A Database Aiding Probe Design System for Virus Identification

Feng-Mao Lin[1], Pak-Leong Chan[2], Yu-Chung Chang[3], Hsien-Da Huang[4], Jorng-Tzong Horng[5]

Keywords: probe design, microarray, database system, virus

1 Introduction.

To identify viruses causing disease becomes more and more important today. If viruses had been identified early, virologists would have had sufficient time to treat them. However, a few systems have been developed previously to design probes to identify virus sequences [1-5]. User can select all the virus-specific oligonucleotide probes under the group of viruses. However, user cannot select viruses of different groups. A system for identifying different host categories of viruses in reasonable time is crucial. Probe design for virus sequence identification is very time-consuming, especially in avoiding the cross-hybridization of designed probes against the non-target sequences. We propose a faster algorithm to implement the program of LIS [6]. We apply LIS algorithm to calculate the similarity between each pair of probe and non-target sequences. Our algorithm is faster than the method of using BLAST only. In addition, we make use the database technology aiding for designing an oligonucleotide microarray that can identify virus sequences selected by users.

2 Materials and methods.

We downloaded two kinds of data form different databases. One is virus taxonomy data from the universal virus database of the international committee on taxonomy of virus (ICTVDB) [7]. Another is the virus sequence from NCBI GenBank database. Finally, we integrated virus taxonomy data and virus sequence data to our local database.

The process of probe design contains two sections. They are probe candidates generation and probe selection. A virus sequence was divided into many fragments by sliding a window with 5 nucleotides a time along the whole virus sequence. The size of window is ranged at 20 to 60 nucleotides. Sequence fragments will be stored in probe candidate table in our local database, if and only if the sequence fragment satisfies all the following criteria [8]:

1. Number of any single base (As, Cs, Ts or Gs) does not exceed half of the fragment length.
2. The length of any contiguous As, Cs, Ts, or Gs does not exceed a quarter of the fragment length.
3. GC-content of sequence fragment is in the range of 40% to 60%.
4. No self-complementary is within the sequence fragment.

[1] Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: meta@db.csie.ncu.edu.tw

[2] Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: leong@db.csie.ncu.edu.tw

[3] Department of Biotechnology, Ming Chuan University, Taiwan. E-mail: d80106@mcu.edu.tw

[4] Department of Biological Science and Technology & Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu, Taiwan. E-Mail: bryan@mail.nctu.edu.tw

[5] Department of Computer Science and Information Engineering and Department of Life Science, National Central University, Taiwan. E-Mail: horng@db.csie.ncu.edu.tw

3 Results.

The system was built under Linux Red Hat 9.0 and MySQL 3.x. The LIS program was implemented in c programming language. It was compiled with a Gnu-compiler and ran on Linux workstation. The web interface was implemented in PHP. Virus sequences can be selected in a web page. As the user selects the virus sequences, inputs the temperature and length of probe, the system will select all the probes candidates belonging to the virus sequences selected by the user from the probe candidate database.

To verify probe with low LIS-identity is specific, we selected 7 species virus under a genus (coronavirus) from virus sequence database. The length of probe was set to 50mer and the temperature threshold was set between 75^{o}C and 85^{o}C. 9,097 probes candidates were selected from probe candidate database. The LIS-identity of each probe was calculated against its non-target sequence. We defined cross-hybridization as a probe that is large than 70% similarity to its non-target sequence. Table 1 shows the number of probes and the percentage of cross-hybridization of every 5 LIS-identity. The lower the LIS-identity is, the less possibility the probe has cross-hybridization to its non-target sequences. Although the LIS-identity is approximately proportional to the similarity of probe to its non-target sequence, there are still some cases that the LIS algorithm cannot calculate the similarity of probe to its non-target sequence accurately.

LIS-identity	Number of data	Identity > 70%	Possibility of being cross-hybridization
10~15	31,420	134	0.42%
15~20	28,653	627	2.1%
20~25	4,031	1,133	28%
25~30	1,932	1,267	65%
30~35	947	879	92%
35~40	453	453	100%
40~45	182	182	100%
45~50	41	41	100%

Table 1. Show the data distribution of each scope of LIS-identity and the correlation between LIS-identity and cross-hybridization.

References

[1] Chang, P. C. and Peck, K. 2003. Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes. *Bioinformatics* 19:11 1311-7.

[2] Kaderali, L. and Schliep, A. 2002. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics* 18:10 1340-9.

[3] Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* 28:22 4552-7.

[4] Rouillard, J. M., Herbert, C. J., and Zuker, M. 2002. OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* 18:3 486-7.

[5] Wang, X. and Seed, B. 2003. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* 19:7 796-802.

[6] Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. 1999. Alignment of whole genomes. *Nucleic Acids Res* 27:11 2369-76.

[7] Buechen-Osmond, C. and Dallwitz, M. 1996. Towards a universal virus database - progress in the ICTVdB. *Arch Virol* 141:2 392-9.

[8] Li, F. and Stormo, G. D. 2001. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 17:11 1067-76.