Database to Dynamically Aid Probe Design for Virus Identification

Feng-Mao Lin, Hsien-Da Huang, Yu-Chung Chang, Ann-Ping Tsou, Pak-Leong Chan, Li-Cheng Wu, Meng-Feng Tsai, and Jorng-Tzong Horng

***Abstract*—Viral infection poses a major problem for public health, horticulture, and animal husbandry, possibly causing severe health crises and economic losses. Viral infections can be identified by the specific detection of viral sequences in many ways. The microarray approach not only tolerates sequence variations of newly evolved virus strains, but can also simultaneously diagnose many viral sequences. Many chips have so far been designed for clinical use. Most are designed for special purposes, such as typing enterovirus infection, and compare fewer than 30 different viral sequences. None considers primer design, increasing the likelihood of cross hybridization to similar sequences from other viruses. To prevent this possibility, this work establishes a platform and database that provides users with specific probes of all known viral genome sequences to facilitate the design of diagnostic chips. This work develops a system for designing probes online. A user can select any number of different viruses and set the experimental conditions such as melting temperature and length of probe. The system then returns the optimal sequences from the database. We have also developed a heuristic algorithm to calculate the probe correctness and show the correctness of the algorithm. (The system that supports probe design for identifying viruses has been published on our web page http://bioinfo.csie.ncu.edu.tw/.)**

***Index Terms*—Database, probe design, virus identification.**

I. Introduction

VIRAL infection poses a major problem for public health, horticulture, and animal husbandry, possibly causing severe health crises and economic losses. Viral infections can be identified by the specific detection of viral sequences in two ways: the first is an amplification-based method, such as using the polymerase chain reaction (PCR), the reverse transcription-polymerase chain reaction (RT-PCR), or nested PCR, for example, and the second is the hybridization-based approach, such as the use of southern blotting, northern blotting, dot blotting, and DNA chips. The former not only provides the advantages of fast and specific detection and a lower detection limit, but also has the following weaknesses: 1) the clinicians must assess which viruses are suspected in an infectious event; 2) the nucleotides on the nearest 3′-end of the designed primers are very important to the successful extension of the primer; and 3) although multiplex PCR can be used to detect many viral sequences simultaneously, diagnosing the viral sequences of over 20 different species or strains in a single reaction is currently very difficult. The hybridization-based methods not only tolerate sequence variations of newly evolved virus strains, but can also simultaneously diagnose more viral sequences in a single reaction than can multiplex PCR. Many chips have so far been designed for clinical use. Most are designed for special purposes, such as typing enterovirus infection, and compare fewer than 30 different viral sequences. None considers primer design, increasing the likelihood of cross hybridization to similar sequences from other viruses. To prevent this possibility, this paper establishes a platform and a database that provides users with specific probes for all known viral genome sequences to facilitate the design of diagnostic chips.

Microarray (also called gene chip, DNA chip, and DNA microarray) technology emerged a few years ago. One of its main applications is in diagnosing pathogens. Typically, a microarray is a glass slide or a piece of nylon membrane, on which thousands to tens of thousands of DNA sequences can be spotted. Such spotted DNA sequences are called probes. They can be used to detect different viral infections and distinguish which serotypes or strains are simultaneously involved in a hybridization reaction. In Southern hybridization, a collection of restriction fragments is transferred from an agarose gel to a nylon membrane and the specific ones being studied are detected by hybridization probing [1]. A hybridization probe is a labeled DNA molecule whose sequence is complementary to the target DNA that we wish to detect. Because the probe and target DNAs are complementary they can hybridize by base pairing of the complementary nucleotides.

Two main DNA microarray formats are widely used. They are the oligonucleotide array format and the cDNA array format. In the cDNA microarray format, probes are whole or partial cDNA sequences (300–5000 bases long) immobilized on a solid surface. In the oligonucleotide array format, probes are DNA or RNA sequences that are 20–80 bases long.

Manuscript received July 15, 2004; revised January 25, 2005, April 28, 2005, June 20, 2005, and November 17, 2005. This work was supported in part by the National Science Council, Taiwan, R.O.C., under Contract NSC94-3112-B-008-002 and Contract NSC94-2213-E-008-006.

F.-M. Lin was with the Department of Computer Science and Information Engineering, National Central University, Jhongli City 320, Taiwan, R.O.C. (e-mail: meta@db.csie.ncu.edu.tw).

H.-D. Huang is with the Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, R.O.C. (e-mail: bryan@mail.nctu.edu.tw).

Y.-C. Chang is with the Department of Biotechnology, Ming Chuan University, Taoyuan 333, Taiwan, R.O.C. (e-mail: d80106@mcu.edu.tw).

A.-T. Tsou is with the Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei 112, Taiwan, R.O.C. (e-mail: aptsou@ym.edu.tw).

P.-L. Chan was with the Department of Computer Science and Information Engineering, National Central University, Jhongli City 320, Taiwan, R.O.C. He is now with ZyXEL, Hsinchu 300, Taiwan, R.O.C. (e-mail: leong@db.csie.ncu.edu.tw).

L.-C. Wu and M.-F. Tsai are with the Department of Computer Science and Information Engineering, National Central University, Jhongli City 320, Taiwan, R.O.C. (e-mail: Richard@db.csie.ncu.edu.tw; mftsai@csie.ncu.edu.tw)

J.-T. Horng is with the Department of Life Science and the Department of Computer Science and Information Engineering, National Central University, Jhongli City 320, Taiwan, R.O.C. (e-mail: horng@db.csie.ncu.edu.tw).

Digital Object Identifier 10.1109/TITB.2006.874202

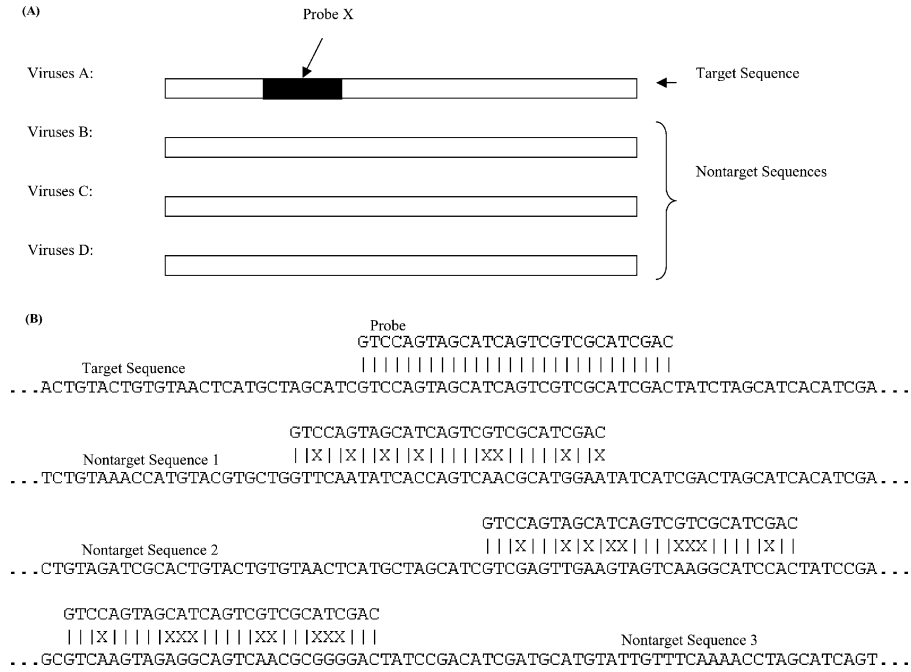

Fig. 1. Example of target and nontarget sequences. (a) Example showing a target sequence of the "probe X" and the nontarget sequences of the "probe X." (b) Example showing that a probe is fully aligned with a target sequence and partially aligned with the nontarget sequences.

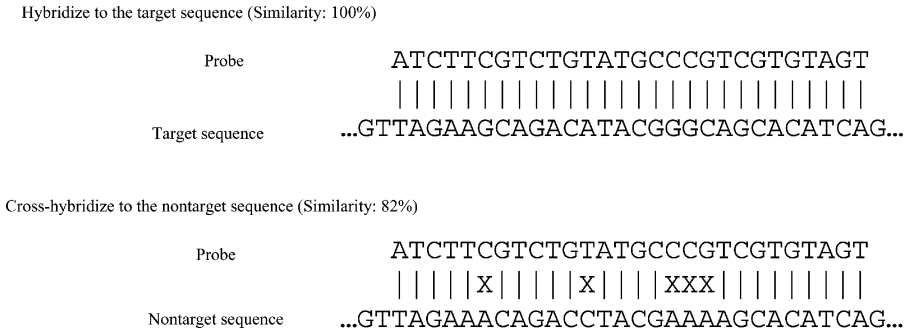

Fig. 2. Example showing how a probe hybridizes to a target sequence and cross hybridizes to a nontarget sequence.

The probe design process is to select several substrings, called probes, from each virus sequence. The "target sequence" of a probe is the virus sequence from which the probe was obtained. It will hybridize with one of a population of sequences of interest. The remaining sequences in the population are "nontarget sequences." The chosen probes should match a substring of their target sequence exactly and not match any substring of any nontarget sequence in the population. Fig. 1(a) shows a probe X that is a part of a virus sequence A. Fig. 1(b) shows how probe X only partially matches nontarget sequences B, C, and D.

One of the current challenges of the microarray technology is the prevention of cross hybridization. Targets are cDNA sequences tagged with fluorescent dyes that are hybridized to the probes. If the target sequence is very similar to nontarget sequences, the probe may hybridize to the target sequence and cross hybridize to the nontarget sequences. Thus, an important aspect of the microarray experiments is the quality of probe design. The best way to assess the quality of probes is by experimental measurements, but this is too expensive. Instead, in the recent studies, the major concern is the similarity between the probe and the nontarget sequences [2], [3]. In this paper a selected probe should be fully hybridized to a target sequence and have at least 30% difference from the nontarget sequences. Fig. 2 is an example of a probe that hybridizes to a target sequence and cross hybridizes to a nontarget sequence. The similarity is the number of matching columns divided by the probe length. The "X" symbol in Fig. 2 indicates that the nucleotides cannot be hybridized to each other. The "|" symbol indicates that the nucleotides can be hybridized.

Probes should be selected according to the criteria of specificity, melting temperature, and sensitivity. The following three main factors that influence virus probe selection were considered:

1) melting temperature or free energy of the oligonucleotide probe;
2) length of contiguous similarity with any other nontarget sequences in the oligonucleotide probe;

3) similarity between each pair consisting of the probe and a nontarget sequence.

All probes must be treated under the same hybridization conditions. Temperature is one of the most important factors. The melting temperature (T_{m}) is the temperature at which the two strands of a double-stranded nucleic acid molecule or base-paired hybrid detach due to complete breakage of hydrogen bonds. This temperature can be obtained using the nearest neighbor model [4], which is defined by the following formula:

$$T_{\mathrm{m}} = \frac{\Delta H}{\Delta S + R\log(c/4)} - 273.15 \tag{1}$$

where ΔH and ΔS represent enthalpy and entropy, respectively, R is the molar gas constant, and c is the total molar concentration of the annealing oligonucleotide.

The second factor that influences the oligonucleotide probe design is the length of contiguous similarity of the probe with any other sequence in the oligonucleotide probe. One report of the sensitivity and specificity of a 50-mer oligonucleotide microarrays [2] suggested that all probes with a 75% overall sequence similarity with their nontarget sequences and contiguous complementary base pairs with a length of under 14 are sufficiently specific to be selected.

The third factor that affects the probe design is the similarity between each probe and nontarget sequences. Although contiguous similarity with other sequences is the primary factor that causes cross hybridization [2], a probe with high similarity to many nontarget sequences will exhibit cross hybridization. Some tools, such as OligoArray [5] and OligoPicker [6], use BLAST to find out the probes whose similarities with their nontarget sequences are high.

A viral sequence must have at least one identifying probe and each probe must hybridize to only a single sequence. The optimal probes should be those that hybridize with their target viral sequences perfectly, but do not hybridize effectively with nontarget sequences.

Many algorithms exist for selecting optimal probes, including a method based on the matching frequency of the sequence landscape [7], a method based on a hash table and the BLAST [6], a method based on the longest common factor (LCF) between a probe and nontarget sequence [8], a method based on the melting temperature of a probe [9], and a method based on unique segments [3]. A common factor of two strings s, t is a string that is both a substring of s and t. A common factor is an LCF if no longer common factor exists [8].

This study uses the longest increasing subsequence (LIS) algorithm, which is faster than the alignment algorithm, to calculate the similarity of each probe with its nontarget sequences. The set of optimal probes can identify their target viral sequence in a short time, which may not cause the webpage to timeout. Section III presents a detailed comparison.

II. System and Methods

The Appendix gives the definitions of a number of terms used in this paper.

TABLE I
DATA FOR DIFFERENT VIRUS GENOMES

Type of virus genome	Number of virus sequences
Deltavirus	1
dsDNA virus	331
dsRNA virus	219
ssRNA negative-strand virus	120
ssRNA positive-strand virus	472
ssDNA virus	43
retroid virus	78
satellites virus	43
unclassified virus	5

TABLE II
TAXONOMIC DATA ON VIRUSES

Virus taxonomy	Number of data
Family	98
Genus	249
Species and Subspecies	2003

A. Data Preparation

The proposed system uses two databases. One is of taxonomic data about viruses, taken from the universal virus database of the International Committee on the Taxonomy of Viruses (ICTVDB) [11]. The other is the viral sequences from the NCBI GenBank database. We use the data retrieval tool IntKey downloaded from the ICTVDB to retrieve virus taxonomy data from the ICTVDB. We download virus DNA sequences from NCBI GenBank. Virus taxonomy data and data about viral sequences are integrated in the local database, in which three tables (family, genus, species) store taxonomic data and one table stores the DNA sequences of viruses. The sequence table contains 1535 complete virus genomes. The average length of the viral sequences is 11 142 nucleotides. The sequence table provides the genomes of the viral sequences and the natural hosts of the virus. Table I shows the data for different viral sequences and Table II shows the data for different levels of the viral taxonomy.

B. Generating Probe Candidates

A viral sequence is divided into many fragments by sliding a window along it in steps of five nucleotides. The size of the window is from 20 to 60 nucleotides. Sequence fragments are

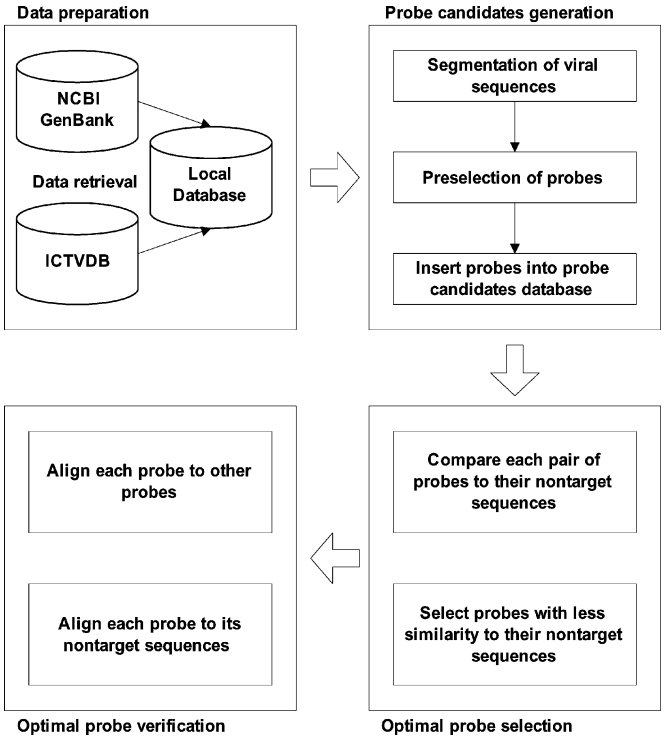

Fig. 3. Database system for designing probes for viruses.

stored in the local database, if and only if, the sequence fragment satisfies all the following criteria [7]:

1) number of occurrences of any single base (As, Cs, Ts, or Gs) does not exceed half of the length of the fragment;
2) length of any section of contiguous As, Cs, Ts, or Gs does not exceed a quarter of the length of the fragment;
3) GC-content of the sequence fragment ranges from 40% to 60%;
4) sequence fragments are not self-complementary.

The database includes about 10 million probe candidates from 1535 viral sequences. The melting temperature of each probe is calculated by MELTING [12]. A user selects a set of viral sequences and inputs the length and the experimental temperature for probe design. All the probe candidates that belong to the set of viral sequences and satisfy the conditions are selected from the probe candidate database.

C. System Flow

Fig. 3 shows an overview of the system used for the probe design. The system has four main phases.

Data preparation: Viral sequences and viral taxonomies are downloaded from the GenBank (NCBI) and ICTVDB, respectively.

Generating candidate probes: Viral sequences are divided into segments by sliding a window along the sequence of five nucleotides at a time. The segments are preselected by the probe filter [7] and inserted into the database of candidate probes. Then, the melting temperatures of all candidate probes are calculated using MELTING [12].

Selecting the optimal probe: After a user has selected the target sequences for designing the viral probe, all of the candidate probes are selected according to the input parameters (melting temperature, range of melting temperature, and length of the probe). The optimal probes are those that are not very similar to their nontarget sequences. The LIS algorithm is used to find the optimal probe.

Verification of the optimal probe: The optimal probes are verified by two processes. One is the alignment of probe to the other probes. This process ensures that the optimal probes will not match the same region of the target sequence. The other is the alignment of each probe to its nontarget sequences. If the alignment score is high, the optimal probe is discarded.

All four phases are integrated as a web-based system. This system supports the cross selection of species of viruses, instant optimal probe selection, and the online verification of the result by local alignment.

D. Algorithm and Implementation

The algorithm was coded in C/C++ and compiled with the GNU C compiler. The system runs on a computer with an AMD Athlon XP CPU running at 1.8 GHz with 1-Gb main memory. The operating system was Redhat 9.0 and the database management system was MySQL.

LIS Algorithm: The most time-consuming part of probe design is determining the most similar regions between the probe and nontarget sequences. Many methods use an alignment tool like BLAST to calculate the similarity of each probe with nontarget sequences. However, this method of calculation is not efficient. We apply a fast method to calculate the similarity of probes to nontarget sequences. Before describing the method, we define the suffix. Let S be a string with length $|S|$ and let $S[i,j]$ be the substring from the ith character to the jth character of S. A suffix is a substring of the form $S[i,|S|]$, where $1 \leqq i \leqq |S|$. An investigation of the alignment of whole genomes [13] applied a suffix tree and the LIS algorithm to find the parts of two sequences that were most similar to each other. Recently, a fast method of alignment, based on BLAST and the LIS algorithm, was published [14]. It uses BLAST to identify some conserved regions in the two sequences and applies the LIS to combine these conserved regions. The two sequences can thus be globally aligned. Both methods efficiently determine the similarity between the two sequences. The LIS algorithm used in the above-mentioned two studies is applied here to evaluate efficiently the similarity between the probe and nontarget sequences. The algorithm is very fast, so the process is completed in a short time and the results can be published to a web page.

The LIS problem is defined as follows: Given a sequence S, the LIS of S is the longest subsequence in which each number is greater than its predecessor. For example, the LIS of a sequence $S = (7, 9, 1, 6, 2, 4, 8)$ is $(1, 2, 4, 8)$. The algorithm for determining the LIS is the LIS algorithm. There are two major implementations of the LIS algorithm. One is a dynamic programming technique and its time complexity is $O(n2)$. By using a binary search, the time complexity of generating the LIS can be reduced to $O(n \log n)$ [15], where n is the number of

Input: File containing virus sequences selected by the user
File containing probe candidates of sequences selected by the user

Output: File containing similarity of each pair of probe to its nontarget sequence

1. **For** each virus sequence of the input file
2. Generate the suffix array from the sequence
3. **For** each probe candidate of the input file
4. Divide the probe into 4-nucleotide substrings (tag) by sliding a window one nucleotide at a time
5. Compare the tag with the first four nucleotides of each suffix of the sequence
6. **if** (tag matches the suffix of the sequence)
7. Add the number of the tag and the position of the nontarget sequence to the number sequence S from which the LIS will be generated.
8. **end if**
9. Find the maximum LIS at each position of S
10. Calculate the similarity of the probe to the nontarget sequence from the LIS
11. The similarity of probe to the nontarget sequence is the element of result set
12. **Next**
13. **Next**

Fig. 4. FindLIS algorithm.

elements in the sequence. Here, we used the faster algorithm to implement the LIS in the C programming language. The program, called findLIS, identifies the most similar parts of a probe and a nontarget sequence. It first generates a suffix array for one of the sequences selected by a user. Each probe candidate is divided into fragments, called tags, by sliding a window one nucleotide at a time along the whole probe. The length of the tags is set to four. If a tag matches the first four nucleotides of a suffix in the suffix array, then both the number of the tag and the number of suffix of the sequence are recorded. A long sequence of numbers is generated when all the tags match the suffixes of the sequence. Since the suffixes of the sequence are sorted lexically, we can use a binary search to find the matching positions of tags of the probe and suffixes of the sequence. When these matching positions are generated, the LIS can be found in this sequence of numbers. The most significant region is the subsequence in which the LIS is located. The regions of the sequence covered by the LIS include all parts that are matched by most tags of the probe. That is, the region of the probe that is most similar to the nontarget sequences is with the LIS. The LIS of each pair that consists of the probe and a nontarget sequence can be obtained by comparing the tags of the probe with the suffixes of the nontarget sequence. Fig. 4 shows the findLIS algorithm and Fig. 5 shows an example of using the findLIS program to identify the most similar regions of the probe and the nontarget sequence. The length of the probe is selected by a user and the candidate lengths of probes are 20 bp, 30 bp, 40 bp, 50 bp, and 60 bp.

E. Selecting the Optimal Probe

The LIS similarity is defined as the number of nucleotides of a probe that matches the nontarget sequence in the LIS. The LIS similarity can be obtained from the LIS of the number sequence generated by the program findLIS. The optimal probe is the one whose LIS similarity is least.

F. Verifying the Selection of the Optimal Probe

Two processes are required to confirm the optimal probes selected by the proposed system. The first process is the alignment of the pairs of the optimal probes. If optimal probes with high similarity are selected, then the probes will identify the same region of the target sequence or the neighboring regions. That is, the probes overlap in the region of the target sequence. The alignment of the pairs of optimal probes reveals that a probe can be discarded if it has high similarity with other probes, ensuring that the only one optimal probe will base pair with the extensible regions of the target sequence. The second process is the local alignment of the probe with its nontarget sequences. The local alignment tool MATCHER [16] is used to verify the quality of the optimal probes. If the similarity of the probe to the nontarget sequences calculated by MATCHER is high, cross hybridization will occur. If the similarity of probe with its nontarget sequence is high, the user can discard the probes from the set of optimal probes. Both processes are implemented in the web service so the user can verify optimal probes using the web interface.

III. Results

A web interface was designed. Users may select probe sequences to identify the viruses of interest. The system takes about 180 min to design probes for 100 viruses including selecting candidate probes and selecting the optimal probes. Although some methods and tools exist for designing probes for microarrays, few online systems have been developed and few allow users to select sequences dynamically across different

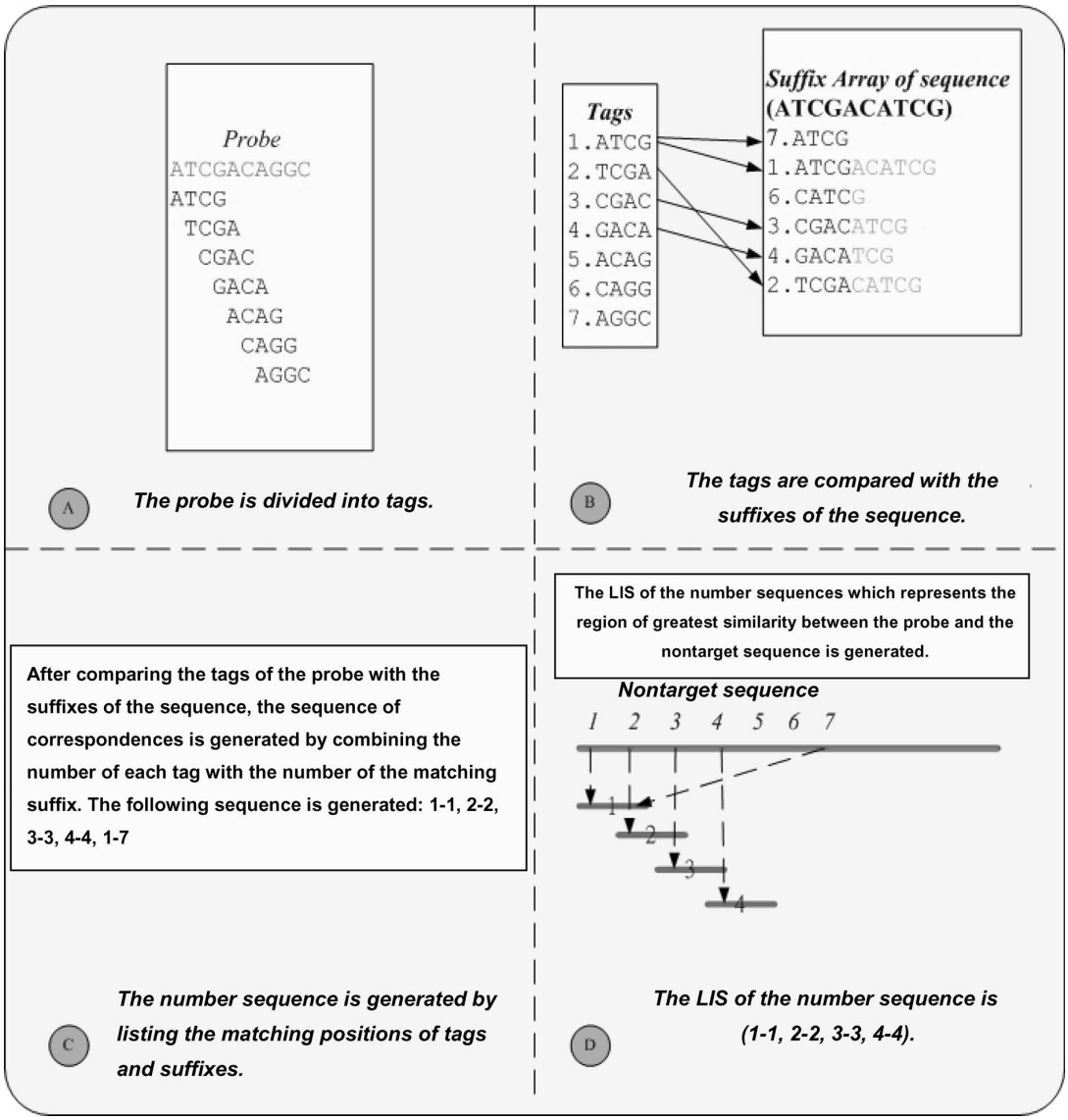

Fig. 5. Example of using the findLIS program to determine the region of greatest similarity between a probe and a nontarget sequence.

virus genera and virus families. Table III compares the tools and methods for designing oligonucleotide probes.

One hundred viruses were randomly selected and the experimental melting temperatures of 75°C–78°C were used to confirm that the LIS similarity of a probe and a nontarget sequence is directly proportional to the similarity between them. A total of 27 377 probe candidates with lengths of 50 bp were selected. Empirical data have suggested that under appropriate hybridization conditions and target concentrations which are commonly employed for microarray studies, the cross hybridization between the probe and nontargeted sequences can be estimated by the similarities between the probe's target sequence and nontarget sequences [2]. Several studies have concluded that a threshold of around 70% sequence similarity can be used as a reference for cross-hybridization prediction. The LIS similarity of each pairing of a probe and a nontarget sequence was calculated by the program findLIS. The correspondence between LIS similarity and similarity is shown in Fig. 6, which compares the similarity between a probe and a nontarget sequence with the average LIS similarity. When the similarity of a probe with its nontarget sequence is large, the average LIS similarity of a probe with its nontarget sequence is also large. According to Fig. 6, the average LIS similarity that corresponds to a 70% similarity between the probe and its nontarget sequence is about 25 bp (indicating a match to the nontarget sequence over 25 bp). In Fig. 6, when the average LIS similarity of probe is below 20 bp, the corresponding similarity of the probe is below 70%. A probe with LIS similarity shorter than 20 bp is thus determined to be effective for use in probe design.

IV. Discussion and Conclusion

Most methods use the BLAST program as the primary tool to avoid cross hybridization. They spend time calculating the

TABLE III
COMPARISON OF DIFFERENT TOOLS AND METHODS FOR OLIGONUCLEOTIDE PROBE DESIGN

Name	Reference sequence	Cross hybridization	T_m Calculation	App. Type	Usage	References
OligoDB	Human cDNA	BLAST	The nearest neighbor model	Web	Detect transcription profiling human gene	[17]
Oligopicker	RefSeq, TIGR	BLAST	Simple formula (Schildkraut 1965)	Tool	Detect RNA expression	[6]
VirOligo	Virus from GenBank	BLAST, hash technique	Simple formula (Bolton and McCarthy, 1962)	Web	Virus sequence identification	[18]
OligoArray	mRNA, CDS, exon	BLAST	The nearest neighbor model	Tool	Detect gene expression	[5]
Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes	TIGR human THC and mouse TC	unique segment	The nearest neighbor model	Tool	Detect gene expression	[3]
PROBSEL	Virus sequence	suffix tree(preselection) T_m (Optimal Selection)	The nearest neighbor model	Tool	Virus sequence identification	[9]
Selection of optimal DNA oligos for gene expression array	Phage genome	suffix array sequence landscape	The nearest neighbor model	Web	Gene expression	[7]
PROMIDE	General	suffix array (LCF)	The nearest neighbor model	Tool	Gene expression	[8]
Rapid Large-Scale Oligonucleotide Selection for Microarry	General	jump in LCF, free energy	The nearest neighbor model	Tool	Gene expression	[19]
Database for virus probe design (our approach)	Virus sequences from GenBank	selection of minimum similarity calculated by the longest increasing subsequence	The nearest neighbor model	Web	virus sequence identification	This work

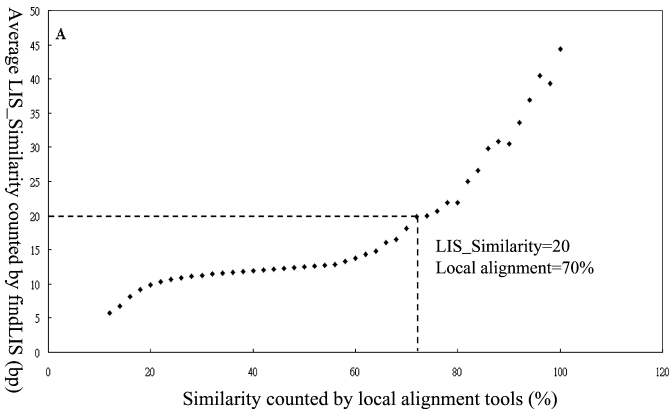

Fig. 6. Correspondence between LIS similarity and similarity.

similarity of the probe with nontarget sequences. The present approach first generates the probe candidates in the database and then uses the LIS algorithm to evaluate the similarity between the probe and nontarget sequences more efficiently. The database technique and the algorithm can be used to finish the process of designing probes for 100 sequences in 3 h. The optimal probes are verified by the alignment tool. Because of the efficient algorithm and database technology, the virus probe design can be carried out online. Although the program find LIS can efficiently calculate the similarity of probes and nontarget sequences, in some cases the program findLIS will fail to calculate this accurately. Three main factors affect the accuracy of LIS similarity calculation. In the assessment above, probes with an LIS similarity of 4–10 are selected and most have similarities with the nontarget sequence of below 70%. However, in some cases, the similarity of probe with its nontarget sequence exceeds 70%. The first factor is the uncovered region. The length of the tag is set to four, and so the tags may fail to cover many regions of the nontarget sequences, in which case the findLIS program cannot calculate the similarity of probes and nontarget sequences accurately. Fig. 7 shows a case of failure of calculating similarity of a probe with a nontarget sequence. The "X" symbol in Fig. 7 indicates that the nucleotides cannot be hybridized to each other. The "|" symbol indicates that the nucleotides can be hybridized. The actual similarity of the probe to its nontarget sequence is 82% but the LIS similarity calculated by findLIS is 26%. Some matching regions with fewer than four nucleotides are omitted. However, if the length of the tag decreases, the time for calculating the similarity of probe to its nontarget sequence will increase. The second factor

```
ATCTCCACCCGGAGCTTGTTCAT
|||X||||||X|||X|||X||||
TAGTGGTGGGGCTCCAACTAGTA
```

Fig. 7. Problem of uncovered region. LIS similarity is 6 (26% similarity) but the actual similarity is 82%.

is several matching positions in a nontarget sequence for one tag. When one tag matches more than one position in a nontarget sequence, one of the various sequences of number may be generated by the comparison of the tag with the suffix. In this study, one of the possible paths is randomly selected. If there are many possible paths, the accuracy of the calculation will be affected. The third factor is that the LIS may not be unique in the number sequence. For example, consider the number sequence $S = (9, 8, 1, 7, 2, 5, 3)$. Either (1, 2, 3) or (1, 2, 5) is a candidate LIS. The maximum LIS similarity can be found in any one of the possible LIS. This study considers only one of the possibilities.

The uncovered region problem is the major cause of failure in calculating the similarity of a probe with a nontarget sequence. Although the three factors will affect the accuracy of calculation, the optimal probes can be verified by the two alignment methods. These processes ensure that the specificity of the optimal probes is high. In this assessment, when the LIS similarity is set between four and ten, the probability of selecting a cross-hybridization probe is 0.0034%. When the user selects probes with a low LIS similarity (4–15), the probes are specific enough to identify the sequences selected by the user in the ICTVDB.

We are now planning to design probes for different groups, genera and families, to facilitate the detection of newly emerged viruses all over the world.

APPENDIX

DEFINITIONS OF THE TERMS USED IN THIS PAPER

Suffix: Let S be a string with length $|S|$ and let $S[i, j]$ be the substring from the ith character to the jth character of S. A suffix is a substring of the form $S[i, |S|]$, where $1 \leqq i \leqq |S|$.

Hybridization: The attachment by base pairing of two complementary polynucleotides.

Melting Temperature (T_{m})*:* The temperature at which the two strands of a double-stranded nucleic acid molecule or base-paired hybrid detach due to complete breakage of hydrogen bond.

LCF: A common factor of two strings s, t is a string that is both a substring of s and t. A common factor is an LCF if no longer common factor exists [8].

Sequence Landscape: The sequence landscape of a sequence is the frequency distribution of the proper subsequences of the sequence [10]. For example, if the sequence is "aagaa," the sequence landscape is ("a," 4), ("g," 1), ("aa," 2), ("ag," 1), ("ga," 1), ("aag," 1), ("aga," 1), ("gaa," 1), ("aaga," 1), ("agaa," 1).

REFERENCES

[1] T. A. Brown, *Genomes*. Manchester, U.K.: Wiley, 1999.

[2] M. D. Kane, T. A. Jatkoe, C. R. Stumpf, J. Lu, J. D. Thomas, and S. J. Madore, "Assessment of the sensitivity and specificity of oligonucleotide (50 mer) microarrays," *Nucleic Acids Res*., vol. 28, pp. 4552–4557, 2000.

[3] P. C. Chang and K. Peck, "Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes," *Bioinformatics*, vol. 19, pp. 1311–1317, 2003.

[4] J. SantaLucia, Jr., "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc. Nat. Acad. Sci. USA*, vol. 95, pp. 1460–1465, 1998.

[5] J. M. Rouillard, C. J. Herbert, and M. Zuker, "OligoArray: Genome-scale oligonucleotide design for microarrays," *Bioinformatics*, vol. 18, pp. 486–487, 2002.

[6] X. Wang and B. Seed, "Selection of oligonucleotide probes for protein coding sequences," *Bioinformatics*, vol. 19, pp. 796–802, 2003.

[7] F. Li and G. D. Stormo, "Selection of optimal DNA oligos for gene expression arrays," *Bioinformatics*, vol. 17, pp. 1067–1076, 2001.

[8] S. Rahmann, "Rapid large-scale oligonucleotide selection for microarrays," in *Proc. 1st IEEE CSB Conf.* Aug. 14–16, 2002, pp. 54–63.

[9] L. Kaderali and A. Schliep, "Selecting signature oligonucleotides to identify organisms using DNA arrays," *Bioinformatics*, vol. 18, pp. 1340–1349, 2002.

[10] S. Levy, L. Compagnoni, E. W. Myers, and G. D. Stormo, "Xlandscape: The graphical display of word frequencies in sequences," *Bioinformatics*, vol. 14, pp. 74–80, 1998.

[11] C. Buechen-Osmond and M. Dallwitz, "Towards a universal virus database—progress in the ICTVdB," *Arch. Virol*., vol. 141, pp. 392–399, 1996.

[12] N. L. Novere, "MELTING, computing the melting temperature of nucleic acid duplex," *Bioinformatics*, vol. 17, pp. 1226–1227, 2001.

[13] A. L. Delcher, S. Kasif, R. D. Fleischmann, J. Peterson, O. White, and S. L. Salzberg, "Alignment of whole genomes," *Nucleic Acids Res*., vol. 27, pp. 2369–2376, 1999.

[14] H. Zhang, "Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm," *Bioinformatics*, vol. 19, pp. 1391–1396, 2003.

[15] D. Gusfield, *Algorithms on Strings, Trees and Sequences*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[16] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European molecular biology open software suite," *Trends Genet*., vol. 16, pp. 276–277, 2000.

[17] R. Mrowka, J. Schuchhardt, and C. Gille, "Oligodb—interactive design of oligo DNA for transcription profiling of human genes," *Bioinformatics*, vol. 18, pp. 1686–1687, 2002.

[18] K. Onodera and U. Melcher, "VirOligo: A database of virus-specific oligonucleotides," *Nucleic Acids Res*., vol. 30, pp. 203–204, 2002.

[19] S. Rahmann, "Fast and sensitive probe selection for DNA chips using jumps in matching statistics," in *Proc. IEEE CSB'03 Conf.,* Aug. 11–14, 2003, pp. 57–64.

Feng-Mao Lin was born in E-Lan, Taiwan, R.O.C. He was working toward the Ph.D. degree in the Department of Computer Science, National Central University, Jhongli City, Taiwan, between 2001 and 2005.

He is currently serving in the military.

Hsien-Da Huang was born in Taoyuan, Taiwan, R.O.C., in 1975. He received the Ph.D. degree in computer science and information engineering from National Central University, Jhongli City, Taiwan, R.O.C., in 2003.

Since 2003, he has been with the Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu, Taiwan, R.O.C. His current research interests include bioinformatics, database systems, and data mining.

Yu-Chung Chang was born in Keelung, Taiwan, R.O.C., on August 30, 1960. He received the M.S. degree from National Chung-Hsing University, Taichung, Taiwan, and the Ph.D. degree in microbiology and immunology from National Yang-Ming University, Taipei, Taiwan, in 1999.

He is a veterinarian. During 1999–2002, he participated in the annotation works of the human and Ganoderma lucidum genome projects conducted at National Yang-Ming University. In 2002, he joined the Department of Biotechnology, Ming Chuan University, Taoyuan, Taiwan, as a member of the faculty, and is currently an Associate Professor. His current research interests are microbiology, genomics, and bioinformatics.

Ann-Ping Tsou was born in Tainan, Taiwan, R.O.C., in 1950. She received the Ph.D. degree in microbiology from Harvard University, Cambridge, MA, in 1982.

From 1985 to 1993, she was a Staff Scientist with the Discovery Research Division, Syntex (USA) Inc. In 1993, she joined the Institute of Biotechnology in Medicine, National Yang-Ming University, Taipei, Taiwan, R.O.C., where she is currently an Associate Professor. Her current research interests include transcription regulation and functional genomics of hepatocellular carcinoma.

Pak-Leong Chan was born on May 4, 1976. He received the Master's degree in computer science and information engineering from National Central University, Jhongli City, Taiwan, R.O.C., in 2004.

He is currently an Engineer with ZyXEL, Hsinchu, Taiwan.

Li-Cheng Wu was born in Taipei, Taiwan, R.O.C., in 1973. He received the Ph.D. degree in computer science and information engineering from National Central University, Jhongli City, Taiwan, R.O.C., in 2004.

He is currently a Postdoctoral Fellow in the Department of Computer Science and Information Engineering, National Central University. His current research interests include bioinformatics, database systems, and data mining.

Meng-Feng Tsai was born in Taipei, Taiwan, R.O.C., in 1967. He received the Ph.D. degree in computer science from the University of California, Los Angeles, in 2004.

In 2004, he joined the Department of Computer Science and Information Engineering, National Central University, Jhongli City, Taiwan, as an Assistant Professor. His current research interests include data warehousing, database systems, data mining, and bioinformatics.

Jorng-Tzong Horng was born in Nantou, Taiwan, R.O.C., in 1960. He received the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1993.

In 1993, he joined the Department of Computer Science and Information Engineering, National Central University, Jhongli City, Taiwan, R.O.C., where he became a Professor in 2002. His current research interests include database systems, data mining, genetic algorithms, and bioinformatics.