MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 291-308

ISSN 0340 - 6253

Evaluating the Different Combinatorial Constraints in DNA Computing Based on Minimum Free Energy

Qiang Zhang[1], Bin Wang[1, 2], Xiaopeng Wei[1, 2]

[1]*Key Laboratory of Advanced Design and Intelligent Computing (Dalian University), Ministry of Education, Dalian, 116622, China*

[2]*School of Mechanical and Engineering, Dalian University of Technology, Dalian, 116024 China*

zhangq@dlu.edu.cn

(Received May 6, 2010)

Abstract: The design of DNA sequences is a key process in DNA computing, DNA-based steganography and other DNA-based applications. The criteria, to constrain the DNA sequence design, include different combinatorial constraints proposed by different researchers. However, the unique and all-purpose constraint (or constraints) has not been found yet. In this paper, we firstly obtain some DNA sequences sets from our algorithm. Then, we employ the minimum free energy (MFE, for short) to evaluate the different combinatorial constraints, because the MFE criterion is the minimum value among free energies of all the possible structures and the effective approach to control the generation of unexpected secondary structure of DNA sequences may cause error. Comparisons of the results suggests that the $C1$ combinatorial constraint could be the best combinatorial constraints and the CC (or WW) constraint be the most important single constraint for designing DNA sequence sets.

1. Introduction

With the development of biotechnology, especially DNA-based biotechnology, DNA molecule have been more and more applied in computer science, communication and other subjects, which bring flying development in many aspects. To the DNA sequence design, DNA computing, firstly proposed by Adleman in 1994,

is an important one in these applications [1]. Subsequently, Clelland proposes a method of DNA-based steganography that mean hiding of secret messages among other information to conceal their existence with DNA microdot [2]. In his paper, he designs the DNA sequences for plaintext encoding where a unique base triplet is assigned to each letter of the alphabet, each numeral and some special characters. Dietrich use the DNA sequences as storage medium based on the property of DNA molecule [3]. With the continuous deepening of the study, the DNA-based method will apply in the more and wider areas in the future.

A single strand of DNA consist of four bases (nucleotides), adenine (A), guanine (G), thymine (T), cytosine (C). Two single strands of DNA can form (under appropriate conditions) a double strand by hybridization reaction, if the respective bases are the Watson-Crick complements of each other-A matches T and C matches G; also 3' end matches 5' end. The hybridization reaction between two DNA sequences is very important for DNA computing, because its efficiency and accuracy directly influences the reliability of DNA computing. However, the false hybridization is unavoidable to emerge because of limit of biologic technology. False hybridization reaction in DNA computing can be assorted two categories: One is false positive, the other is false negative. The former is the hybridization reaction between two unmatched DNA sequences. The latter is that two match DNA sequences do not hybridize each other. The false positive is result of the lack of similarity between DNA sequences [4,5]. The false negative is result of the mistake in the biochemical operation. The mean ways that are used to reduce the emergence of false hybridization reaction can be stated as follows:

(1) optimizing the DNA sequences and decreasing the similarity between the DNA sequences;
(2) using specific biochemical methods;
(3) enhancing the accuracy of biochemical operation.

1.1 Design DNA Sequence

The design of DNA sequence is an approach for a robust computation or application by designing DNA sequences that satisfy some constraints to avoid unexpected false hybridization reactions. The purpose is to design DNA sequences that are used as elemental components of computation, DNA-based steganography and other DNA-based applications and to improve the veracity and the reliability of DNA-based applications.

The research of designing DNA sequence has obtained a lot of enormous progress in many fields, but there are a number of problems not solved. The problem that is urgent to solve is that how to combine the recognition of information specific in DNA-based applications with varied kind of factors of biochemical reaction and to build a unique and all-purpose standard for designing DNA sequences. In this paper, we employ the minimum free energy (MFE, for short) criterion to evaluate the constraints used in DNA sequence design. The significance of our work is to find the best constraint (or constraints) to design DNA sequence based on the MFE criterion and to improve the accuracy of hybridization reaction between the two single DNA sequences.

1.2 The Significance of Designing DNA Sequence

The significance of researching DNA sequence design can be briefly stated as follows:

(1) The problem of designing DNA sequence is to produce the DNA sequences that satisfy the constraints. Therefore, it could ensure the quality of DNA coding and use the shortest DNA sequences to code every information unit.

(2) According to the actual needs, it can obtain the better DNA sequences and use them to decrease the emergence of false hybridization reaction and improve the accuracy of DNA –based application.

(3) According to the results, we could use the least DNA sequences to express the data in the DNA data storage. It could decrease the redundancy of DNA data storage that could be used in many wide fields.

2 Related Works

There are two main problems of designing DNA sequence: one is to research the quantity of DNA sequences, namely obtaining the better DNA sequences used in DNA-based applications; the other is to find the quality of DNA sequences, namely designing DNA sequence sets.

The researching the quantity of DNA sequences is earlier than the sets of DNA sequences. Baum [4] proposes a new method-the minimal same subsequences that are coded between DNA sequences used as information unit should be more than a constant. This method could decrease the nonspecific hybridization reaction between two single DNA sequences. Deaton [5,6] proposes that the DNA coding should be combined with biochemistry reactions and research of the reliability of coding problems from information theory. Moreover, he firstly proposes the algorithm of design DNA sequences based on genetic algorithm. Wood [7] proposes a method of designing DNA sequences which has error correction function and use it in DNA computing. Hartemink [8,9] proposes the designing method based on distance constraints (such as Hamming distance, we will be described in subsequent chapter) and free-energy criterion. S.Y. Shin [10,11] uses the algorithm of Multi-objective evolutionary to design DNA sequences and develops a system that named NACST by genetic algorithm. W.B. Liu [12] proposes the algorithm that used the template frame to design DNA sequences. We used the improved genetic algorithm to design DNA sequences used in DNA computing [13]. In [13], we use the combinatorial constraints that include distance constraints, GC content and continuity constraint. X.C. Zhang uses the invasive weed optimization technology to design DNA sequences [14]. R. Zhang employs an improved particle swarm optimization algorithm to solve the problem of DNA sequence design and integrate it into required combinatorial and thermodynamic constraints for DNA computing [15]. Their experimental results show that the proposed method is effective and convenient for the user to design and select effective DNA sequences in silicon for controllable DNA computing.

The main problem of designing DNA sequences set is to research the amount of

DNA sequences satisfy the certain constraints. The ways designing the set of DNA sequence include two main methods: one is theoretical derivation. It could obtain the structural method of DNA sequence set which satisfy the constraints and the approximate upper or lower bounds, such as [16-18]. The other is to use the intelligent algorithm to search and obtain the improved lower bounds of DNA sequence sets, such as stochastic local search algorithm [19,20], hybrid randomized neighborhoods search algorithm [21], dynamic neighborhood search algorithm [22], variable neighborhood search algorithm [23], improved genetic algorithm [24] and swarm optimization algorithm (PSO) algorithm [25].

In the theoretical derivation, the main idea is to apply the research results of 2-component code and q-component code to the DNA coding and improve them [26], such as Sphere-Packing bound, Singleton upper bounds, Gilbert-Varshamov lower bounds, Plotkin lower bounds and so on. There are some introductions and some corresponding derivation in the [16] and [17]. Applying these results, they could reduce the values range of DNA sequence sets. In [17], the authors deeply research the theoretical bounds that satisfy the Hamming distance (WW, for short). In [16], the authors deeply research the theoretical bounds which respectively satisfy the WW, WW and reverse-complement Hamming distance (WC, for short), and give the relation between them. In [25], the authors research the GC content, WC and GC content constraints. In [27], the authors use the method that combines linear construction with stochastic local search algorithm. They improve the some lower bounds that satisfy the GC content, WC and GC content constraints.

In the research of intelligent algorithm, the main idea is to use intelligent algorithm to search DNA sequences sets that satisfy constraints. In the single constraint, this method usually is used to improve the lower bounds. Because the theoretical research is hard to find the relation between the combinatorial constraints, such as the relation between GC content and other distance constraints, intelligent algorithm could improve the upper and lower bounds in the combinatorial constraints. In [19], the authors use the stochastic local search algorithm to improve the lower bounds that satisfy the WW and WC combinatorial constraints. The results are

compared with the theoretical value. At the same time, they also improve the bounds that satisfy the WW and WC constraints and obtain the approximate bounds that satisfy the WC and GC content constraints. In [21], the authors improve the stochastic local search algorithm and the results that are from the [19]. In [22], the authors use the dynamic neighborhood search to improve the lower bounds that satisfy the WC and GC content constraints. In [23], the authors use the variable neighborhood search algorithm improve the lower bounds satisfy GC content constraint, and GC content and WC combinational constraints.

2.1 Minimum Free Energy

A DNA sequence s is a string composed of alphabet $\Sigma=\{\mathrm{A,G,C,T}\}$. A DNA sequence or sequences maybe form secondary structures by the Watson-Crick property, which are also called conformations. Each conformation of a sequence (or sequences) has a Gibbs standard free energy. The Minimum Free Energy (MFE, for short) of a sequence or sequences is the minimum value among free energies of all possible conformations of a sequence (or sequences). It is known that a conformation with a small Gibbs standard free energy is more stable than the one with larger Gibbs standard free energies [28].

$\Delta G(u,v)$ denotes the value of MFE between two DNA sequences u,v, which can be calculated by PairFold [29]. In addition, s' denotes the Watson-Crick reverse-complement sequence of DNA sequence s. S is the set of DNA sequences s, $\mathrm{S'}$ is the set of s'. In order to calculate the free energy gap δ, we need some definitions that are stated as following:

(1) Sequence-Sequence Constraint: for all pairs of u_i, v_j in S,

$$\Delta G_{ww}(u_i)=\min_{1\le j\le n}\{\Delta G(u_i,v_j)\} \tag{1}$$

(2) Sequence-Complement Constraint: for all pairs of u_i in S, $v_j^{'}$ in $\mathrm{S'}$, and $i\neq j$,

$$\Delta G_{wc}(u_i)=\min_{1\le j\le n, i\neq j}\{\Delta G(u_i,v_j^{'})\} \tag{2}$$

(3) Complement-Complement Constraint: for the pairs of $u_i^{'}, v_j^{'}$ in S', $\Delta G_{cc}(u_i) = \min_{1 \le j \le n} \{\Delta G(u_i^{'}, v_j^{'})\}$ (3)

(4) Sequence-Self-Complement Constraint: for all pairs of u_i in S, $u_i^{'}$ in S', and $u = (u')'$, $\Delta G_{ws}(u_i) = \min_{1 \le i \le n} \{\Delta G(u_i, u_i^{'})\}$ (4)

(5) Free energy gap: the free energy gap is denoted by δ. For two DNA sequences u and v, $\min\{\Delta G(u,v), \Delta G(u,v'), \Delta G(u',v')\} - \Delta G(u,u') \ge \delta$ (5)

where $u, v \in S$, $u', v' \in S'$. Roughly speaking, the larger δ is, the larger the gap between the free energy of desired and undesired hybridizations, and thus the better (the quality of DNA sequences set) the set is [30, 36].

Tulpan et al. [30] early research the DNA sequence set design based on MFE, which uses PairFold package. They describe a new algorithm for design of DNA sequence sets, for using in DNA computations or universal microarrays. Their algorithm can design sets satisfy any of several thermodynamic and combinatorial constraints, which aim to maximize desired hybridizations between strands and their complements, while minimizing undesired cross-hybridizations. In the Garzon et al. paper [31], authors report results of a *tour de force* to conduct an exhaustive search to produce DNA sequence sets that are arguably of sizes comparable to that of maximal sets while guaranteeing high quality, as measured by the minimum Gibbs energy between any pair of DNA sequences. By comparing their experimental results with previous work, the results has been improved the lower bounds of DNA sequence sets. Subsequently, in [28], Kawashimo et al. use dynamic neighborhood searches to design DNA sequence sets and further improve the Garzon's value [31]. In the Kawashimo's paper, authors introduce techniques to reduce such time-consuming evaluations of MFE, by which the proposed dynamic neighborhood search strategy become applicable to the thermodynamical constraints in practice. Recently, they propose a

new speeding up local-search type algorithms for designing DNA sequences based on MFE [32]. Comparing the results, their algorithm succeeded in generating better DNA sequence sets than exiting methods.

2.2 The Constraints

In this part, we introduce the distance constraints that are frequently used in designing DNA sequence. The Hamming distance constraint criterion mainly includes word-word Hamming distance constraint (WW, for short), word-complement Hamming distance (WC, for short) and complement-complement Hamming distance (CC, for short).

Garzon firstly proposed the definition problem of designing DNA sequences for DNA computing [33]. The definition is as follow: in the alphabet $\Sigma=\{A,G,C,T\}$, there exists a set S with the length of n and size of $|S|=4^n$. A subset $C \subseteq S$ and let u, v any two codes in the C satisfy

$$\tau(u,v) \geq d \tag{6}$$

d is a positive integer, τ is the constraint criteria (or criterion) for design DNA sequences, such as Hamming distance criterion.

2.2.1 Word-word Hamming distance (WW, for short)

Word-word Hamming distance constraint: for the DNA sequences u, v with given length n (written from the 5'- to the 3'-end), $H(u,v)$ denotes the Hamming distance between u and v. $WW(u_i)$ denotes the minimal of $H(u_i, v_j)$ in all DNA sequences and should not be less than certain parameter d,

$$WW(u_i) = \min_{1\leq j\leq n, j\neq i} \left\{ H(u_i, v_j) \right\} \geq d \tag{7}$$

2.2.2 Word-complement Hamming distance (WC, for short)

Word-complement Hamming distance: for the DNA sequences u, v with given length n (written from the 5'- to the 3'-end), $H(u, v')$ denotes the Hamming distance between u and v'. $WC(u_i)$ denotes the minimal of $H(u_i, v_j')$ in all DNA sequences and should not be less than certain parameter d, i.e.

$$WC(u_i) = \min_{1 \le j \le n, j \ne i} \left\{ H(u_i, v_j') \right\} \ge d \tag{8}$$

2.2.3 Complement-complement Hamming distance (CC, for short)

Complement-complement Hamming distance constraint: for the DNA sequences u, v with given length n (written from the 5'- to the 3'-end), $H(u', v')$ denotes the Hamming distance between u' and v'. $CC(u_i')$ denotes the minimal of $H(u_i', v_j')$ in all DNA sequences and should not be less than certain parameter d,

$$CC(u_i') = \min_{1 \le j \le n} \{H(u_i', v_j')\} \ge d \tag{9}$$

2.2.4 GC content constraint

GC content constraint affects the chemical properties of DNA sequences. A fixed percentage of the nucleotides within each DNA sequence is either G or C. Using this constraint, we assume that this percentage is $(\lfloor n/2 \rfloor / n)$ %.

2.3 The Combinatorial Constraints

$C0$: This combinatorial constraints include the WW, WC, CC and GC content constraints.

$C1$: This combinatorial constraints include the WC, CC and GC content constraints (Delete the WW constraint).

$C2$: This combinatorial constraints include the WW, CC and GC content constraints (Delete the WC constraint).

$C3$: This combinatorial constraints include the WW, WC and GC content

constraints (Delete the CC constraint).

$C4$: This combinatorial constraints include the WW, WC and CC constraints (Delete the GC content constraint).

$C5$: This combinatorial constraints include the WW and WC constraints.

$C6$: This combinatorial constraints include the WW and CC constraints.

$C7$: This combinatorial constraints include the WW and GC content constraints.

$C8$: This combinatorial constraints include the WC and CC constraints.

$C9$: This combinatorial constraints include the WC and GC content constraints.

$C10$: This combinatorial constraints include the CC and GC content constraints.

In this paper, we employ the improved genetic algorithm to design DNA sequence sets that satisfy the combinatorial constraints and gauge the quality of the DNA sequences sets by the free energy gap based on MFE that is calculated by the PairFold package [31]. We choose the best combinatorial constraints from $C0$ to $C10$ by comparing their scores.

3 Design of Algorithm

According to the introduction above, different algorithms could be used to design DNA sequence sets by many various authors. In this paper, we employ the improved genetic algorithm to design DNA sequence sets based on the combinatorial constraints from $C0$ to $C10$. The improved genetic algorithm could conquer the shortages and enhance the global search capability of traditional genetic algorithm based on the characteristic of DNA sequence set. The improved areas can be briefly stated as follows:

(1) Initializing the populations of algorithm with the evenly distributed method. It can enhance the multiformity of populations based on global field. According to the number of populations, the populations are evenly distributed in the value scope by the evenly distributed method.

(2) Randomly re-initializing the populations when they satisfy certain condition. It

can enhance the ability of conquering premature convergence. The time of re-initializing the populations is only once, because with the times increasing, the convergence of algorithm will be decreased.

(3) In the mutation process, we adjust the probability of mutation operator with dynamic method. In the traditional genetic algorithm, the algorithm adopts unique value to process the mutation operation, which could certainly decrease the convergence of algorithm.

The optimization problem is defined by the problem of maximum value, and we employ average weight to deal with the function of evaluation. We denote fitness function $f(i)$ is $f(i)=\sum_{j=1}^{m}\omega_j f_j(i)$,

$$f_j(i)\in\{WW(u_i),WC(u_i),CC(u_i)\} \tag{10}$$

Where $\omega_j=1$ is the weight of each constraint, m is the number of constraints and $f_j(i)$ are the constraints which have been selected.

The main process of algorithm is that: initializing DNA sequences with evenly distributed method, selecting the sequences that satisfy the constraint (or constraints) from these sequences, generating new DNA sequences by selection, crossover and mutation operator, lastly obtaining the DNA sequence sets. Fig.1 is the flowchart of algorithm.

The steps of designing DNA word set by the improved genetic algorithm are stated as follows:

Step 1: Setting parameters and initializing population with evenly distributed method.

Step 2: Calculating the value of fitness function. We employ the MeanF to denote the mean of the fitness function. If MeanF < $\sum_{i=1}^{m} f(i)/\mathrm{m}$ then randomly re-initializing the populations.

Step 3: Generating the next generation population by selection、crossover and mutation. The algorithm uses the random tournament selection in the

selection process and three-point crossover strategy in the crossover process. The size of tournament is equal to 2, and the number of times to repeat is equal to the 10% of the total number of population in the random tournament selection. In the mutation process, if one of the fitness is larger than MeanF, its probability of mutation is 0.01. In addition, if it is equal to MeanF, its probability of mutation is 0.03. Else, its probability of mutation is 0.3. It is the process of dynamic adjustment of the probability. If the generation is less than 200, go to step2; if not, then go to step4.

Step 4: Ending and outputting results.

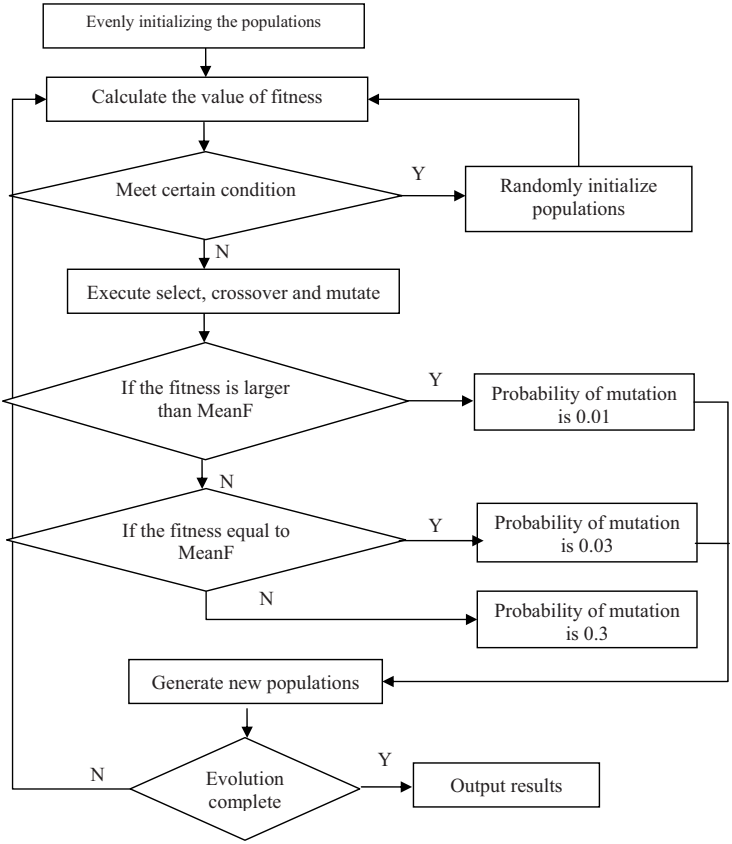

Fig.1 The flowchart of algorithm.

Note that our algorithms have succeeded in many different combinatorial constraints for designing the DNA sequence sets and obtained the better results than previous work [24, 34, 35]. Therefore, the algorithm used in this paper has enough effect to design DNA sequence sets that satisfy the different combinatorial constraints.

4 Experimental results

The parameters of improved genetic algorithm used in our example are: the size of population is 1000; the probability of crossover is 0.45; the initial probability of mutate which is initialized is 0.01. In order to control the time of running algorithm, the generation is 200. We use the PairFold package [22] to calculate the MFE of two DNA sequences. The setting temperature is 37 C. In order to increase the reliability of our experimental results, we did fifty experiments for every value and employed the mean of the fifty experiments as final results.

The bold face values in the Table 1 and Table 2 are the energy gaps of DNA sequence sets that satisfy the constraints from $C0$ to $C10$, respectively. n denotes the length of DNA sequences and d denotes the distance. The numbers in the brackets are the values that are sorted in descending order by the each row. For an overall evaluation of each combinatorial constraint, we calculate the energy gaps by the different lengths and distances. The last row is the overall evaluation for each combinatorial constraint and is equal to the sum of the all sorted value for each column.

Table 1. Results satisfying the constraints from $C0$ to $C4$.

n	d	C0	C1	C2	C3	C4
4	4	**0.676(3)**	**0.696(2)**	**0.542(5)**	**0.716(1)**	**0.632(4)**
5	4	**0.410(5)**	**0.494(2)**	**0.526(1)**	**0.456(4)**	**0.476(3)**
6	4	**0.598(2)**	**0.380(4)**	**0.620(1)**	**0.496(3)**	**-0.026(5)**
7	4	**0.114(4)**	**0.282(1)**	**0.192(3)**	**0.282(1)**	**-0.092(5)**
8	4	**-0.320(5)**	**-0.264(2)**	**-0.266(4)**	**-0.264(2)**	**-0.206(1)**
5	5	**1.048(5)**	**1.360(1)**	**1.300(2)**	**1.236(3)**	**1.110(4)**

6	5	**1.772(1)**	**1.418(4)**	**1.632(2)**	**1.608(3)**	**0.468(5)**
7	5	**1.338(2)**	**1.246(3)**	**1.342(1)**	**1.246(3)**	**0.956(5)**
8	5	**0.144(1)**	**0.070(3)**	**0.008(5)**	**0.030(4)**	**0.076(2)**
6	6	**2.248(1)**	**2.176(2)**	**2.130(3)**	**2.044(4)**	**1.844(5)**
7	6	**2.648(1)**	**2.330(3)**	**2.316(4)**	**2.410(2)**	**2.068(5)**
8	6	**0.882(3)**	**1.148(1)**	**0.996(2)**	**0.652(5)**	**0.816(4)**
7	7	**2.728(2)**	**2.722(3)**	**2.566(5)**	**2.586(4)**	**2.788(1)**
8	7	**2.048(3)**	**2.112(2)**	**2.864(1)**	**2.042(4)**	**1.598(5)**
8	8	**3.096(3)**	**2.980(4)**	**2.870(5)**	**3.102(2)**	**3.400(1)**
score		41	37	44	45	55

Table 2. Results satisfying the constraints $C1$ and from $C5$ to $C10$.

n	d	C1	C5	C6	C7	C8	C9	C10
4	4	**0.696(1)**	**0.526(2)**	**0.046(4)**	**0.046(4)**	**0.526(2)**	**-0.214(7)**	**0.046(4)**
5	4	**0.494(3)**	**0.574(1)**	**0.0084(4)**	**-0.144(5)**	**0.510(2)**	**-0.670(7)**	**-0.144(5)**
6	4	**0.380(1)**	**0.018(3)**	**-0.194(6)**	**-0.174(4)**	**0.022(2)**	**-0.740(7)**	**-0.174(4)**
7	4	**0.282(1)**	**-0.018(3)**	**-0.420(4)**	**-0.442(5)**	**0.020(2)**	**-0.892(7)**	**-0.442(5)**
8	4	**-0.264(2)**	**-0.246(1)**	**-0.386(4)**	**-0.564(5)**	**-0.276(3)**	**-0.974(7)**	**-0.564(5)**
5	5	**1.360(1)**	**1.060(3)**	**0.410(6)**	**1.032(4)**	**1.060(2)**	**-0.050(7)**	**0.964(5)**
6	5	**1.418(1)**	**0.454(7)**	**0.628(4)**	**0.696(3)**	**0.508(5)**	**0.500(6)**	**0.724(2)**
7	5	**1.246(1)**	**0.780(3)**	**0.280(6)**	**0.522(4)**	**1.024(2)**	**-0.722(7)**	**0.486(5)**
8	5	**0.070(5)**	**0.084(4)**	**0.136(3)**	**0.204(2)**	**0.008(6)**	**-0.798(7)**	**0.262(1)**
6	6	**2.176(1)**	**1.600(3)**	**1.284(6)**	**1.394(5)**	**1.772(2)**	**0.048(7)**	**1.514(4)**
7	6	**2.330(1)**	**2.142(2)**	**1.178(6)**	**1.536(5)**	**2.054(3)**	**-0.442(7)**	**1.812(4)**
8	6	**1.148(3)**	**0.840(5)**	**1.022(4)**	**1.288(1)**	**0.652(6)**	**-0.606(7)**	**1.212(2)**
7	7	**2.722(2)**	**2.558(3)**	**1.562(6)**	**2.384(4)**	**2.788(1)**	**0.192(7)**	**2.340(5)**
8	7	**2.112(6)**	**2.268(5)**	**2.368(4)**	**2.658(2)**	**2.546(3)**	**-0.276(7)**	**2.712(1)**
8	8	**2.980(4)**	**3.324(1)**	**2.578(6)**	**3.018(3)**	**2.968(5)**	**0.610(7)**	**3.142(2)**
score		33	46	73	56	46	104	54

5 Conclusions

In this paper, we firstly present an improved genetic algorithm to design DNA sequence sets that satisfy the different combinatorial constraints based on minimum free energy criterion. We employ the energy gap to evaluate the quality of DNA sequence sets, namely the larger the energy gap of DNA sequence set is, the larger the gap between the free energy of desired and undesired hybridizations is, and thus the better the set is [30,36]. For an overall evaluation of each combinatorial constraint from $C0$ to $C10$, we calculate the energy gaps by the different lengths and distances. Summary, the smaller the score of the constraints is, the greater the impact of the constraints on the quality of DNA sequence set is, and vice versa.

Comparing all the constraints from Table 1 and Table 2, the results suggest that the $C1$ constraint is the best combinatorial constrain for designing DNA sequence set. In the Table 2, the score of $C5$ is equal to $C8$ and the score of $C7$ is nearly equal to $C10$ (having the same energy gaps in some rows). Comparing the definition of WW and CC constraint, we conjecture that the impact of the WW constraint on the designing DNA sequence sets would be the same as the impact of the CC constraint. So the impact of $C6$ would approximate the impact of one constraint in $C6$. Because the score of $C9$ is the largest one, $C9$ is the worst combinatorial constraint. However, in Table 1, the $C0$ and $C3$ constraints both include the $C9$ constraint and are not the worst combinatorial constraints. For these reasons, the CC (or WW) constraint is the most important single constraint for designing DNA sequence sets.

In the further, we have some works need to do. We will try to proof our conjecture in theory, use the $C1$ combinatorial constraint to compare other constraints, such as edit distance constraint and find the best constraint or combinatorial constraints to design DNA sequence sets.

ACKNOWLEDGMENT

This paper is supported by the National Natural Science Foundation of China (Grant No. 30870573) and the National High Technology Research and Development Program ("863"Program) of China (No.2009AA01Z416).

References

[1] L. Adleman, Molecular computation of solution to combinatorial problems, *Science* **266** (1994) 1021–1024.

[2] C. T. Celland, V. Risca, C. Bancroft, Hiding messages in DNA microdots, *Nature* **399** (1999) 533–534.

[3] A. Dietrich, W. Been, Memory and DNA, *J. Theor. Biol.* **208** (2001) 145–149.

[4] E. B. Baum, DNA sequences useful for computation, *Proceedings of 2nd DIMACS Workshop on DNA Based Computers* 1996, pp. 122–127.

[5] R. Deaton, R. C. Murphy, J. A. Rose, M. Garzon, D. R. Franceschetti, S. E. Jr. Stevens, A DNA based implementation of an evolutionary search for good encodings for DNA computation, *Proceedings of the 1997 IEEE International Conference on Evolutionary Computation*, 1997, pp. 267–272.

[6] R. Deaton, Franceschetti, M. Garzon, A. J. Rose, R. C. Murphy, S. E. J. Stevens, Information transfer through hybridization reaction in DNA based computing, *Proceedings of the Second Annual Conference, California*, 1997, pp. 463–471.

[7] D. H. Wood, Applying error correcting codes to DNA computing, *4th DIMACS workshop on DNA based computers, Pennsylvania*, 1998, pp. 109–110.

[8] A. J. Hartemink, D. K. Gifford, Thermodynamics simulation of deoxyoligonucleotide hybridize for DNA computation, *3rd DIMACS meeting on DNA based computers, Pennsylvania,* 1997, pp. 23–25.

[9] A. Hartemink, D. K. Gifford, J. Khodor, Automated constraint-based nucleotide sequence selection for DNA computation, *4th DIMACS workshop on DNA based computers, Pennsylvania*, 1998, pp. 287–297.

[10] S. Y. Shin, D. Min, Evolutionary sequence generation for reliable DNA computing, Evolutionary Computation, 2002. CEC'02. *Proceedings of the 2002 Congress on.*

[11] S. Y. Shin, I. H. Lee, D. Kim, B. T. Zhang, Multi-objective evolutionary optimization of DNA sequences for reliable DNA computing, *IEEE Transac. Evolution. Comput.* **9** (2005) 143–158.

[12] W. B. Liu, S. D. Wang, L. Gao J. Xu, DNA sequence design based on template strategy, *J. Chem. Inf. Comput. Sci.***43** (2003) 2014–2018.

[13] B. Wang, Q. Zhang, R. Zhang, Design of DNA sequence based on improved genetic algorithm, *Lecture Notes in Computer Science LNCS 5226*, 2008, pp. 9–14.

[14] X. C. Zhang, Y. F. Wang, G. Z. Cui, Y. Niu, J. Xu, Application of a novel IWO to the design of encoding sequences for DNA computing, *Comput. Math. Appl.* **57** (2009) 2001–2008.

[15] R. Zhang, Q. Zhang, B. Wang, Improved particle swarm optimization algorithm for designing DNA codewords, *Inter. Interdisc. J.* **12** (2009) 497–505.

[16] A. Marathe, A. Condon, R. Corn, On combinatorial DNA word design, *J. Comput. Biol.* **18** (2001) 201–220.

[17] T. G. Bogdanova, A. E. Brouwer, S. N. Kapralov, P. R. J. Ostergard, Error-correcting codes over an alphabet of four elements, *Des. Cod. Crypt.* **23 (**2001) 333–342.

[18] B. Paolo, L. Anne, M. Vincenzo, M. Victor, Superposition based on Watson–Crick–like complementarity, *Theor. Comput. Sys.* **39** (2006) 503–524.

[19] D. C. Tulpan, H. Hoos, A. Condon, Stochastic local search algorithms for DNA word design, *Lecture Notes in Computer Science, DNA 8, LNCS 2568*, 2002, pp. 229–241.

[20] Y. M. Chee, S. Ling, Improved lower bounds for constant GC-content DNA codes, *IEEE Trans. Inf. Theor.* **54** (2008) 391–394.

[21] D. C. Tulpan, H. Hoos, Hybrid randomized neighborhoods improve stochastic local search for DNA code design, *Proc. Advances in Artificial Intelligence, 16th Conference of the Canadian Society for Computational Studies of Intelligence, Lecture Notes in Computer Science LNCS 2671*, 2003, pp. 418–433.

[22] S. Kawashimo, H. Ono, K. Sadakane, M. Yamashita, DNA sequence design by dynamic neighborhood searches, *Lecture Notes in Computer Science, DNA 12, LNCS4287*, 2006, pp. 157–171.

[23] R. Montemanni, D. H. Smith, Construction of constant GC-content DNA codes via a variable neighborhood search algorithm, *J. Math. Model. Algor.* **7** (2008) 311–326.

[24] Q. Zhang, B. Wang, R. Zhang, C. X. Xu, Genetic algorithm-based design for DNA sequences sets, *Chinese J. Comput.* **31** (2008) 2193–2199.

[25] R. Zhang, Q. Zhang, B. Wang, DNA sequence sets design by particle swarm optimization algorithm, *Inter. J. Innov. Comput. Inf. Control* **5** (2009) 2249–2255.

[26] O. D. King, Bounds for DNA codes with constant GC-content, *Elec. J. Comb.* **10** (2003) #R33: 1–13.

[27] P. Gaborit, O. D. King, Linear constructions for DNA codes, *Theor. Comput.* **334** (2005) 99–113.

[28] S. Kawashimo, H. Ono, K. Sadakane, M. Yamashita, Dynamic neighborhood searches for thermodynamically designing DNA sequence, *Lecture Notes Comput. Sci. DNA 13*, *LNCS4848*, 2008, pp. 130–139.

[29] M. Andronescu, Z. Zhang, A. Condon, Secondary structure prediction of interacting RNA molecules, *J. Molec. Biol.* **345** (2005) 987–1001.

[30] D. C. Tulpan, M. Andronescu, S. B. Chang, M. R. Shortreed, A. Condon, H. H. Hoos, L. M. Smith, Thermodynamically based DNA strand design, *Nucleic Acids Res.* **33** (2005) 4951–4964.

[31] M. Garzon, V. Phan, S. Roy, A. Neel, In search of optimal codes for DNA computing, *Lecture Notes Computer Science, LNCS4287*, 2006, pp. 143–156.

[32] S. Kawashimo, Y. K. Ng, H. Ono, K. Sadakane, M. Yamashita. Speeding up local–search type algorithms for designing DNA sequences under thermodynamical constraints, *Lecture Notes Computer Science, DNA 14, LNCS5347*, 2009, pp. 168–178.

[33] M. Garzon, R. Deaton, L. F. Nino, S. E. Stevens, M. Wittner, Genome encoding for DNA computing, *Proc. Third Genetic Programming Conf., Madison*, 1998, pp. 684–690.

[34] Q. Zhang, B. Wang, Designing DNA sequences satisfy combinational constraints, *J. Comput. Theor. Nanosci.* in press.

[35] B. Wang, Q. Zhang, R. Zhang, C. X. Xu, Improved lower bounds of DNA coding, *J. Comput. Theor. Nanosci.* **7** (2010) 638–641.

[36] M. R. Shortreed, S. B. Chang, D. G. Hong, M. Phillips, B. Campion, D. C. Tulpan, M. Andronescu, A. Condon, H. H. Hoos, L. M. Smith. A thermodynamic approach to designing structure–free combinatorial DNA word sets, *Nucleic Acids Res.* **33** (2005) 4965–4977.