MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **66** (2011) 371-380

ISSN 0340 - 6253

On the Bounds of DNA Coding with H-Distance

Qiang Zhang and Bin Wang

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

zhangq@dlu.edu.cn

(Received October 18, 2010)

Abstract

As a new research field, DNA computing has received much more and more attention by the researchers all over the world. DNA computing is a new biologic computing method, which uses DNA molecule as computing medium and biochemical reaction as computing tool. Because DNA computing is carried out by hybridization reaction, the quality of DNA sequences is very important for DNA computing. DNA coding could decrease the emergence of false negative and false positive, which affects the reliability and the accuracy of DNA computing. The problem of codeword design plays an important role in the DNA coding. In this paper, the improved lower bounds of DNA codeword is established for the capacity of DNA to encode information using a combinatorial model of DNA homology given by the so-called h-distance. By comparing our experimental results with the previous works, the results improve the lower bounds and further shorten the value range of DNA codeword.

1 Introduction

The combination of computer science and biological science-DNA computing is a new field which uses DNA molecule as computing medium and biochemical reaction as computing tool. In 1994, Dr Adleman released *Molecular Computation of Solutions to Combinatorial Problems in Science*, which indicates DNA computing comes into being [1,23,24]. Because the structure of DNA molecule could store an enormous amount of data and have the ability of parallel reaction, DNA computing will replace the traditional electronic computer in the future. Along with the development of biologic technology, DNA computing will solve more and more complex problems, especially NP problems. Finally, it will product a new DNA computer which could bring the flying development of Mathematics, Computer Science

and other subjects.

Hybridization reaction is the important step in DNA computing between DNA sequences satisfied the principle of Watson-Crick complement, which directly influences the reliability and accuracy of DNA computing [26–30]. Then, the good DNA sequences are closely related to improving the efficiency and reliability of DNA computing. To obtain good DNA sequences, researchers always introduce the characteristic physical chemistry and combinatorial constraints. For these reasons, the DNA coding problem is an important step for DNA computing. About the DNA coding, the first researcher is Baum who proposes a new method which used the DNA sequences to code every information unit, and defines that the minimal length of same subsequences should be more than a constant [2]. This method could decrease the nonspecific hybridization between DNA sequences. Deaton [3,4] proposes the DNA coding should be combined with biochemistry reaction. According to the information theory, he further researches the reliability of coding problems. What's more, he proposes the coding method based on genetic algorithm. Garzon [5] firstly proposed the definition of coding problem in DNA computing. Wood [6] proposes a method of designing DNA sequences which has an error correction function. Hartemink [7,8] proposes a coding method based on constraints and a coding designing method based on Gibbs energy standard. Soo-Yong Shin [9] also develops a system which is named NACST based on genetic algorithm. L.Wenbin [10] proposes a method to optimize the template frame in DNA computing.

Although the research of DNA computing has obtained a lot of enormous progress in many fields, there are a number of problems which are not solved, such as DNA word sets which satisfy the distance constrains and thermodynamics constrains, biologic technology problem and so on. The problem which is urgent to solve is that how to combine the recognition of information specific in DNA computing with varied kind of biochemical reaction factors and build a standard of DNA coding. At present, the coding problem has been noticed by more and more researchers. However, there is no the better method to solve the coding problem. In the DNA coding constrains, the most common constrains are distance constrains and thermodynamics constrains. The distance constrains commonly includes Hamming distance constrain, shift distance constrain, h-distance and so on. The thermodynamics constrains commonly includes GC content, Gibbs energy ΔG, melting temperature T_M and so on [9].

1.1 The significance of researching DNA word sets

In the DNA coding, most researchers are to produce DNA sequences which satisfy the constraints, but they do not deeply research the DNA word sets which satisfy one or more constraints. That is the core point which we propose in this paper. Its significance can be briefly stated as follows:

（1）The problem of designing DNA word sets is to produce the DNA sequences which satisfy the constraints. So 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 improve the accuracy of DNA computing.

（3）According to the researchful 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.

1.2 Progress in DNA word sets

The main problem of DNA word sets is to research the bounds of DNA word sets. The ways obtained the bounds of DNA word sets include two main methods: one is theoretical derivation. It could obtain the structural method of DNA sequences which satisfy the constraints and the approximate upper or lower bounds, such as [11–13]. The other is to use the intelligent algorithm to search the DNA word sets which satisfy the constraints and obtain the improved lower bounds of DNA word sets, such as stochastic local search algorithm[14,15], hybrid randomized neighborhoods search algorithm [16], dynamic neighborhood search algorithm [17] and so on.

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 [18], 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 [11] and [12]. Applying these results, they could reduce the values range of DNA word sets. In [12], the authors deeply research the theoretical bounds which satisfy the HD. In [11], the authors deeply research the theoretical bounds which respectively satisfy the HD, HD and RC, HD and HR, and give the relation between them. In [18], the authors research the GC content, RC and GC content constraints. In [19], the authors use the method which combines linear construction with stochastic local search algorithm. They improve the some lower

bounds which satisfy the GC content, RC and GC content constraints.

In the research of intelligent algorithm, the main idea is to use intelligent algorithm to search DNA sequences sets which 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 [14], the authors use the stochastic local search algorithm to improve the lower bounds which satisfy the HD and RC combinatorial constraints. The results are compared with the theoretical value. At the same time, they also improve the bounds which satisfy the HD and RC constraints and obtain the approximate bounds which satisfy the RC and GC content constraints. In [16], the authors improve the stochastic local search algorithm and the results which are from the [14]. In [17], the authors use the dynamic neighborhood search to improve the lower bounds which satisfy the RC and GC content constraints.

In this paper, we put emphasis on the problem of DNA word sets which satisfy h-distance constraint and use dynamic genetic algorithm to improve lower bounds of DNA word sets. By comparing our experimental results with previous researchful work, our results improve the lower bounds and further shorten the value range of DNA word sets.

2 The H-Distance Constrains

Garzon firstly proposes the definition of DNA coding problem in DNA computing [5]. The definition is as follow: In the alphabet $\Sigma=\{\mathrm{A,G,C,T}\}$, it exists a set S with the length of n. The size of S is $|S|=4^n$. A subset of S, $C\subseteq S$ and let x_i, x_j any two codes in the C satisfy

$$\tau_{i\ j} \geq \qquad (1)$$

k is the positive integer and τ is the criterion of estimating the quality of coding, such as Hamming distance, shift distance and so on. The better quality of coding is, the more number of constraints, and then it must lead to decrease the number of DNA sequences which satisfy constraints.

2.1 The H-Distance Constrains

To cope with DNA coding problem, a much simple and computationally tractable model of the Gibbs energy landscapes given by the hybridization distance (namely h-distance), was introduced in [5] as a measure of hybridization affinity. Hybridization reaction, naturally dependents on many other reaction conditions such as temperature, salinity, and kinetic factors, are reduced to a single numerical threshold m. Hybridization occurs if and only if their h-distance does not exceed d. For example, under the tightest stringency $d = 0$, two strings can only hybridize when they are perfect complements, whereas under the most relaxed stringency $d = n$, any two strings will bind when they encounter each other. The h-distance constraint is defined as follows [20]:

$$h(x, y) = \min_{-(n-1)<k<n-1} \{|k| + H(x, \sigma^k(y^{wc}))\} \tag{2}$$

where, x, y are two DNA strands of a given length n (written from the 5'- to the 3'-end), σ^k is the (right-) left-shift by k positions (if $k < 0$, respectively.), y^{wc} is the Watson-Crick-complement of y obtained by reversing y and exchanging A-T's, C-G's and vice versa, and H is the ordinary Hamming distance. The h-distance considers hybridization in all possible frame-shifts, which is more realistically restrictive than simple models in which hybridization is considered only in the perfect alignment. Measure 0 indicates perfect complementarities. A large measure indicates that even when x finds itself in the proximity of y, they contain few complementary base pairs, and are less likely to hybridize.

2.2 The Bounds of the H-Distance constraint

According to the definition above, V.Phan provided lower and upper bounds of DNA codeword sets which satisfy the h-distance constraint [20].

$$\frac{4^{n-d+1}}{d\binom{n}{d-1}} \le |S| \le V \tag{3}$$

where $V = 4^n$, the number of possible sequence sets of size $t-1$ that can be formed with n code sequence is:

$$\binom{n}{d-1} = \frac{n!}{(d-1)!(n-d+1)!} \tag{4}$$

Depending on the above inequation (3), we calculate the size of codes that satisfy the

h-distance constraint as a function of sequence sets length n and threshold m. Table 1 is the lower bounds of DNA word sets which satisfy the h-distance constraint as a function of word length n and h-distance d. These bounds result from the V.Phan's paper [20]. There do not exist values in the blank which have '-'.

Table1. Theoretical bounds on the size of codes that satisfy h-distance constraint [20]

n / d	3	4	5	6	7	8
3	1	-	-	-	-	-
4	1	1	-	-	-	-
5	3	1	1	-	-	-
6	6	1	1	1	-	-
7	17	2	1	1	1	-
8	25	4	1	1	1	1

3 Dynamic Genetic Algorithm

In this paper, we use dynamic genetic algorithm to improve lower bounds of DNA word sets which satisfy combinatorial constraints. Genetic algorithm (for short GA) is stochastic search algorithm based on nature selection and genetic mechanism. GA could solve a number of problems, because the process of nature evolution is a process of learning and optimizing. The main idea of this algorithm is that: the process of nature evolution (from simple to complex, from low class to upper class) is natural and parallel; the intention is to adapt to environment. The biological populations begin to evolve by survival of the fittest and genetic variation. Genetic algorithm carries out the evolution of biology by selection, crossover and mutation [21,22,25].

The dynamic genetic algorithm can conquer the premature problem of genetic algorithm and be used to improve lower bounds of DNA word sets which satisfy the constraints. We control the evolution by controlling the fitness function. The improved areas can be briefly stated as follows:

(1) Initializing the populations of algorithm with evenly distributed method.

(2) In the mutation process, we adjust the probability of mutation operator with dynamic method.

The main process is that: initializing DNA sequences with evenly distributed method, selecting the sequences which satisfy the Hamming distance and GC content constraints from these sequences, generating new DNA sequences by selection, crossover and mutation operator, last obtaining the DNA word sets. Fig.1 is the flowchart of algorithm.

The steps of solving the sequence design by Dynamic Genetic algorithm are as follows:

Step1: Setting parameters and initializing population with evenly distributed method.

Step2: Calculating the value of fitness function.

Step3: Generating next population by selection、crossover and mutation. In the selection process, the algorithm use randomly selecting strategy. In the mutation process, If one of fitness value satisfies $F(i) \le d-3$, where $F(i)$ is the fitness value, d is Hamming distance, it will be generated again. And if $F(i) = d-2$, its probability of mutation is 0.2. Else, its probability of mutation is 0.01. If the size of $F(i)$ is more than 50, go to step2; if not then go to step4.

Step4: end and output results.

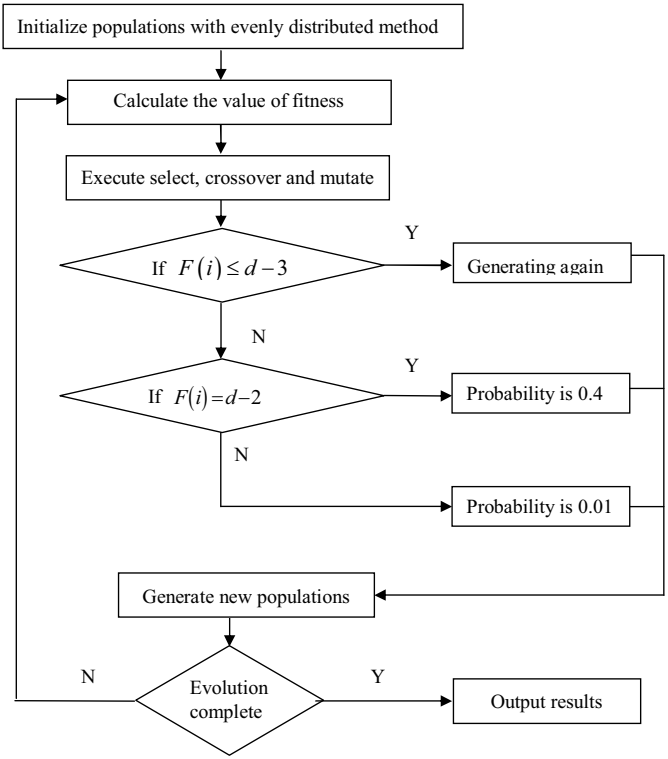

Figure1. The flowchart of algorithm

4 DNA Codeword Sets Obtained by Dynamic Genetic Algorithm

The parameters of dynamic GA used in our example are: the size of population is 1000. The length of DNA sequences is n. The probability of crossover is 0.45. The probability of mutate which is initialized is 0.05. In order to control the time of running algorithm, the generation is less than 1000, where n is the length of DNA codeword and d is the h-distance threshold.

Table 2 shows the maximal lower bounds of DNA codeword sets obtained by the dynamic genetic algorithm when we performed ten trials for every value. In this table, entries in bold face exceed the theoretical lower bounds from Table 1. The values are the minimal generation of dynamic genetic algorithm in the bracket. There do not exist values in the blank which have '-'.

Comparing Table 1 with Table 2, we could find that the results of Table 2 improve the lower bounds which satisfy the h-distance constraint, and further shorten the value range of the bounds of DNA codeword sets. From the above empirical results, we could obtain the shortest DNA sequences which meet the h-distance constraint, when we have known the scale of the practical problem. We could use these sequences to deal with the NP problem, such as using in Adleman's experimentation, reduce the redundancy degree of storing information and improve the veracity and efficiency of DNA computing.

Table 2. The lower bounds of DNA codeword sets from our algorithm

n \ d	3	4	5	6	7	8
3	**8**(12)	-	-	-	-	-
4	**26**(36)	**16**(10)	-	-	-	-
5	**144**(28)	**59**(37)	**32**(2)	-	-	-
6	**436**（50）	**265**（19）	**96**（15）	**62**(10)	-	-
7	**903**（2）	**566**（15）	**261**（36）	**155**（6）	**78**（2）	-
8	**652**（10）	**499**（15）	**403**（23）	**324**（1）	**204**(5)	**194**(5)

5 Conclusions

The design of DNA sequence sets, or sets of short DNA strands that satisfy h-distance constraint, is motivated by the tasks of storing information in DNA strands that are used for computation or as molecular bar codes in chemical libraries. Design of DNA codeword sets could produce enough good DNA sequences, which is important in order to minimize errors due to non-specific hybridization between distinct sequence and their complements, and also important to obtain a higher information density, and larger sets of sequences for large-scale application. In this paper, we use dynamic genetic algorithm to design DNA codeword sets

that satisfy the h-distance constraint. By comparing Table 1 with Table 2, it could prove the efficiency of dynamic genetic algorithm, improve the lower bounds of DNA word sets and further shorten the value range of DNA coding bounds. Furthermore, in the field of theoretical research of DNA sequence sets, the constructive method of DNA sequences that satisfy constraints is not proposed, and it is difficult to be used in practical DNA computing. By using dynamic genetic algorithm, DNA sequences that satisfy constraints could be produced. These sequences could decrease the emergence of false negative and false positive, and improve the efficiency and reliability of DNA computing. However, we have some further work need to do. Such as, if we increase the number of initializing populations, there maybe obtain better results. In future work, we will improve our algorithm and combine the more constraints.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (Grant No. 30870573), by the National High Technology Research and Development Program ("863"Program) of China (No.2009AA01Z416), and by the open fund of Key Laboratory of Advanced Design and Intelligent Computing (Dalian University), Ministry of Education, Dalian University (No. ADIC2010012).

Reference

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

[2] E. B. Baum, DNA sequences useful for computation, *Proc. 2nd DIMACS Workshop DNA Based Comput.*, 1996, pp. 122–127.

[3] R. Deaton, R. C. Murphy, J. A. Rose, M. Garzon, D. R. Franceschetti, S. E. Stevens, A DNA based implementation of an evolutionary search for good encodings for DNA computation, *Proc. 1997 IEEE Int. Conf. Evolution. Comput.*, 1997, pp. 267–272.

[4] R. Deaton, D. R. Franceschetti, M. Garzon, J. A. Rose, R. C. Murphy, S. E. Stevens, Information transfer through hybridization reaction in DNA based computing, *Proc. Second Annual Conf.*, California, 1997, pp. 463–471.

[5] M. Garzon, R. Deaton, P. Neathery, R. C. Murphy, S. E. Stevens, D. R. Franceschetti, A new metric for DNA computing, *Proc. 2nd Annual Genetic Prog. Conf. GP-97*, Morgan Kaufmann, Stanford Univ., 1997, pp. 472–487.

[6] D. H. Wood, Applying error correcting codes to DNA computing, *4th DIMACS Workshop on DNA Based Computers, Pennsylvania*, 1998, pp. 109–110.

[7] A. Hartemink, E. J. 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.

[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] S. Y. Shin, D. M. Kim, I. H. Lee, B. T. Zhang, Evolutionary sequence generation for reliable DNA computing, *Proc. Conf. Comput. Intel.*, WCCI, 2002, pp. 79–84.

[10] W. B. Liu, L. C. Chen, B. G. Bai, Research on optimizing the template frame in DNA computing, *Acta Electron. Sinica* **35** (2007) 1490–1494.

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

[12] G. T. Bogdanova, A. E. Brouwer, S. N. Kapralov, P. R. J. Ostergard, Error–correcting codes over an alphabet of four elements, *Design. Code Cryptogr.* **23** (2001) 333–342.

[13] B. Paolo, A. Labella, V. Manca, V. Mitrana, Superposition based on Watson–Crick–like complementarity, *Theor. Comput. Sys.* **39** (2006) 503–524.

[14] D. Tulpan, H. Hoos, A. Condon, Stochastic local search algorithms for DNA word design, *Proc. 8th DNA Based Comput.*, 2002, LNCS 2568, pp. 229–241.

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

[16] D. Tulpan, H. Hoos, Hybrid randomized neighborhoods improve stochastic local search for DNA code design, *Proc. Adv. Artif. Intell.*, 2003, LNCS 2671, pp. 418–433.

[17] S. Kawashimo, H. Ono, K. Sadakane, M. Yamashita, DNA sequence design by dynamic neighborhood searches, in: C. Mao, T. Yokomori (Eds.), *DNA Computing*, Lecture Notes in Computer Science 4287, Springer–Verlag, Berlin, 2006, pp. 157–171.

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

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

[20] V. Phan, M. H. Garzon, The capacity of DNA for information encoding, DNA10, LNCS 3384, 2005, pp. 281–292.

[21] W. Wang, X. D. Zheng, Q. Zhang, J. Xu, The optimization of DNA encodings based on GA/SA algorithms, *Progress Natural Sci.* **17** (2007) 739–744.

[22] Q. Zhang, B. Wang, X. P. Wei, X. Y. Fang, C. J. Zhou, DNA word set design based on minimum free energy, *IEEE Trans. Nanobiosci*, accepted.

[23] D. Tulpan, M. Andronescu, S. Leger, Free energy estimation of short DNA duplex hybridizations, *BMC Bioinform.* **11** (2010) 105: 1–22.

[24] I. S. Jeong, K. W. Park, S. H. Kang, H. S. Lim, An efficient similarity search based on indexing in large DNA databases, *Comput. Biol. Chem.* **34** (2010) 131–136.

[25] Q. Zhang, B. Wang, X. Wei, Evaluating the different combinatorial constraints in DNA computing based on minimum free energy, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 291–308.

[26] V. Phan, M. H. Garzon, On codeword design in metric DNA spaces, *Natur. Comput.* **8** (2009) 571–588.

[27] T. Nakano, J. Q. Liu, Design and analysis of molecular relay channels: An information theoretic approach, *IEEE Trans. Nanobiosci.* **9** (2010) 213–221.

[28] S. Kawashimo, Y. K. Ng, H. Ono, K. Sadakane, M. Yamashita, Speeding up local–search type algorithms for designing DNA sequences under thermodynamical constraints, DNA 14, LNCS 5347, 2009, pp. 168–178.

[29] Z. J. Zhang, DV–curve: A novel intuitive tool for visualizing and analyzing DNA sequences, *Bioinform.* **25** (2009) 1112–1117.

[30] W. W. M. Lam, K. C. C. Chan, Discovering interesting molecular substructures for molecular classification, *IEEE Trans. Nanobiosci.* **9** (2010) 77–89.