DNA Secondary Structures for Probe Design*

Yanga Byun and Kyungsook Han

School of Computer Science and Engineering, Inha University, Inchon 402-751, Korea
quaah@hanmail.net, khan@inha.ac.kr

Abstract. Visualizing DNA secondary structures is essential to fast and efficient design of probes for DNA chips. There are several programs available for visualizing single-stranded RNA secondary structures, but these programs cannot be used to draw DNA secondary structures formed by several hundred to thousand primers and target genes. We have developed an algorithm and program for visualizing DNA secondary structures formed by multiple strands. We believe the program will be a valuable tool for designing primers and probes in DNA chips.

1 Introduction

The high-throughput analysis of genes using DNA chips has a great impact on modern biological research. Several thousand different primers are required for a DNA chip [1]. DNA primers are DNA sequence fragments consisting of 4 types of nucleotides: adenine (A), guanine (G), cytosine (C), and tymine (T). Target genes and primers form secondary structures by hydrogen bonds between complementary base pairs. The secondary structure is an important criterion for the selection of the primer since interaction between primers should be avoided to conserve the maximum sensitivity of the primer and the spot on a DNA chip.

Several programs are available for drawing RNA secondary structures [2–4], but none of these can be used to draw DNA secondary structures because the programs are intended for drawing single-stranded RNA. DNA itself is a double-stranded molecule and primer design should be able to consider secondary structures formed by multiple primers and target genes. We have developed a program called DNAdraw for fast and accurate selection of primers to be used in DNA chips. The input for the program is the DNA or cDNA sequence(s) of the target gene, candidate primer sequences, and their secondary structures. Experimental results demonstrate that DNAdraw is capable of automatically producing a clear and aesthetically appealing drawing of DNA secondary structures. This paper describes an algorithm and its implementation.

2 Algorithm

In the structure data, a pair of parentheses represents a base pair. The parenthesis pairs used in DNAdraw are '()', '[]', and '{}'. In visualizing DNA structure, we

* This work was supported by the Korea Science and Engineering Foundation (KOSEF) under grant R01-2003-000-10461-0.

J. Pach (Ed.): GD 2004, LNCS 3383, pp. 504–507, 2004.

Fig. 1. (A) Example of a structure with 2 simple stem-loops (SL1 and SL2) and a composite stem-loops (SL3). (B) Tree structure of Fig. 1A.

call a structure element enclosed by matching parentheses a *stem-loop* (Fig. 1A). A *simple stem-loop* corresponds to a single hairpin loop-stem, and a *composite stem-loop* contains one or more other stem-loops. From the standpoint of graph theory, a drawing of DNA secondary structures can be considered as a tree with simple stem-loops as leaf nodes of the tree (Fig. 1B). Computation starts with a leaf node.

The algorithm of DNAdraw is outlined as follows: (1) stem-loops are identified from the input structure data; (2) the position and shape of a simple stem-loop are computed; and (3) the position and shape of a composite stem-loop are computed.

Base pairs of a stem in a simple stem-loop are stacked on the y-axis. In Fig. 2A, n represents the number of bases in the loop region plus 2 (for the base pair at the end of a stem). If the loop region contains a terminal base either at $5'$ or $3'$ end of a strand, 10 is added to n to make space between the base and other parts. L represents the distance between adjacent bases of a loop. L is also the distance between a pair of bases of a stem. Then, the angle a is π/n and the radius R of the loop is $L/2sin(a)$.

To determine the loop center, we first compute the midpoint P_m of points P_1 and P_2. If we use N to represent the unit vector directed toward the loop center C from a point P_m, vector N can be obtained by rotating the vector $P_2 - P_1$ clockwise with respect to P_m and then by normalizing the vector. The distance d between C and P_m is determined by equation (1). From the distance d, vector N, and the position vector P_m, we can compute the position vector C representing the loop center.

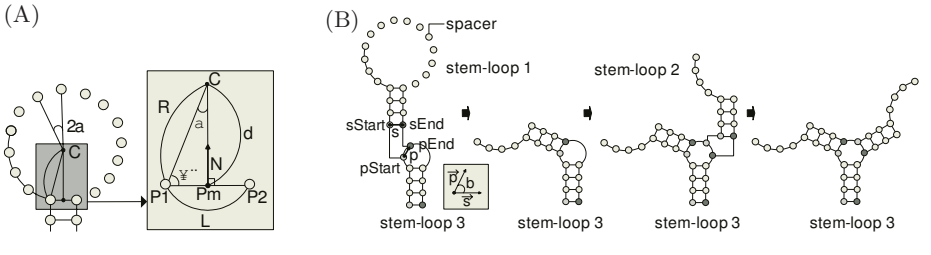

Fig. 2. (A) The radius, angle, and center of a loop in a simple stem-loop. (B) Example of inserting two simple stem-loops into a composite stem-loop by rotating and translation the simple stem-loops.

$$\theta = \pi/2 - a, \quad d = \|\overrightarrow{C} - \overrightarrow{P_m}\| = R \cdot \sin(\theta), \quad \overrightarrow{C} = d \cdot \overrightarrow{N} + \overrightarrow{P_m} \tag{1}$$

Consider a composite stem-loop pSL containing a simple stem-loop sSL. In Fig. 2B we use $sStart$ and $sEnd$ to represent the position of the first and the last base of sSL before being enclosed in pSL; $pStart$ and $pEnd$ to represent the position at which the first and the last base of sSL to be located in pSL. Let s be the unit vector in the direction of $sEnd - sStart$ and p the unit vector in the direction of $pEnd - pStart$. The simple stem-loop sSL can be inserted into the composite stem-loop pSL by rotating sSL by the angle between s and p with respect to $sStart$ and then translating it by the vector of $pEnd - sStart$. Fig. 2B shows an example of enclosing two simple stem-loops in a composite stem-loop.

3 Results and Discussion

DNAdraw is written in Microsoft Visual C#, and is executable on any Windows system (see Figs. 3 and 4). DNAdraw takes as input the DNA sequence with its structure data in bracket view. In the input data below, 5 and 3 denote the start and termination of the DNA sequence, respectively.

Input Format 1: Both ends of the DNA sequence are denoted either by 5 or 3.

```
# primer1  // optional sequence name
3-ATGCCGTAGGTA-5
5-TAGGTGAGCCAT-3
3-CTCAGCATTGCA-5
3-(((((((((::::-5
5-))))((((::::-3
3-)))))))):::-5
```

Input Format 2: The sequence in each line is ended by a slash ("/") character, with the sequence direction from the 5′ end to 3′ end.

```
# primer2  // optional sequence name
ATGCCGTAGGTA/
TAGGTGAGCCAT/
CTCAGCATTGCA
(((((((((::::/
))))((((::::/
)))))))):::::
```

In summary, we have developed a new algorithm for visualizing DNA secondary structures with multiple DNA strands and have implemented the algorithm in a web-based program called DNAdraw. For given secondary structures, DNAdraw identifies all simple stem-loops and composite stem-loops enclosing other stem-loops. Stem-loops are inserted into their enclosing composite stem-loops by rotation, and/or translation operations. The DNAdraw algorithm is the first capable of automatically drawing DNA structures with multiple strands. DNAdraw will be a valuable tool for fast and accurate design of primers and probes in DNA chips.

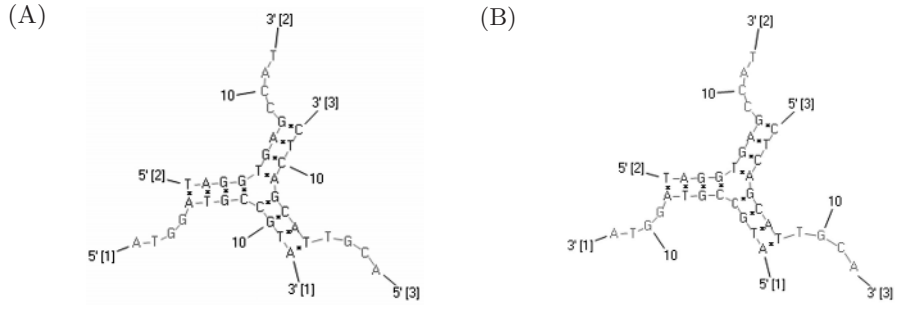

Fig. 3. (A) Structure drawing for input data format 1. (B) Structure drawing for input data format 2.

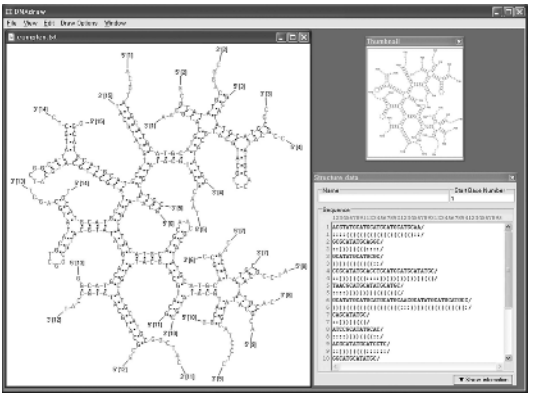

Fig. 4. Hypothetical DNA secondary structure with 15 DNA strands.

References

[1] Lander, E. S.: Array of hope, Nature Genetics **21** (1999) 3-4
[2] De Rijk, P., Wuyts J., De Wachter, R.: RnaViz 2: an improved representation of RNA secondary structure. Bioinformatics **19** (2003) 299-300
[3] Han, K., Kim, D., Kim, H.-J.: A vector-based method for drawing RNA secondary structure. Bioinformatics **15** (1999) 286-197
[4] Matzura, O. and Wennborg, A.: RNAdraw: an integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. Computer Applications in the Biosciences **12** (1996) 247-249