Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/070649

A Dynamic Programming Algorithm for Circular Single-stranded DNA Tiles Secondary Structure Prediction

Zhang Kai[1,2,*], *Huang Xinquan*[1,2], *Shi Xiaolong*[3], *Qiang Xiaoli*[4], *Song Tao*[3,*], *Shi Xinzhu*[1], *Chen Zhihua*[3]

[1]School of Computer Science, Wuhan University of Science and Technology, Wuhan 430081, P. R. China
[2]Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430081, P. R. China
[3]Department of Control Science and Engineering, Huazhong University of Science and Technology , Wuhan 430074, P. R. China
[4]College of Computer Science, South-Central University for Nationalities , Wuhan 430074, P. R. China

Received: 17 Apr. 2013, Revised: 18 Aug. 2013, Accepted: 21 Aug. 2013
Published online: 1 Nov. 2013

Abstract: The design of DNA sequences is critical for many research fields such as DNA self-assembly, DNA hybridization arrays, DNA computing, and PCR-based applications. DNA secondary structure prediction is the key part for these DNA nanotechnologies. In this paper, we present a dynamic programming algorithm to predict the secondary structure of single-stranded DNA tiles. The algorithm calculates all possible maximum matches based on the nearest-neighbour model and global energy minimization. Experimental results show that the algorithm performers significantly to predict secondary structures for single-stranded DNA tiles.

Keywords: Bio-computing, Secondary structure prediction, DNA self-assembly, Dynamic programming.

1 Introduction

Bio-computing is a new branch of natural computing, whose aim is to abstract ideas from the functioning of DNA molecular, RNA molecular or cell membrane to construct computing models. In recent years, many theoretical and practical bio-computing models have been developed, such as DNA computing [1-3], DNA self-assembly [4-6], membrane computing [7-9], spiking neural P systems [10-14]. Among them, DNA self-assembly have arisen widely research interests all over the world. With novel mechanical and chemical function, DNA molecules have shown great potential as a design medium for the construction of nano-scale devices [15, 16], and widely used in many laboratories such as assembly [17-19], switching [20-22], circuitry [23, 24], DNA chips [25, 26]. Among all approaches to design stable DNA sequences, computational algorithms play an important role [15, 27].

In these research areas, it is very important to select DNA sequences with required secondary structure, since high quality DNA sequences can prevent the interference between different hybridization, and improve the reliability and effectiveness of experimental. DNA folding algorithm can be regarded as an adjunct to digitally represent DNA sequences, but not a replacement of reality physical data. The designed DNA sequences should be theoretically rapidly assessed before attempting laboratory validation. So, we need to design effective search algorithms to identify and predict rational novel nucleic acid structures for selecting promising sequences.

Recent years several dynamic programming algorithms have been proposed to predict secondary structures of DNA or RNA sequences. Zuker [28] proposed an algorithm for prediction of RNA secondary structure. The time complexities of the algorithms is $O(n^4)$. Andronescu et al. [29] developed the PairFold algorithm for secondary structure prediction of minimum free energy. Pervouchine [30], Alkan et al. [31] and Kato et al. [32] presented different dynamic programming algorithms with different scoring functions for predicting secondary structures, respectively. However, these algorithms cannot deal with circular single-stranded DNA tiles.

* Corresponding author e-mail: zhangkai@wust.edu.cn, songtao0608@hotmail.com

In order to avoid this defect, we present a dynamic programming algorithm to predict the secondary structure of single-stranded DNA tiles. The algorithm calculates all possible maximum matching of single strands DNA tile based on the nearest-neighbor model and global energy minimizations. Experimental results show that the algorithm performs significantly to predict structures for single-stranded DNA tiles.

2 The Algorithm of Circular Single-Strand DNA Tiles Secondary Structure Prediction

In this section, we will present a dynamic programming algorithm for predicting single-strand DNA tiles secondary structures. Before going through the details of the algorithms, let us begin with definitions of DNA secondary structure and the prediction problem.

In biological experiment, single strand DNA molecules dismissed randomly in vitro. A single stranded DNA molecule is an unbranched polymer composed of only four types of subunits. These subunits are the deoxyribonucleotides containing the bases adenine (A), cytosine (C), guanine (G) and thymine (T). The nucleotides composing DNA bind to each other in pairs via hydrogen bonds in a process known as hybridization. Each nucleotide pairs up with its unique complement (the Watson-Crick complement), so C pairs up with G and A with T.

A DNA secondary structure for a given DNA sequence $X = 5' - x_1x_2\cdots x_n - 3'$ of length n is defined to be a set S of ordered pairs (i, j), with $1 \le i \le j \le n$, such that the following conditions are satisfied:

(1) **Watson-Crick Constraint**:

If $(i, j) \in S$, then $\{x_i, x_j\} \in \{\{A,T\},\{G,C\}\}$;

(2) **No base triples Constraint**:

If (i, j) and (i,k) belong to S, then $j = k$; if (i, j) and (k, j) belong to S, then $i = k$;

(3) **Threshold requirement for hairpins**:

If (i, j) belongs to S, then $j - i > \theta$, for a fixed value $\theta \ge 0$; i.e. there must be at least unpaired bases in a hairpin loop.

2.1 Dynamic Programming Algorithms

It is assumed that a structure with more number of base-pairs is more stable than that with fewer base-pairs, so we define the score of a secondary structure simply as the number of base-pairs in sequence. The optimal DNA secondary structure problem can be formally formulated as follows.

Given a DNA string X of length n, compute a secondary structure S with maximum number of base-pairs. We need to compute the optimal secondary structure score according to the scoring function, denoted by $M(i, j)$, for every substring $x[i\cdots j], 1 \le i < j \le$. Since the structures have recursive sub-structures, we should consider all of the conditions in which x_i and x_j can form base-pairs (or not). As in the case of alignment, there are following several possibilities.

(1) Nucleotide x_j does not form a base-pair. In this case, we have $M(i, j) = M(i, j-1)$, as shown in Fig.1.

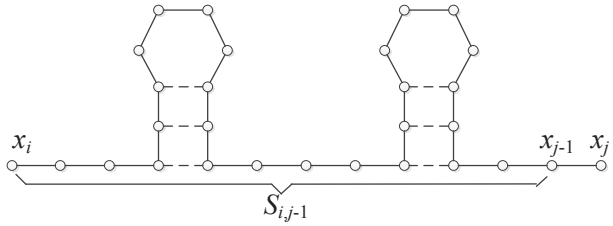

Fig. 1: x_j cannot form a Watson-Crick complement with any x_i

(2) Nucleotide x_i and x_j are complementary, and pair with each other. Consequently, we get $M(i, j) = 1 + M(i+1, j-1)$, as shown in Fig.2.

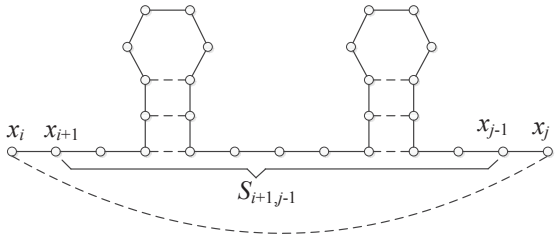

Fig. 2: x_i and x_j are Watson-Crick complementary

(3) If there exists $i < k < j$ such that x_k and x_j are complementary and can pair up, then all base-pairs in a secondary structure of $x[i\cdots j]$ must have the property that either $x[i\cdots k-1]$ or $x[k+1\cdots j-1]$ presents. In this case, then $M(i, j) = 1 + M(i,k-1) + M(k+1, j-1), i < k < j$.

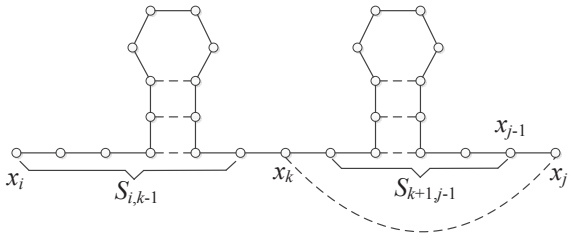

Fig. 3: x_k and x_i are Watson-Crick complementary

There are three possible cases, but only the cast that optimal structure $S(i,j)$ holds the maximum value is considered. The optimal score $M(i,j)$ is the number of base-pairs of $S(i,j)$. Because $M(i,j)$ is the optimal scores of smaller subsequences, in this way we need to record these scores, but not the combinatorial explosion of possible structures. Mathematically this recursion is as follows:

$$M_{i,j} = \max \begin{cases} M_{i,j-1} \\ M_{i,j-1} + \rho(x_i, x_j) \\ \max\limits_{i+1 \le k \le j-4} \{M_{i,k-1} + M_{k+1,j} + \rho(x_i, x_j)\} \end{cases}$$

$$\rho(x_i, x_j) = \begin{cases} 1, \ if\{x_i, x_j\} \in \{\{\mathrm{A},\mathrm{T}\}\{\mathrm{G},\mathrm{C}\}\} \\ 0, \ else \end{cases}$$

2.2 Free Energy Constraint and Nearest-Neighbor Model

After a group of admissible structures have been defined, we need to enumerate all the structures that can be formed with n nucleotides. Subsequently, we choose the most stable structure which has the minimum free energy. (Free energy is a measure of DNA double stranded stability. Since DNA hybridization usually emits heat, free energy changes are usually negative that is $\Delta G < 0$.) It is easy to obtain that the higher the absolute value is, the more stable DNA double stranded will be. The algorithm based on free energy used the nearest-neighbor thermodynamic model, the formula is as follows:

$$\Delta G = \sum_i n_i \Delta G(i) + \Delta G(ini\,\mathrm{GC}) + \Delta G(ini\,\mathrm{AT}) + \Delta G(sym)$$

where $\Delta G(i)$ denotes the standard free energy changes for the 10 possible Watson-crick NNs (e.g., $\Delta G(1) = \Delta G(\mathrm{AA/TT})$, $\Delta G(2) = \Delta G(\mathrm{TA/AT})$, $\cdots$ etc.), n_i is the number of occurrences of each nearest neighbor, and $\Delta G(sym)$ equals to +0.43 kcal/mol (if the duplex is self-complementary) or equals to 0 (if it is non-self-complementary). The primary energetic factor for hybridization is not the energy of the hydrogen bonding between nucleotide bases, but is the nearest neighbor base stacking energies. These base stacking energies must be measured, and are not unique. Nevertheless, from an energetic point of view, they are the parameters of choice to determine the potential for hybridization between oligonucleotides.

2.3 Circular DNA Secondary Structure Prediction

A circular DNA molecule consists of a chain of nucleotides linked together, which are not only some DNA tiles computation model, but also various small viruses consist of single-stranded circular DNA. The starting and ending nucleotides are linked together. Given an arbitrary point, a secondary structure can be defined as in Section 2, except that the starting and ending bases cannot pair up with each other because they are adjacent in the circular sequence.

Our algorithm starts with a target circular DNA sequence. We calculate the maximum base-pair matching by our dynamic programming algorithm, and all the sequence loop will be taken into account. Subsequently, we evaluate and predict the DNA secondary structure by nearest neighbor model and free energy minimization. The pseudo program code of our algorithm is given as bellow:

program Dynamic Programming Algorithm for Circular DNA Tiles Secondary Structure Prediction
input : Single Strand DNA Sequence X
output : Secondary Structure with Minimum Free Energy
begin
for $i = 1$ **to** n **then**
begin
generate new sequence X' with circular X loop
$X' = 5'-x_2x_3\ldots x_{n-1}x_nx_1-3'$
Calculate the maximum matching structure S of X' using Dynamic Programming
Calculate the free energy E of S using Nearest Neighbor thermodynamic model
if $E < MinE$ **then**
begin
$MinE = E$, keep the minimum free energy
$MinS = S$, keep the corresponding structure
end
end
Return most stable structure $MinS$;
end.

2.4 Example Results

To demonstrate the applicability of our algorithm, we select a DNA sequence to test our algorithm by analyzing its secondary structures and free energy. The sequence is 5′-CCGAACGCCGCAAAAAGCGGCGAACGCGCGA AAAACGCGCGAACGGGCGCGCCCGAACGGCGCA AAAAGCG-3′ with length 71. There are 71 kinds of candidate structures, and we demonstrate three candidate prediction results using our algorithm. The prediction result is shown in Fig.4, which shows that from position 0 the maximum match is 22 and free energy -32.31. Another result is from position 27 with maximum match 22 and free energy -30.14, which is shown in Fig.5.

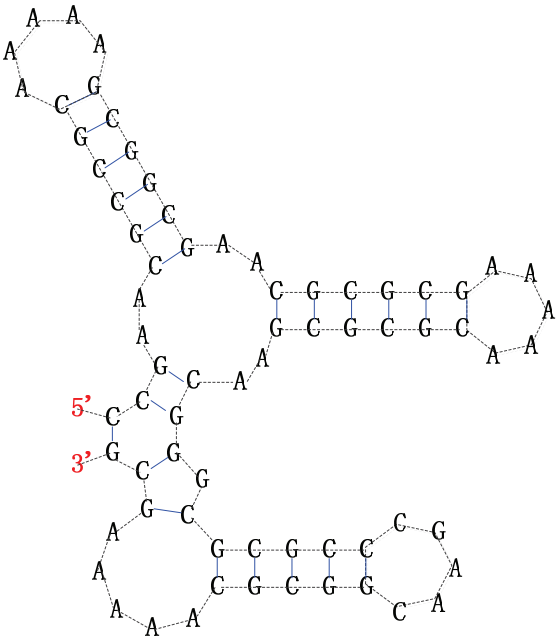

Fig. 4: Calculate from position 0, maximum matching number 22, stack energy is -32.31.

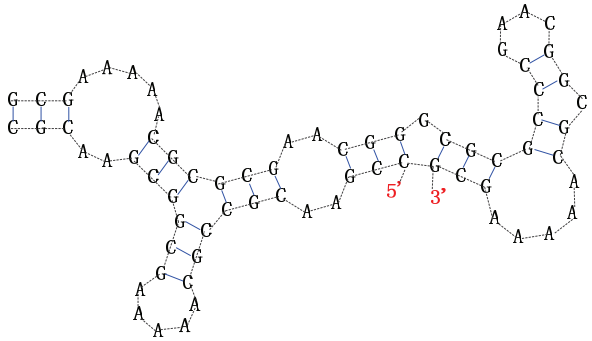

Fig. 5: Calculate from position 27, maximum matching number 22, stack energy is -30.14.

The optimal result is given in Fig.6. It shows the structure from position 49 has maximum match 24 and free energy is -42.17. Among 71 kinds of candidate structures, this candidate solution shows much lower absolute value of free energy ΔG, so this DNA molecule can form a more stable secondary structure.

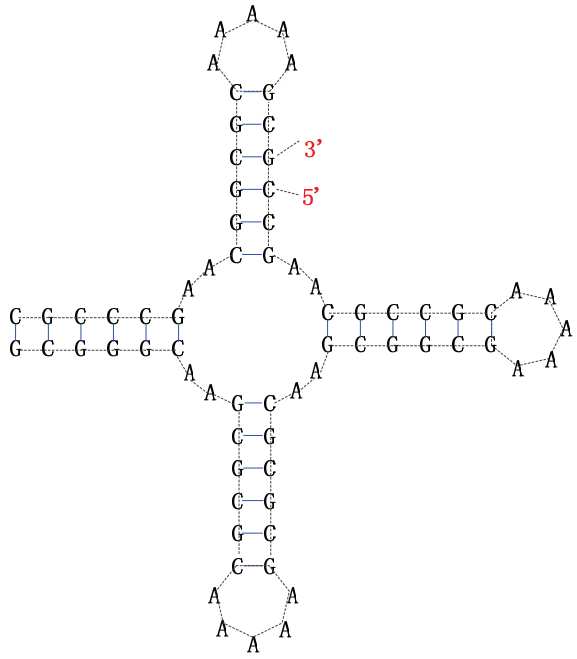

Fig. 6: Calculate from position 49, maximum matching number 24, stack energy is -42.17.

3 Conclusion

In this work, a dynamic programming algorithm for circular DNA tiles structure prediction has been proposed. The nearest-neighbor thermodynamic model is integrated which can select the most stable structure with energy minimization. The time complexities of traditional dynamic programming algorithms is $O(n^4)$, but the time complexities of our algorithm is $O(n^5)$ by taking n start points into account.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61100055, 61033003, 60974112, 60975031, 60975031, 31201121 and 91130034), Natural Science Foundation of Hubei Province (Grant No. 2011CDB233), Innovation Fund of Huazhong University of Science and Technology (Grant No. 2011TS005).

References

[1] Adelman L. M., Molecular computation of solutions to combinatorial problems. Science, **266**, 1021-1024 (1994).
[2] Faulhammer D., Cukras A. R., Lipton R. J., Molecular computation : RNA solutions to chess problems. Proceedings of the National Academy of Sciences, **97**, 1385-1389 (2000).

[3] Braich R. S., Chelyapov N., Johnson C., Rothemund P.W.K. and Adleman L., Solution of a 20-variable 3-SAT problem on a DNA computer. Science, **296**, 499-502 (2002).
[4] Winfree E., Liu F., Wenzler L. A., Design and self-assembly of two-dimensional DNA crystals . Nature, **394**, 539-544 (1998).
[5] LaBean T. H., Yan H., Kopatsch J., Construction, analysis, ligation, and self - assellbly of DNA triple crossover complexes. Joumal of the American Chemical Society, **122**, 1848-1860 (2000).
[6] Mao C. , LaBean T. H. , Reif J. H., Logical computation using algorithmic self-assellbly of DNA triple-crossover molecules . Nature, **407**, 493-496 (2000).
[7] Zhang X., Wang J., Pan L., A Note on the Generative Power of AxonSystems. Int. J. Comput. Commun., **4**, 92–98 (2009)
[8] Zhang X., Zeng X., Pan L., On String Languages Generated by Asynchronous Spiking Neural P Systems. Theor. Comput. Sci., **410**, 2478-2488 (2009)
[9] Zhang X., Jiang Y., Pan L., Small Universal Spiking Neural P Systems with Exhaustive Use of Rules. Int. J. Comput. Commun., **7**, 1-10 (2010)
[10] Pan L., Daniel D. P., Prez-Jimnez M. J., Computation of Ramsey Numbers by P System with Active Membranes. Int. J. Found. Comput. Sci., **22**, 29-38 (2011)
[11] Pan L., Paun G., Perez-Jimenez M. J., Spiking Neural P Systems with Neuron Division and Budding. SCIENCE CHINA Information Sciences, **54**, 1596-1607 (2011)
[12] Pan L., Zeng X., Zhang X., Time-Free Spiking Neural P Systems. Neural Computation, **23**, 1-23 (2011)
[13] Pan L, Zeng X, Zhang X, Jiang Y., Spiking Neural P Systems with Weighted Synapses. Neural Processing Letters, **35**, 13-27 (2012)
[14] Pan L., Wang J., Hoogeboom H. J., Spiking Neural P Systems with Astrocytes, Neural Computation, **24**, 805-825 (2012)
[15] Seeman N. C., Nucleic Acid Junctions and Lattices. J. Theor. Biol., **99**, 237-247 (1982)
[16] Seeman N. C., DNA Engineering and Its Application to Nanotechnology. Trends. Biotechnol, **17**, 437-443 (1999)
[17] Chen J., Seeman N. C., The Synthesis from DNAs of a Molecule with the Connectivity of a Cube. Nature, **350**, 631-633 (1991)
[18] Kallenbach R. K., MaR I., Seeman N. C., An Immobile Nucleic Acid Junction Constructed from Oligonucleotides. Nature, **305**, 829-831 (1983)
[19] LaBean T. H., Yan H., Kopatsch J., Liu F., Winfree E., Reif J. H., Seeman N. C., Construction, Analysis, Ligation and Self-assembly of DNA Triple Crossover Complexes. J. Am. Chem. Soc, **122**, 1848-1869 (2000)
[20] Soukup G. A., Breaker R. R., Engineering Precision RNA Molecular Switches. Proc. Natl Acad. Sci., **96**, 3584-3589 (1999)
[21] Winfree E., Liu F., Wenzler L. A., Seeman N. C., Design and Self-assembly of Two-Dimensional DNA Crystals. Nature, **394**, 539-544 (1998)
[22] Yan H., Zhang X., Shen Z., Seeman N.C., A Robust DNA Mechanical Device Controlled by Hybridization Topology. Nature, **415**, 62-65 (2002)
[23] Stojanovic M. N., Stefanovic D., A Deoxyribozyme-based Molecular Automaton. Nat. Biotechnol., **21**, 1069-1074 (2003)
[24] Yurke B., Turberfield A. J., Mills A. P., Jr Simmel F. C., Neumann J. L., A DNA-fuelled Molecular Machine Made of DNA. Nature, **406**, 605-608 (2000)
[25] Brenner S., Johnson M., Bridgham J., Golda G., Lloyd D.H., Johnson D., Luo S., McCurdy S., Foy M., Ewan M., Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS) on Microbead Arrays. Nat. Biotechnol., **18**, 630-634 (2000)
[26] Shoemaker D. D., Lashkari D. A., Morris D., Mittman M., Davis R. W., Quantitative Phenotypic Analysis of Yeast Deletion Mutants using a Highly Parallel Molecular Bar-coding Strategy. Nature. Genet., **16**, 450-456 (1996)
[27] Seeman N. C., Kallenbach R. K., Design of Immobile Nucleic Acid Junctions. Biophys. J., **44**, 201-209 (1983)
[28] Zuker M., Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., **9**, 133-148 (1981).
[29] Andronescu M., Zhang Z. C., Condon A., Secondary Structure Prediction of Interacting RNA Molecules. J. Mol. Biol., **4**, 987-1001 (2005)
[30] Pervouchine D. D., IRIS, Intermolecular RNA Interaction Search. Genome Inform., **15**, 92-101 (2004)
[31] Alkan C., Karako E., Nadeau J. H., Sahinalp S. C., Zhang K., RNA-RNA Interaction Prediction and Antisense RNA Target Search. J. Comput. Biol., **13**, 267-82 (2006)
[32] Kato Y., Akutsu T., Seki H., A Grammatical Approach to RNA-RNA Interaction Prediction, Pattern Recognition, **42**, 531-538 (2009)

Zhang Kai received the Ph.D. degree from Huazhong University of Science and Technology in 2008. He was on the Post Doctor research of the School of Electronics Engineering and Computer Science at Peking University from 2008 to 2010. Currently, He is an associate professor of the College of Computer Science and Technology at Wuhan University of Science and Technology. His research interests include combinatorial optimization, graph theory, molecular computation and intelligent computing.

Huang Xinquan is a postgraduate student in Wuhan University of Science and Technology. His current research interests include DNA computing and operational research.

Shi Xiaolong is an associate professor of Department of Control Science and Engineering at Huazhong University of Science and Technology. He received his Ph.D. degree from Huazhong University of Science and Technology in 2003. His major research interests include image processing, neural network, pattern recognition and bioinformatics.

Qiang Xiaoli received the Ph.D. degree from Huazhong University of Science and Technology in 2008. She is an associated professor of computer science at South-Central University for Nationalities. Her research interests include bio-computing, bio-informatics.

Song Tao received his docter Degree on system analysis and intergration in 2013 from Huazhong University of Science and Technology. His research interests include DNA computing, DNA encoding and membrane computing.

Shi Xinzhu received her master degree from Wuhan University in 2009. Her research interests include DNA nanotechnology and DNA-Based Computation.

Chen Zhihua received the Ph.D. degree in systems analysis and integration from Huazhong University of Science and Technology in 2009. She is currently an Associated Professor at the Department of Control Science and Engineering, Huazhong University of Science and Technology. Her research interests include DNA computing, information security, and optimization algorithms.