result discussed in section ***V.A***. However, the performance is not as good as the situation of absence of specific sequence restriction. Since our inference is based on the ideal situation, but specific sequence constraint is too strong to break the ideal situation, it directly leads to the deterioration of the phenomenon. In addition, we can find that the DNA sequences with a long base length (such as EcoRI and HindIII) are worse for *Sim* and *Hm* than the sequence with a shorter base length (such as AluI and HaeIII).

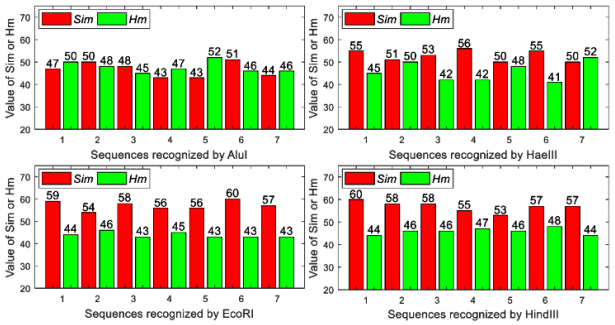

FIGURE 10. ***Sim*** **and** ***Hm*** **comparison for BFA-A under different endonuclease constraint.**

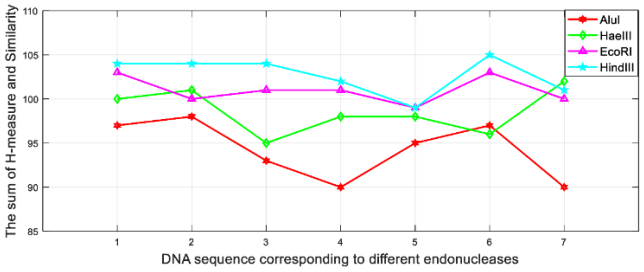

FIGURE 11. The value of Sim+Hm comparison for BFA-A under different endonuclease constraint.

Considering that *Tm* affects the efficiency and difficulty of experiment, we report the average and variance of *Tm* under a specific sequence constraint, as shown in Fig. 12. It can be seen that BFA-A performs well at *Tm*, even under specific sequence constraint.

To summarize this subsection, under the constraint of specific base sequences, BFA-A performs good for *Tm*. However, BFA-A performs poorly for *Sim* and *Hm* in absence of specific base sequences constraint. Considering that there is

TABLE 6. DNA sequences generated by BFA-A under the constraint of 4 kinds of endonuclease.

Endonuclease	Sequence from 5' to 3'	*Sim*	*Hm*	*Tm*	*GC*	*Con*	*Hair*
AluI	CTATCACC**AGCT**CTACTAGG	47	50	60.29	50%	0	0
	GACTAGTC**AGCT**ACTCCTAG	50	48	59.92	50%	0	0
	GTAGTAGG**AGCT**GGTAAGAG	48	45	60.51	50%	0	0
	CGTAGGTA**AGCT**GAGTAGAG	43	47	60.86	50%	0	0
	CCTCTCTG**AGCT**ATAGTCTC	43	52	59.99	50%	0	0
	GTAGGTAC**AGCT**GTAGGTAG	51	46	60.50	50%	0	0
	CACTACCT**AGCT**TCTACCTC	44	46	60.51	50%	0	0
	Group result	326	334			0	0
HaeIII	GTACTAGT**GGCC**ACTACTAC	55	45	60.79	50%	0	0
	CTTACTAA**GGCC**TTACCTCC	51	50	61.10	50%	0	0
	GAGTGTAA**GGCC**TACTAGAG	53	42	60.51	50%	0	0
	AGTAGAGA**GGCC**TAGACTAG	56	42	60.81	50%	0	0
	GATAAGTA**GGCC**ACCTAAGG	50	48	61.17	50%	0	0
	GTAGTAGA**GGCC**ATAGACAG	55	41	60.58	50%	0	0
	CTAGCCTA**GGCC**TATAGATC	50	52	59.84	50%	0	0
	Group result	370	320			0	0
EcoRI	CCAGACC**GAATTC**TACCTAC	59	44	61.45	50%	0	0
	GGTAGTA**GAATTC**TAGGCGG	54	46	61.23	50%	0	0
	GGAGAGA**GAATTC**CTGTGAG	58	43	61.38	50%	0	0
	CGTCCAG**GAATTC**TCTACTC	56	45	61.44	50%	0	0
	GGTAGTA**GAATTC**TAGGCGG	56	43	61.23	50%	0	0
	GTAGACG**GAATTC**GTACCTC	60	43	61.78	50%	0	0
	GAGCGGA**GAATTC**ACTCTAC	57	43	62.08	50%	0	0
	Group result	400	307			0	0
HindIII	GACGAGG**AAGCTT**ACTAGAG	60	44	61.15	50%	0	0
	CTACAGT**AAGCTT**GGCTAGG	58	46	61.74	50%	0	0
	CTAGGAG**AAGCTT**CTGGTAG	58	46	60.80	50%	0	0
	CCTAGAG**AAGCTT**GAGAGTG	55	47	61.38	50%	0	0
	CTACAGT**AAGCTT**GGCTAGG	53	46	61.74	50%	0	0
	GAGACGT**AAGCTT**CTACTCC	57	48	61.43	50%	0	0
	GGTAGTG**AAGCTT**CCTTACC	57	44	61.96	50%	0	0
	Group result	398	321			0	0

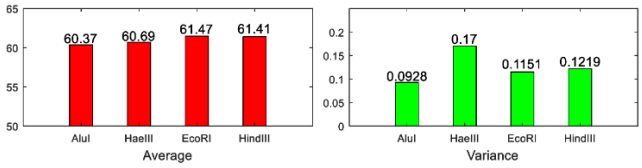

FIGURE 12. *Tm* performance for BFA-A under the endonuclease constraint.

a certain correlation between *Sim* and *Hm*, we can consider only one constraint to reduce the complexity of the algorithm.

VI. CONCLUSION

In this paper, we proposed a DNA sequence coding method named BFA-A to generate DNA sequences with low mismatch probability and stable melting temperature, which solves the problem that DNA sequences are prone to mismatch and experimental temperature instability in DNA computing. In order to obtain DNA sequences with high quality, we introduced a variety of constraints and DNA sequence scoring strategies. By introducing bacterial activity regulation mechanism and competitive exclusion mechanism, the global search ability and convergence of the algorithm are significantly improved, and thus a better DNA sequences are obtained. A large number of simulation experiments show that the DNA sequences obtained by BFA-A have improved performance on *Sim*, *Hm* and *Tm*. As part of our future work, we will pay close attention to the following issues:

- We will consider more constraints in DNA sequence design to adapt to more practical problems;
- We will try more new methods to enrich the technical routes in this field;
- We will explore the application scenarios of the BFA-A method to enable it to be more commonly used in a variety of problems.

REFERENCES

[1] L. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, no. 5187, pp. 1021–1024, Nov. 1994, doi: 10.1126/science.7973651.

[2] L. Qian, E. Winfree, and J. Bruck, "Neural network computation with DNA strand displacement cascades," *Nature*, vol. 475, no. 7356, pp. 368–372, Jul. 2011, doi: 10.1038/nature10262.

[3] K. M. Cherry and L. Qian, "Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks," *Nature*, vol. 559, no. 7714, pp. 370–376, Jul. 2018, doi: 10.1038/s41586-018-0289-6.

[4] Y.-J. Chen, N. Dalchau, N. Srinivas, A. Phillips, L. Cardelli, D. Soloveichik, and G. Seelig, "Programmable chemical controllers made from DNA," *Nature Nanotechnol.*, vol. 8, no. 10, pp. 755–762, Sep. 2013, doi: 10.1038/nnano.2013.189.

[5] X. Liu, R. Aizen, R. Freeman, O. Yehezkeli, and I. Willner, "Multiplexed aptasensors and amplified DNA sensors using functionalized graphene oxide: Application for logic gate operations," *ACS Nano*, vol. 6, no. 4, pp. 3553–3563, Mar. 2012, doi: 10.1021/nn300598q.

[6] Z. Hu, J. Jian, Y. Hua, D. Yang, Y. Gao, J. You, Z. Wang, Y. Chang, K. Yuan, Z. Bao, Q. Zhang, S. Li, Z. Jiang, and H. Zhou, "DNA colorimetric logic gate in microfluidic chip based on unmodified gold nanoparticles and molecular recognition," *Sens. Actuators B, Chem.*, vol. 273, pp. 559–565, Nov. 2018, doi: 10.1016/j.snb.2018.06.073.

[7] R. Peng, X. Zheng, Y. Lyu, L. Xu, X. Zhang, G. Ke, Q. Liu, C. You, S. Huan, and W. Tan, "Engineering a 3D DNA-logic gate nanomachine for bispecific recognition and computing on target cell surfaces," *J. Amer. Chem. Soc.*, vol. 140, no. 31, pp. 9793–9796, Jul. 2018, doi: 10.1021/jacs.8b04319.

[8] Y. Amir, E. Ben-Ishay, D. Levner, S. Ittah, A. Abu-Horowitz, and I. Bachelet, "Universal computing by DNA origami robots in a living animal," *Nature Nanotechnol.*, vol. 9, no. 5, pp. 353–357, Apr. 2014, doi: 10.1038/NNANO.2014.58.

[9] A. J. Thubagere, W. Li, R. F. Johnson, Z. Chen, S. Doroudi, Y. L. Lee, G. Izatt, S. Wittman, N. Srinivas, D. Woods, E. Winfree, and L. Qian, "A cargo-sorting DNA robot," *Science*, vol. 357, no. 6356, Sep. 2017, Art. no. eaan6558, doi: 10.1126/science.aan6558.

[10] X. Liu, F. Zhang, X. Jing, M. Pan, P. Liu, W. Li, B. Zhu, J. Li, H. Chen, L. Wang, J. Lin, Y. Liu, D. Zhao, H. Yan, and C. Fan, "Complex silica composite nanomaterials templated with DNA origami," *Nature*, vol. 559, no. 7715, pp. 593–598, Jul. 2018, doi: 10.1038/s41586-018-0332-7.

[11] M. Babaei, "A novel text and image encryption method based on chaos theory and DNA computing," *Natural Comput.*, vol. 12, no. 1, pp. 101–107, Aug. 2012, doi: 10.1007/s11047-012-9334-9.

[12] Q. Zhang, L. Guo, and X. Wei, "Image encryption using DNA addition combining with chaotic maps," *Math. Comput. Model.*, vol. 52, nos. 11–12, pp. 2028–2035, Dec. 2010, doi: 10.1016/j.mcm.2010.06.005.

[13] Q. Zhang and X. Wei, "A novel couple images encryption algorithm based on DNA subsequence operation and chaotic system," *Optik*, vol. 124, no. 23, pp. 6276–6281, Dec. 2013, doi: 10.1016/j.ijleo.2013.05.009.

[14] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, pp. 77–80, Jan. 2013, doi: 10.1038/nature11875.

[15] S. Hoshika, N. A. Leal, M.-J. Kim, M.-S. Kim, N. B. Karalkar, H.-J. Kim, A. M. Bates, N. E. Watkins, H. A. SantaLucia, A. J. Meyer, S. DasGupta, J. A. Piccirilli, A. D. Ellington, J. SantaLucia, M. M. Georgiadis, and S. A. Benner, "Hachimoji DNA and RNA: A genetic system with eight building blocks," *Science*, vol. 363, no. 6429, pp. 884–887, Feb. 2019, doi: 10.1126/science.aat0971.

[16] G. Yang, B. Wang, X. Zheng, C. Zhou, and Q. Zhang, "Iwo algorithm based on niche crowding for DNA sequence design," *Interdiscip. Sci., Comput. Life Sci.*, vol. 9, pp. 342–349, Mar. 2016, doi: 10.1007/s12539-016-0160-0.

[17] A. Panjkovich and F. Melo, "Comparison of different melting temperature calculation methods for short DNA sequences," *Bioinformatics*, vol. 21, no. 6, pp. 711–722, Mar. 2005, doi: 10.1093/bioinformatics/bti066.

[18] E. de Klerk and D. V. Pasechnik, "Approximation of the stability number of a graph via copositive programming," *SIAM J. Optim.*, vol. 12, no. 4, pp. 875–892, Jan. 2002, doi: 10.1137/S1052623401383248.

[19] J. Rose, R. Deaton, M. Garzon, R. C. Murphy, D. Franceschetti, and S. E. Stevens, "The effect of uniform melting temperatures on the efficiency of DNA computing," in *Proc. DIMACS*, Philadelphia, PA, USA, 1997, pp. 35–42.

[20] F. Tanaka, A. Kameda, M. Yamamoto, and A. Ohuchi, "Design of nucleic acid sequences for DNA computing based on a thermodynamic approach," *Nucleic Acids Res.*, vol. 33, no. 3, pp. 903–911, Feb. 2005, doi: 10.1093/nar/gki235.

[21] J. Ren and Y. Yao, "DNA computing sequence design based on bacterial foraging algorithm," in *Proc. 5th Int. Conf. Bioinf. Res. Appl.*, Hong Kong, Dec. 2018, pp. 1–7.

[22] A. J. Hartemink, D. K. Gifford, and J. Khodor, "Automated constraint-based nucleotide sequence selection for DNA computation," *Biosystems*, vol. 52, nos. 1–3, pp. 227–235, Oct. 1999, doi: 10.1016/s0303-2647(99)00050-7.

[23] A. G. Frutos, A. J. Thiel. A. E. Condon, L. M. Smith, and R. M. Corn, "DNA computing at surfaces: Four base mismatch word design," in *Proc. DIMACS*, Philadelphia, PA, USA, 1997, pp. 238–239.

[24] M. Arita and S. Kobayashi, "DNA sequence design using templates," *New Gener. Comput.*, vol. 20, no. 3, pp. 263–277, Sep. 2002, doi: 10.1007/BF03037360.

[25] W. Liu, S. Wang, L. Gao, F. Zhang, and J. Xu, "DNA sequence design based on template strategy," *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 2014–2018, Nov. 2003, doi: 10.1021/ci025645s.

[26] U. Feldkamp, S. Saghafi, W. Banzhaf, and H. Rauhe, "DNA sequence generator: A program for the construction of DNA sequences," in *Proc. 7th Int. Workshop DNA-Based Comput., DNA Comput.*, Tampa, FL, USA, 2001, pp. 23–32.

[27] A. Marathe, A. E. Condon, and R. M. Corn, "On combinatorial DNA word design," *J. Comput. Biol.*, vol. 8, no. 3, pp. 201–220, Feb. 2001, doi: 10.1089/10665270152530818.

[28] F. Tanaka and M. Nakatsugawa, "Developing support system for sequence design in DNA computing," in *Proc. 7th Int. Workshop DNA-Based Comput., DNA Comput.*, Tampa, FL, USA, 2001, pp. 129–137.

[29] Q. Zhang, B. Wang, X. Wei, X. Fang, and C. Zhou, "DNA word set design based on minimum free energy," *IEEE Trans. Nanobiosci.*, vol. 9, no. 4, pp. 273–277, Dec. 2010, doi: 10.1109/TNB.2010.2069570.

[30] Q. Zhang and X. Xia, "The quality optimization of DNA sequences with improved genetic algorithm," *J. Comput. Theor. Nanosci.*, vol. 10, no. 5, pp. 1192–1195, May 2013, doi: 10.1166/jctn.2013.2827.

[31] D. F. Luo and D. J. Luo, "The research of DNA coding sequences based on invasive weed optimization," *Sci. Tech Eng.*, vol. 13, no. 13, pp. 3545–3551, Jul. 2013, doi: 10.3969/j.issn.1671-1815.2013.13.005.

[32] S.-Y. Shin, I.-H. Lee, D. Kim, and B.-T. Zhang, "Multiobjective evolutionary optimization of DNA sequences for reliable DNA computing," *IEEE Trans. Evol. Comput.*, vol. 9, no. 2, pp. 143–158, Apr. 2005, doi: 10.1109/TEVC.2005.844166.

[33] S. Xu and Q. Zhang, "Optimization of DNA coding based on GA/PSO algorithm," *Comput. Eng.*, vol. 34, no. 1, pp. 218–220, Jan. 2008, doi: 10.3969/j.issn.1000-3428.2008.01.075.

[34] C. Xu, Q. Zhang, B. Wang, and R. Zhang, "Research on the DNA sequence design based on GA/PSO algorithms," in *Proc. 2nd Int. Conf. Bioinf. Biomed. Eng.*, Shanghai, People's Republic of China, May 2008, pp. 816–819.

[35] Q. Guo, B. Wang, C. Zhou, X. Wei, and Q. Zhang, "DNA code design based on the bloch quantum chaos algorithm," *IEEE Access*, vol. 5, pp. 22453–22461, Oct. 2017, doi: 10.1109/ACCESS.2017.2760882.

[36] K. Liu, B. Wang, H. Lv, X. Wei, and Q. Zhang, "A BPSON algorithm applied to DNA codes design," *IEEE Access*, vol. 7, pp. 88811–88821, Jun. 2019, doi: 10.1109/ACCESS.2019.2924708.

[37] Z. Yin and C. Ye, "Research on DNA encoding based on cultural particle swarm optimization algorithm," *Comput. Eng.*, vol. 37, no. 3, pp. 10–12, Mar. 2011, doi: 10.3969/j.issn.1000-3428.2011.03.004.

[38] Z. Yin and C. Ye, "Cultural evolution based particle swarm optimization algorithm for DNA sequence design," *Comput. Eng. Appl.*, vol. 47, no. 1, pp. 40–42, Jan. 2011, doi: 10.3778/j.issn.1002-8331.2011.01.011.

[39] J.-H. Xiao, Y. Jiang, J.-J. He, and Z. Cheng, "A dynamic membrane evolutionary algorithm for solving dna sequences design with minimum free energy," *Match Commun. Math. Comput. Chem.*, vol. 70, no. 3, pp. 987–1004, 2013.

[40] A. C. H. Choong and N. K. Lee, "Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method," in *Proc. Int. Conf. Comput. Drone Appl. (IConDA)*, Kuching, Malaysia, Nov. 2017, pp. 60–65.

[41] Y. Wang, Y. Shen, X. Zhang, G. Cui, and J. Sun, "An improved non-dominated sorting genetic algorithm-II (INSGA-II) applied to the design of DNA codewords," *Math. Comput. Simul.*, vol. 151, pp. 131–139, Sep. 2018, doi: 10.1016/j.matcom.2018.03.011.

[42] Z. Liu, B. Wang, C. Zhou, X. Wei, Q. Zhang, Z. Yin, X. Fang, and Z. Zheng, "An improved iterated hybrid search for DNA codes design," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Guangzhou, People's Republic of China, Oct. 2018, pp. 99–105.

[43] S. A. H. Nair and P. Aruna, "Comparison of DCT, SVD and BFOA based multimodal biometric watermarking systems," *Alexandria Eng. J.*, vol. 54, no. 4, pp. 1161–1174, Dec. 2015, doi: 10.1016/j.aej.2015.07.002.

[44] S. Y. Yao and Q. T. Han, "Application of BFOA in maintenance task scheduling of ordnance equipments," *J. Nav. Aeronaut. Astron. Univ.*, vol. 26, no. 5, pp. 558–560, Apr. 2011, doi: 10.3969/j.issn.1673-1522.2011.05.018.

[45] K. M. Passino, "Biomimicry of bacterial foraging for distributed optimization and control," *IEEE Control Syst. Mag.*, vol. 22, no. 3, pp. 52–67, Mar. 2002, doi: 10.1109/MCS.2002.1004010.

[46] S. Dasgupta, S. Das, A. Abraham, and A. Biswas, "Adaptive computational chemotaxis in bacterial foraging optimization: An analysis," *IEEE Trans. Evol. Comput.*, vol. 13, no. 4, pp. 919–941, Aug. 2009, doi: 10.1109/TEVC.2009.2021982.
[47] J. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 4, pp. 1460–1465, Feb. 1998, doi: 10.1073/pnas.95.4.1460.
[48] R. Deaton, M. Garzon, R. C. Murphy, J. A. Rose, D. R. Franceschetti, and S. E. Stevens, "Reliability and efficiency of a DNA-based computation," *Phys. Rev. Lett.*, vol. 80, no. 2, pp. 417–420, Jan. 1998, doi: 10.1103/PhysRevLett.80.417.
[49] J. SantaLucia and D. Hicks, "The thermodynamics of DNA structural motifs," *Annu. Rev. Biophys. Biomolecular Struct.*, vol. 33, no. 1, pp. 415–440, Jun. 2004, doi: 10.1146/annurev.biophys.32.110601.141800.
[50] A. J. Maniotis, K. Valyi-Nagy, J. Karavitis, J. Moses, V. Boddipali, Y. Wang, R. Nuñez, S. Setty, Z. Arbieva, M. J. Bissell, and R. Folberg, "Chromatin organization measured by AluI restriction enzyme changes with malignancy and is regulated by the extracellular matrix and the cytoskeleton," *Amer. J. Pathol.*, vol. 166, no. 4, pp. 1187–1203, Apr. 2005, doi: 10.1016/S0002-9440(10)62338-3.
[51] R. H. M. Parland, H. M. Engelking, C. J. Jones, and G. D. Pearson, "Cleavage of type 2 adenovirus DNA BY HaeIII endonuclease. II. Map of HaeIII sites in EcoRI fragments C and E," *Biochim Biophys Acta.-Nucleic Acids Protein Synth.*, vol. 518, no. 3, pp. 424–439, May 1978, doi: 10.1016/0005-2787(78)90161-2.
[52] M. Nasri and D. Thomas, "Relaxation of recognition sequence of specific endonuclease HindIII," *Nucleic Acids Res.*, vol. 14, no. 2, pp. 811–821, Jan. 1986, doi: 10.1093/nar/14.2.811.
[53] X. Wang, L. T. Yang, Y. Wang, L. Ren, and M. J. Deen, "ADTT: A highly efficient distributed tensor-train decomposition method for IIoT big data," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 1573–1582, Mar. 2021, doi: 10.1109/TII.2020.2967768.
[54] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and J. Deen, "A tensor-based multi-attributes visual feature recognition method for industrial intelligence," *IEEE Trans. Ind. Informat.*, vol. 17, no. 3, pp. 2231–2241, Mar. 2020, doi: 10.1109/TII.2020.2999901.
[55] L. Ren, Z. H. Meng, X. K. Wang, L. Zhang, and T. L. Yang, "A data-driven approach of product quality prediction for complex production systems," *IEEE Trans. Ind. Informat.*, early access, Jun. 9, 2020, doi: 10.1109/TII.2020.3001054.
[56] G. H. Wadhams and J. P. Armitage, "Making sense of it all: Bacterial chemotaxix," *Nat. Rev. Mol. Cell Biol.*, vol. 5, pp. 1024–1037, Dec. 2004, doi: 10.1038/nrm1524.
[57] X. Liu, X. Xie, S. Wang, J. Liu, D. Yao, J. Cao, and K. Li, "Efficient range queries for large-scale sensor-augmented RFID systems," *IEEE/ACM Trans. Netw.*, vol. 27, no. 5, pp. 1873–1886, Oct. 2019, doi: 10.1109/TNET.2019.2936977.
[58] X. Liu, S. Chen, J. Liu, W. Qu, F. Xiao, A. X. Liu, J. Cao, and J. Liu, "Fast and accurate detection of unknown tags for RFID systems–hash collisions are desirable," *IEEE/ACM Trans. Netw.*, vol. 28, no. 1, pp. 126–139, Feb. 2020, doi: 10.1109/TNET.2019.2957239.
[59] X. Liu, J. Zhang, S. Jiang, Y. Yang, K. Li, J. Cao, and J. Liu, "Accurate localization of tagged objects using mobile RFID-augmented robots," *IEEE Trans. Mobile Comput.*, early access, Dec. 24, 2019, doi: 10.1109/TMC.2019.2962129.

YAO YAO received the B.S. degree from Northeast Petroleum University, Daqing, China, in 2017. He is currently pursuing the Ph.D. degree in computer science with the Dalian University of Technology, China.

His research interests include Mobile Edge Computing and Biocomputing.

JIANKANG REN (Member, IEEE) received the B.Sc., M.E., and Ph.D. degrees in computer science from the Dalian University of Technology, China, in 2008, 2011, and 2015, respectively.

He was a Visiting Scholar with the Computer and Information Science Department, University of Pennsylvania, USA, from September 2013 to September 2014. He is currently an Associate Professor with the School of Computer Science and Technology, Dalian University of Technology. His research interests include cyber-physical systems (CPS), mobile edge computing, dispersed computing, and computational intelligence.

RAN BI received the B.S. degree from the Department of Mathematics, Harbin Institute of Technology, Harbin, China, and the M.S. and Ph.D. degrees from the Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China.

She is currently an Associate Professor with the School of Computer Science and Technology, Dalian University of Technology, Dalian, China. Her research interests include wireless networking, sensory data management, the Internet of Things, and vehicular networking.

QIAN LIU (Member, IEEE) received the B.S. and M.S. degrees from the Dalian University of Technology, Dalian, China, in 2006 and 2009, respectively, and the Ph.D. degree from The State University of New York at Buffalo (SUNY-Buffalo), Buffalo, NY, USA, in 2013.

She was a Postdoctoral Fellow with the Ubiquitous Multimedia Laboratory, SUNY-Buffalo, from 2013 to 2015. She was a Postdoctoral Fellow with the Chair of Media Technology and the Chair of Communication Networks, Technical University of Munich, from 2016 to 2017. She is currently an Associate Professor with the Department of Computer Science and Technology, Dalian University of Technology, China. Her current research interests include multimedia transmission over MIMO systems, IEEE 802.11 wireless networks and LTE networks, device-to-device communication, energy-aware multimedia delivery, and the Tactile Internet.

Dr. Liu received the Best Paper Runner-up Award at the 2012 International Conference on Complex Medical Engineering and was the Finalist for the Best Student Paper Award at the 2011 IEEE International Symposium on Circuits and Systems. She received the Alexander von Humboldt Fellowship in 2015.