A DATABASE TO AID PROBE DESIGN FOR VIRUS IDENTIFICATION

FENG-MAO LIN[1], HSIEN-DA HUANG[2], YU-CHUNG CHANG[3], PAK-LEONG CHAN[1],
JORNG-TZONG HORNG[1], MING-TAT KO[4]

[1]Department of Computer Science and Information Engineering, National Central University,
No.300, Jungda Rd, Jhongli City, Taoyuan, Taiwan 320
ChungLi, 320, Taiwan

[2]Department of Biological Science and Technology & Institute of Bioinformatic, National Chiao-Tung University, 1001 Ta Hsueh Road, Hsinchu, Taiwan 300
Hsin-Chu, 300, Taiwan

[3]Department of Biotechnology, Ming Chuan University, 250 Chung Shan North Road, Section 5,
Taipei, 111, Taiwan

[4]Institute of Information Science, Academia Sinica, 128 Sec. 2, Academia Rd, Nankangm,
Taipei, 115, Taiwan

Viral infection poses a major problem for public health, horticulture and animal husbandry, possibly causing severe health crises and economic loss. Viral infections can be identified by the specific detection of viral sequences in two ways, the first is the 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 provides the advantages of fast and specific detection and a lower detection limit, but also has some the following weakness; (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 of the extension of the primer; (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 method can 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 fat been designed for clinical use. Most are designed for special purpose, such as typing enterovirus infection, and compare fewer than 30 different viral sequences. None considers all primer design, increasing the likelihood of cross hybridization of similar sequences with other viral sequences. To prevent this possibility, this work establishes a platform and database that provides users with specific probes of all known viral genome sequences, to designing their diagnostic chips. This work develops a system for designing probes online. A user can select any number of different viruses and set their experimental conditions. Including, for example, melting temperature, length of probe. The system then return the optimal sequences to suspected viral infections to be automatically identified from database. The system that supports probe design for identifying virus has been published on our web page http://bioinfo.csie.ncu.edu.tw. Contact: horng@db.csie.ncu.edu.tw.

1 Introduction

1.1. *Microarray*

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 slide of glass or a piece of nylon membrane, above which thousands to tens 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.

One of the current challenges of microarray technology is the prevention of cross-hybridization. If a probe is very similar to its non-target sequences, cross-hybridization may occur. An oligonucleotide probe is shorter than cDNA and can be more easily and economically prepare. An oligonucleotide can be better spotted on a microarray than can cDNA for spotting, in solving the problem of cross hybridization and identifying more viruses.

1.2. *Criteria of probe design*

Probes were selected according to the criteria of specificity, melting temperature, and sensitivity. The following three main factors that influence virus probe selection are considered. The melting temperature or free energy of the oligonucleotide probe. The length of contiguous identity with any other non-target sequences in the oligonucleotide probe. The identity between each pair of the probe and the non-target sequence.

All probes must be treated under the same hybridization conditions. Temperature is one of the most important factors. The melting temperature T_m can be obtained using the nearest neighbor model [1].

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

The third factor that affects probe design is the identity between each probe and its non-target sequence. Although contiguous identity with other sequences is the primary factor that causes cross-hybridization [2], a probe with a high identity with other non-target sequences always definitely causes cross-hybridization. Some tools, such as OligoArray [3], and OligoPicker [4] use the BLAST to find out the probes whose identities with their non-target sequences are high.

1.3. *Optimal probe for microarray*

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

Many algorithms exist for selecting optimal probes, including a method based on the matching frequency of the sequence landscape [5], a method based on a hash table and the BLAST [4], a method based on the longest common factor between probe and non-target sequence [6], a method based on the melting temperature of a probe [7], and a method based on unique segments [8]. This study uses the longest increasing subsequence algorithm (LIS), which is faster than the alignment algorithm, to calculate the identity of each probe with its non-target sequence. The set of optimal probe can identify its target viral sequence in a reasonable time. The results section presents a detailed comparison.

Probe design is time consuming and few online systems exist. We design a database of candidate probes which contains appropriate probes and the melting temperature of the probe is also calculated. A fast algorithm for calculating identity between the probe and the non-target sequence is proposed so that the process of selection of optimal probe can be finished in a short time. The selection of optimal probe can be processed online and the cross-hybridization will not occur in our optimal probes. Therefore, user can select viral sequences arbitrarily on web and can immediately obtain optimal probe for selected virus. Many studies propose the identity of probe with non-target sequence as the point to evaluate the degree of cross-hybridization. We also apply the alignment of probe with non-target sequence online and the result is shown as graphic table.

2 System and Methods

2.1. *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 Virus (ICTVDB) [9]. Another is the viral sequence from NCBI GenBank database. We use a data retrieval tool IntKey downloaded from ICTVDB to retrieval virus taxonomy data from ICTVDB. Then 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) stores taxonomic data and one table stores the DNA sequences of viruses. The sequence table contains 1,535 virus complete genomes. The sequence table provides the genomes of the viral sequences and the natural hosts of the virus.

2.2. *Generating probe candidates*

A viral sequence is divided into many fragments by sliding a window in steps of five nucleotides. The size of the window is from 20 to 60 nucleotides. Sequence fragments are stored in the local database, if and only if the sequence fragment satisfies all the following criteria [5]. The number of occurrences of any single base (As, Cs, Ts or Gs) does not exceed half of the length of the fragment. The length of any section of contiguous As, Cs, Ts or Gs does not exceed a quarter of the length of the fragment. The GC-content of the sequence fragment ranges from 40% to 60%. The sequence fragments is not at all self-complementary.

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

2.3. *System flow*

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 fragments by sliding a window five nucleotides at a time. The fragments are preselected by the probe filter [5] and inserted into the database of candidate probes. Then, the melting temperatures of all candidate probes are calculated using MELTING [10].

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 probe). The optimal probes are those that are not very similar to their non-target sequences. The longest increasing subsequence algorithm is used to find out the optimal probe.

Verification optimal probe: As the optimal probes are selected by using the LIS algorithm, the optimal probes are verified by two processes. One is the alignment of probe to the other probes. This process makes sure that the optimal probes will not match the same region of the target sequence. The other is the alignment of each probe to its non-target sequences. If the alignment score is high, the optimal probe can be discarded.

3 Algorithm and Implementation

3.1. *Longest increasing subsequence algorithm*

The most time consuming part for probe design is to find out the most similar regions between probe and its non-target sequence. Many methods use alignment tool like the BLAST to calculate the identity of each probe with its non-target sequence. However the method of calculation of the identity of probe with its non-target is not efficient. We apply a fast method to calculate the identity of probe to its non-target sequence. An investigation of the alignment of the whole genomes [11] applied a suffix tree and the longest increasing subsequence 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 longest increasing subsequence algorithm (LIS) was published [12]. It used the BLAST to identify some conserved regions in the two sequences and applied LIS to combine these conserved regions. The two sequences can thus be globally aligned. Both methods efficiently determine the identity between two sequences. The LIS algorithm used in above two studies is applied herein to valuate efficiently the identity between the probe and its non-target sequence. The algorithm is very fast so that the process of calculating the identity between probe and its non-target sequence can be finished in a short time and the results can be published to a web page.

The algorithm for determining the longest increasing subsequence is the LIS algorithm. For example, the longest increasing subsequence of a sequence S=(7, 9, 1, 6, 2, 4, 8) is (1, 2, 4, 8). There are two major implementations of LIS algorithm. One is dynamic programming technique and its time complexity is $O(n^2)$. By using binary search, time complexity of generating LIS can be reduced to $O(n \log n)$ [13], where n is the number of elements in the sequence of numbers. A program called findLIS in C programming language is implemented to identify the most similar parts of a probe and a non-target sequence, using the LIS algorithm. This program 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 the window one nucleotide at a time along the whole probe. The length of tag is set to four. If a tag matches the first four nucleotides of the 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 compare with suffixes of the sequence. Since the suffixes of the sequence are sorted lexically, we can make use of binary search to find out the matching positions of tags of probe and suffixes of sequence. When the matching positions of tags of probe and suffixes of sequence is generated, the LIS can be found in the sequence of number generated by the matching positions of tags of probe and suffixes of sequence. The most significant region is the subsequence in which the longest increasing subsequence located and the size of the longest increasing subsequence is within the length of the probe. The regions of the sequence covered by LIS all parts that are matched by most tags of the probe. That is, the most similar region of probe to its non-target sequence is within the LIS. The LIS of each pair of the probe and the non-target sequence can be obtained by comparing the tags of a probe with the suffixes of non-target sequence. Figure 1 shows the algorithm of findLIS and Fig. 2 shows the example of using the findLIS program to identify the most similar region of the probe and the non-target sequence.

3.2. *Selecting optimal probe*

LIS-Identity is defined as the number of nucleotides of a probe matches the non-target sequence in the longest increasing subsequence. The LIS-Identity can be obtained form the LIS of the number sequence generated by the program findLIS. The optimal probe is the one whose LIS-identity is low.

3.3. *Verifying the selection of optimal probe*

Two processes are required to confirm the optimal probes selected by the proposed system. The first process is the alignment of any two optimal probes. If optimal probes with high identity 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 result of the alignment of probes reveals that probe with high identity with the other probes can be discarded, ensuring that the optimal probes will base-pair with the extensible regions of the target sequence. The second process is the local alignment of the probe with its non-target sequences. The local alignment tool MATCER [14] is user to verify the quality of the optimal probes. If the identity of probe to its non-target sequences calculated by MATCHER is high, the cross hybridization will occur. If the identity of probe with its non-target sequence is high, the user can discard the probes from the result set of optimal probes. Both processes are implemented in web service so the user can verify optimal probes using the web interface.

```
Input: File containing virus sequences selected by the user
       File containing probe candidates of sequences selected by user
Output: File containing identity of each pair of probe to its non-target sequence
For each virus sequence of the input file
  Generate suffix array from the sequence
  For each probe candidate of input file
    Divide probe to 4 nucleotides subsequence (tag) by sliding window one nucleotide a time
    Compare the tag with the first four nucleotides of suffixes of sequence
    if (tag matches the suffix of the sequence)
      Add the number of tag and the position of non-target sequence to a number sequence $S$ from which the
          maximum LIS will be generated.
    end if
    Find the maximum LIS in each position of $S$
    Calculate the identity of probe to its non-target sequence from the maximum LIS
    The identity of probe to its non-target sequence is the element of result set
  Next
Next
```

Fig. 1. The findLIS algorithm.

4 Results

A web interface was designed. Users may select probe sequences to identify these viruses of interest. The system takes about 180 minutes to design probes for 100 sequences 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 can allow users to select sequences dynamically across different virus genera and virus families. Table 1 compares the tools and methods for designing oligonucleotide probes.

One hundred sequences were randomly selected and the experimental melting temperature of 75oC to 78oC was used to confirm that the LIS_Identity of a probe with its non-target sequence is directly proportional to similarity between the probe and its non-target sequence. 27,377 probe candidates with lengths of 50mer were selected. The LIS_Identity of each pair of probe and its non-target sequences were calculated by the program findLIS. The Correlation between LIS-Identity and identity are shown in Fig. 3. Fig. 3 compares the identity between the probe and its non-target sequence with the average LIS_Identity that is obtained by the program findLIS. When the identity of probe with its non-target sequence is large, the average LIS_Identity of probe with its non-target sequence is also large. It is obvious when the identity of probe to its non-target sequence over 70%. A threshold sequence similarity of approximately 70% sequence can be considered as the threshold for cross-hybridization [2]. According to Fig. 3, the average LIS_Identity that corresponds to a 70% similarity between the probe and its non-target sequence is about 25mer (indicating a match to the non-target sequence over 25mer).

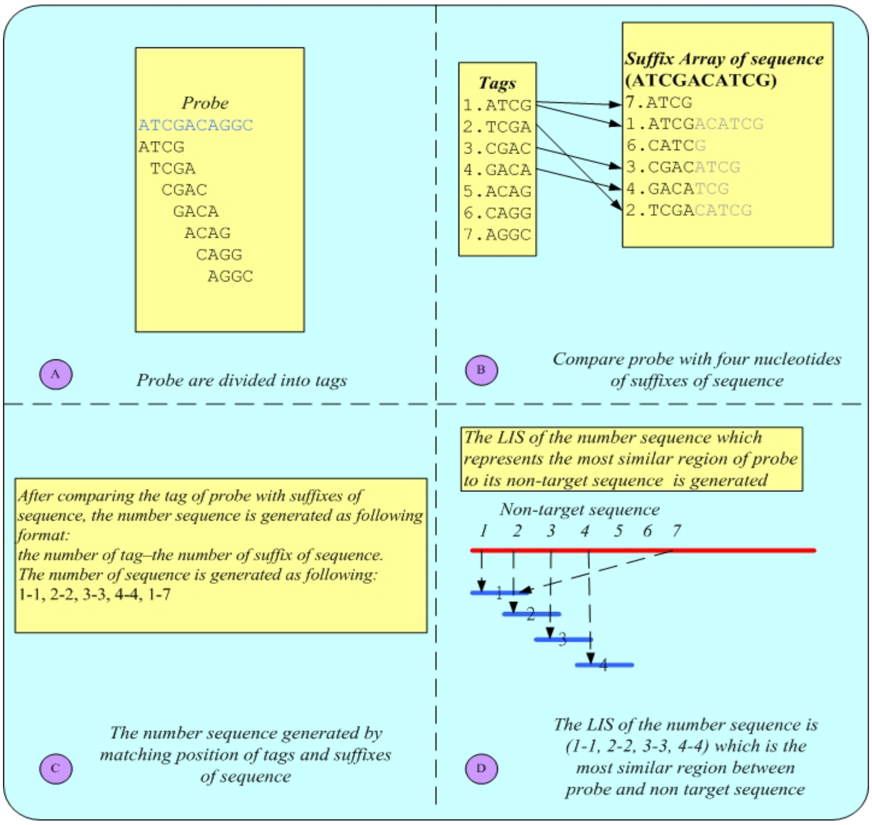

Probe
ATCGACAGGC
ATCG
TCGA
CGAC
GACA
ACAG
CAGG
AGGC
A
Probe are divided into tags
Tags
1.ATCG
2.TCGA
3.CGAC
4.GACA
5.ACAG
6.CAGG
7.AGGC
Suffix Array of sequence
(ATCGACATCG)
7.ATCG
1.ATCGACATCG
6.CATCG
3.CGACATCG
4.GACATCG
2.TCGACATCG
B
Compare probe with four nucleotides of suffixes of sequence
After comparing the tag of probe with suffixes of sequence, the number sequence is generated as following format:
the number of tag–the number of suffix of sequence.
The number of sequence is generated as following:
1-1, 2-2, 3-3, 4-4, 1-7
C
The number sequence generated by matching position of tags and suffixes of sequence
The LIS of the number sequence which represents the most similar region of probe to its non-target sequence is generated
Non-target sequence
1 2 3 4 5 6 7
1
2
3
4
D
The LIS of the number sequence is (1-1, 2-2, 3-3, 4-4) which is the most similar region between probe and non target sequence

Fig. 2. Example of using findLIS program to determine region of greatest similarity between probe and the non-target sequence.

In Fig. 3, we compare the similarity of the probe to its non-target sequence with the average LIS_Identity that is counted by our program findLIS. When the similarity of probe to its non-target sequence is large the average LIS_Identity of probe to its non-target sequence is large also. It is obviously agrees when the similarity of probe to its non-target sequence over 70%. A threshold of around 70% sequence similarity can be considered as cross-hybridization [2]. According to Fig. 3, the average LIS_Identity corresponding to 70% similarity of probe to its non-target sequence is about 25mer (match non-target sequence with 25mer).

5 Discussion and Implementation

Most methods use the BLAST program as the primary tool to avoid cross-hybridization.

They spend much time to calculate the identity of the probe with its non-target sequences. The presented approach first generates the probe candidates in the database and then uses the LIS algorithm to evaluate the identity between the probe and its non-target sequence more efficiently. The database technique and the algorithm can be used to finish the process of designing probes for 100 sequences in three hours. The optimal probes are verified by alignment tool. Because of the efficient algorithm and database technology, the virus probe design can carry out in a web for online using. Although the program findLIS can efficiently calculate the identity of probe with its non-target sequence, in some cases the program findLIS will fail to calculate the identity of probe to its non-target sequence accurately. Three main factors affect the accuracy of LIS_Identity calculation. In the above assessment herein, probes with an LIS_Identity of four to ten are selected, and most have identities with the non-target sequence of below 70%. However, in some cases, the identity of probe with its non-target sequence is exceeds 70%. The first factor is the uncovered region. The length of the tag is set to four, so the tags failed to cover many regions in the non-target sequence several regions in the non-target sequence are not covered by the tags. The program findLIS cannot accurately calculate the identity of probes with its non-target sequence.

Table 1. shows the comparison of different tools and methods for oligonucleotide probe

Journal	Reference sequence	Cross Hybridization detection	Accessibility	Select viral sequence arbitrarily on web	Verificat-ion online	Usage
[15]	human cDNA transcript sequences	BLAST	web (OligoDB)	NO	NO	Detect transcription profiling human gene
[4]	RefSeq, \TIGR	BLAST	Tool (Oligopicker)	N/A	N/A	detect RNA expression
[16]	Virus from GenBank	BLAST, hash technique	Web (VirOligo)	NO	NO	virus sequences identification
[3]	mRNA, CDS, exon	BLAST	Tool (OligoArray)	N/A	N/A	detect gene expression
[8]	TIGR human THC and mouse TC	unique segment	Tool	N/A	N/A	detect gene expression
[7]	virus sequence	Suffix tree (preselection) Tm (Optimal Selection)	Tool	N/A	N/A	virus sequences identification
[5]	Phage genome	suffix array sequence landscape	Web	NO	NO	gene expression
[6]	General	suffix array (LCF)	tool	N/A	N/A	gene expression
[17]	General	jump in LCF, free energy	tool	N/A	N/A	gene expression

Our approac h	virus sequence from GenBank	selection of minimum similarity calculated by the longest increasing subsequence	Web	YES	YES	virus sequences identification

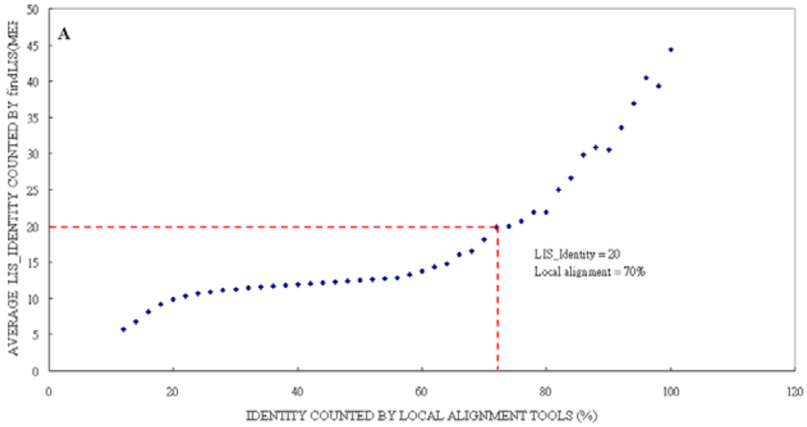
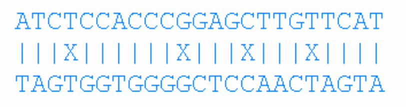


Fig. 3. Correlation between LIS-Identity and similarity

Figure 4 shows the case of failure of calculating identity of probe with its non-target sequence. The actual identity of probe to its non-target sequence is 82% but the LIS_Identity calculated by findLIS is 26%. Some matching regions with shorter than four nucleotides are omitted. However, if the length of the tag decreases, the time for calculating the identity of probe to its non-target sequence will increase. The second factor is multi matching positions in non-target sequence of one tag. When one tag matches more than one position in a non-target sequence, one of 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 calculation of probe with its non-target sequence will be affected. The third factor is the uniqueness of LIS. Unfortunately LIS is not unique in a number sequence. For example, consider a number sequence S = (9, 8, 1, 7, 2, 5, 3). The LIS is either (1, 2, 3) or (1, 2, 5). This maximum LIS_Identity can be found in one of the possible LIS. This study considers only one of the possibilities.



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

Fig. 4. Problem of uncovered region. The LIS_Identity is 6 (26% identity) but the actual identity is 82%.

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

References

1. SantaLucia J, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A*, 95(4):1460-1465, 1998.
2. Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*, 28(22):4552-4557, 2000.
3. Rouillard JM, Herbert CJ, Zuker M OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics*, 18(3):486-487, 2002.
4. Wang X, Seed B Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics*, 19(7):796-802, 2003.
5. Li F, Stormo GD Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics*, 17(11):1067-1076, 2001.
6. Rahmann s Rapid large-scale oligonucleotide selection for microarrays. *In proceeedings of the First IEEE Computer Society Bioinformatics Conference(CSB)* 2002.
7. Kaderali L, Schliep A Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, 18(10):1340-1349, 2002.
8. Chang PC, Peck K Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes. *Bioinformatics*, 19(11):1311-1317, 2003.
9. Buechen-Osmond C, Dallwitz M Towards a universal virus database - progress in the ICTVdB. *Arch Virol*, 141(2):392-399, 1996.
10. Le Novere N MELTING, computing the melting temperature of nucleic acid duplex. *Bioinformatics*, 17(12):1226-1227, 2001.
11. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL Alignment of whole genomes. *Nucleic Acids Res*, 27(11):2369-2376, 1999.
12. Zhang H Alignment of BLAST high-scoring segment pairs based on the longest increasing subsequence algorithm. *Bioinformatics*, 19(11):1391-1396, 2003.
13. Gusfield D: Algorithms on Strings, Trees and Sequences: Cambridge University Press; 1997.
14. Rice P, Longden I, Bleasby A EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, 16(6):276-277, 2000.
15. Mrowka R, Schuchhardt J, Gille C Oligodb--interactive design of oligo DNA for transcription profiling of human genes. *Bioinformatics*, 18(12):1686-1687, 2002.

16. Onodera K, Melcher U VirOligo: a database of virus-specific oligonucleotides. *Nucleic Acids Res*, 30(1):203-204, 2002.
17. Rahmann S Fast and Sensitive Probe Selection for DNA Chips Using Jumps in Matching Statistics. *IEEE Computer Society Bioinformatics Conference (CSB'03)* 2003.