Probe Design for DNA Chips

Ken-ichi Kurata [1] **Akira Suyama** [2]

kurata@genta.c.u-tokyo.ac.jp suyama@dna.c.u-tokyo.ac.jp

[1] Department of Physics, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

[2] Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

1 Introduction

DNA chips are a promising fundamental technology for the post-genome research [1]. A large number of DNA oligonucleotide probes integrated in a small area of the chip surface facilitate speedy simultaneous detection of many target sequences. So, expression profile analysis of thousands of genes can be finished in a few days by using DNA chip technology although it takes years to finish the analysis by using conventional methods such as northern hybridization and SAGE.

High-density DNA chips are made by light-directed combinatorial oligonucleotide synthesis technology, which is based on photolithography technology for semiconductor micro-fabrication [4, 5]. Recently, high-density DNA oligonucleotide chips have been commercially available [2]. Those chips can detect thousands of mRNA transcripts simultaneously. However, as many as 20 pairs of perfect match and center-mismatch oligonucleotide probes are needed to identify each transcript. This is due to low specificity of probe sequences designed by unsatisfactory method [3]. The necessity of many probes not only decreases the detection capacity of DNA chips but also makes DNA chip development difficult and expensive.

Here, we describe a more rational method for designing oligonucleotide probe sequences of DNA chips for gene expression profile analysis, which can reduce the number of DNA probes needed for identification of target sequences. Each transcript is identified with a pair of two oligonucleotide probes, which can hybridize to target transcript specifically under the same hybridization condition. In addition, they can be used as a PCR primer pair, which helps experiments to examine the specificity of designed probe sequences.

2 Methods

Selection of probe sequences from target sequences was based on the criteria of specificity, melting temperature, and secondary structure stability. Five filters were developed to efficiently select probe sequences satisfying these criteria.

The first is a filter of exactly repetitive sequences (RS filter), which efficiently finds sequences appearing in more than one ORF, and removes them from candidates for probe sequences. The second is a filter of frequency of occurrence (FO filter), which estimates the frequency of occurrence of a probe sequence based on the frequency of occurrence of all k-tuples consisting of a probe sequence. Only rarely occurring sequences are selected for probe candidates. Seven-tuples were used in FO filter when designing 30-base probe sequences. The third is a filter of melting temperature (TM filter), which calculates the melting temperature of perfect match probe. The filter classifies probe sequences into fewest groups with a uniform melting temperature. All DNA probes on a chip surface are subjected to the same hybridization and wash conditions. So the uniform melting temperature minimizes false-positive and/or false-negative signals, making accurate target identification possible. The fourth is a filter of secondary structure stability (SS filter), which calculates the free energy of optimal secondary structure of probe. Stable intra-strand secondary structure of probe hinders rapid hybridization to a target sequence, resulting in extensive decrease of the signal intensity. SS filter removes those unfavorable probe sequences. The last is a filter of Hamming distance (HD filter), which calculates the minimum Hamming distance between a probe sequence and a target sequence. This filter examines the specificity of probe more rigorously than FO filter.

Selection of probe sequences using HD filter is slower than the other filters and consumes most computing time because a long target sequence is scanned many times by candidate probe sequences. It is too slow to process a 3 GB human genome sequence on personal computers. However, probe candidates generated by the first four filters usually have good specificity. So the present method can be practically applied to large genome such as human genome.

3 Results and Discussion

Oligonucleotide probes of 30-base long to identify transcripts of 4,289 ORFs of *E. coli* K-12 genome were designed using the present method. T_m distribution of probe candidates obtained with FO filter showed that there was no complete set of specific 30-base probes of a uniform melting temperature. This fact means that no single DNA chip analysis is able to identify 4,289 transcripts accurately because all probes on a chip are subjected to the same hybridization and wash conditions. Probe candidates obtained with TM filter were classified into three groups with uniform melting temperatures differing from each other by more than six degrees. The probe groups with the highest and the second highest melting temperature contained specific probes for 53% and 42% of 4,289 ORFs, respectively. Thus at least three DNA chips and three sets of hybridization and wash conditions are needed for accurate gene expression profile analysis of *E. coli* genome.

Probe candidates are obtained as a PCR primer pair after finishing HD filter. Therefore, the probe specificity is easily examined using PCR experiments. We have chosen two ORFs, yaaJ coding inner membrane transport protein and SSB coding single-stranded DNA binding protein, by chance; and then examined their probe specificity by PCR using *E. coli* genome DNA as template. Both probes showed a single PCR product of correct length. Thus the specificity of probe sequences designed by the present method was confirmed. PCR experiments confirming the specificity of other probes are now under way.

Acknowledgements

This work was performed as a part of the research and development project of Industrial Science and Technology Program supported by NEDO (New Energy and Industrial Technology Development Organization) (SW, YM and YE).

References

[1] *Nature Genetics*, 21(1):supplement, 1999.

[2] Affymetrix, Inc., GeneChip, *http://www.affymetrix.com/.*

[3] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L., Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, 14:1675–1680, 1996.

[4] Pease, A.C., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., Fodor, and S.P., Light-generated oligonucleotide arrays for rapid DNA sequence analysis, *Proc. Natl. Acad. Sci. USA*, 91:5022–5026, 1994.

[5] Suyama, A., DNA chips – Integrated chemical circuits for DNA diagnosis and DNA computes, *Proc. of Third International Micromachine Symposium*, 7–12, 1997.