JOURNAL OF CLINICAL MICROBIOLOGY, Aug. 2006, p. 2857–2862 Vol. 44, No. 8
0095-1137/06/$08.00+0 doi:10.1128/JCM.00135-06

Robust Sequence Selection Method Used To Develop the FluChip Diagnostic Microarray for Influenza Virus

Martin Mehlmann,[1]† Erica D. Dawson,[1]† Michael B. Townsend,[1] James A. Smagala,[1] Chad L. Moore,[1] Catherine B. Smith,[2] Nancy J. Cox,[2] Robert D. Kuchta,[1]* and Kathy L. Rowlen[1,3]

Department of Chemistry and Biochemistry, The University of Colorado at Boulder, UCB #215, Boulder, Colorado 80309[1]; Influenza Branch, Centers for Disease Control and Prevention, 1600 Clifton Rd., Atlanta, Georgia 30333[2]; and InDevR, LLC, 2100 Central Ave., Suite 106, Boulder, Colorado 80301[3]

Received 20 January 2006/Returned for modification 10 April 2006/Accepted 14 May 2006

DNA microarrays have proven to be powerful tools for gene expression analyses and are becoming increasingly attractive for diagnostic applications, e.g., for virus identification and subtyping. The selection of appropriate sequences for use on a microarray poses a challenge, particularly for highly mutable organisms such as influenza viruses, human immunodeficiency viruses, and hepatitis C viruses. The goal of this work was to develop an efficient method for mining large databases in order to identify regions of conservation in the influenza virus genome. From these regions of conservation, capture and label sequences capable of discriminating between different viral types and subtypes were selected. The salient features of the method were the use of phylogenetic trees for data reduction and the selection of a relatively small number of capture and label sequences capable of identifying a broad spectrum of influenza viruses. A detailed experimental evaluation of the selected sequences is described in a companion paper. The software is freely available under the General Public License at http://www.colorado.edu/chemistry/RGHP/software/.

In addition to their already widespread use in differential gene expression experiments, DNA microarrays are increasingly being explored for use in diagnostic applications (3, 23, 28). Current applications of interest include the identification of risk for genetic diseases such as cancer, the detection of drug resistance in a wide variety of species, and the identification and subtyping of viral pathogens (28). An ongoing goal in our laboratory is the development of an oligonucleotide microarray for the rapid identification and subtyping of influenza viruses. While previously reported influenza virus microarrays detected DNA made by reverse transcription of viral RNA and amplification by PCR (9, 20), our approach is based on the direct capture and detection of amplified RNA by use of a two-step hybridization process (Fig. 1). The amplification of viral RNA is performed by reverse transcription-PCR, followed by a runoff transcription as described in the companion paper (26).

Several substantial challenges exist in designing capture and label sequences for influenza virus identification. First, one should use limited numbers of capture and label sequences that will "hit" many viral targets belonging to a specific subtype. This situation is different from that encountered in gene expression studies, in which the capture sequences are derived from a single, specified gene with a known sequence. Second, the influenza virus is an RNA virus with a high mutation rate and is therefore a "moving target"; regions of conservation determined at one point in time will likely change as the virus mutates. The high mutation rate requires a rapid, reliable method that can be used to reduce the currently available data set of interest to a set of sequences that comprise a simple, functional array. Third, the postgenomic era has provided rapidly growing databases of publicly available sequence information. In fact, the National Institutes of Health is funding the Influenza Genome Sequencing Project, aimed at making the complete sequences of thousands of influenza viruses rapidly available (www.niaid.nih.gov/dmid/genomes/mscs/default.htm#influenza). As such databases continually grow and change, a systematic method of extracting the desired information from them is required.

Probe design for oligonucleotide microarrays has been the subject of recent reviews (19, 25), and several software tools for the design of microarray probes have been developed. For example, OligoWiz (12, 27) searches for potential probes by taking into account five different parameters: specificity, melting temperature, the position within the transcript, complexity, and self-annealing ability. The user assigns weights to each of these parameters and a sum score is calculated. The program returns the oligonucleotides with the best scores. Other examples of software tools for probe design programs are OligoArray (17, 18), GoArrays (15), or Probe Select (8). In addition, other programs that are not specifically designed for microarray oligonucleotide sequence selection are available but can be used to find and optimize primers, especially for large-scale sequencing purposes. Examples include PRIDE (5), PrimeArray (14), and PRIMO (10).

The objective of most currently available sequence selection tools, such as those mentioned above, is to find primers or probes targeting a single gene within a single organism. In general, sequences are chosen for an experiment on the basis of their specificity for the target, similarity of hybridization conditions, inability to cross-hybridize, and the "coverage" of the genes of interest by the sequence set.

* Corresponding author. Mailing address: Department of Chemistry and Biochemistry, University of Colorado, UCB215, Boulder, CO 80303. Phone: (303) 492-7027. Fax: (303) 492-5894. E-mail: kuchta@colorado.edu.

† M. Mehlmann and E. D. Dawson contributed equally to this work.

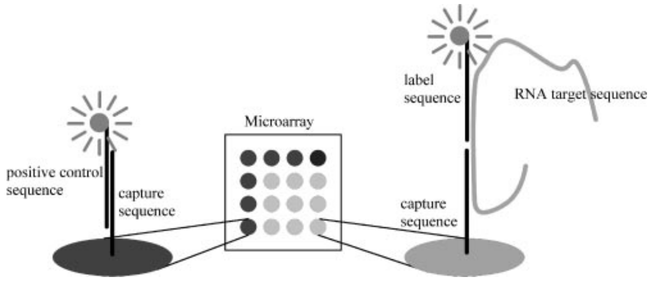

FIG. 1. Scheme of the assay design describing the direct hybridization used for the positive control (left-hand side) and the two-step hybridization process for detection of viral RNA (right-hand side).

For the typing and subtyping of influenza viruses, the objective is more demanding, since the capture and label sequences not only should target a single gene of a specific virus strain but also should target many viruses of the same subtype. To design such capture and label sequences, sequences from a set of virus strains must be examined in order to identify regions that are capable of targeting multiple viruses. Relatively few programs focus on the examination of a set of sequences. Andersson et al. (1a) compared bacterial genomes in an effort to reduce the number of primers required to amplify the genes of two different bacterial genomes by identifying regions of high sequence similarity. However, the algorithm compares only two sequences at a time and does not identify conservation over a large set of sequences. Using PROFILES, Rodriguez et al. (16) calculated "homology profiles" for aligned sequences from foot-and-mouth disease viruses by creating a consensus sequence and recording the number of sequences showing a nucleotide difference from that consensus sequence. These profiles were used to visualize similarities or differences between sequences, and primer pairs were then chosen manually by simply inspecting the "homology profiles."

Primer Premier (PREMIER Biosoft International, Palo Alto, CA) designs primer and microarray sequences for a given set of sequences. A limiting requirement in its application to large databases for highly mutable viruses such as influenza viruses, which often contain incomplete and nonoverlapping regions, is that all sequences in the set must contain data over a specific nucleotide range. In contrast, the method presented herein is more robust, as it allows conserved regions to be identified even when only a fraction of the set contains sequence information at a certain position.

GPRIME (4) is most similar in regards to the aforementioned goals for examining a set of sequences. Beginning with an aligned set of sequences, GPRIME finds homologous regions of a specific length in a data set using an "ambiguity consensus." In the application described by Gibbs et al. (4), the homologous regions were manually selected by examining redundancy values, melting temperatures, gaps, and possible secondary structures. The sequences chosen were compared to those in the EMBL database by using a FASTA search to determine their specificity for the target genomes. Also outlined was a tool that identifies sequence regions where PCR primers could distinguish between two subsets of data by noting differences between the consensus sequences from the two data sets. The sequences chosen were tested for their ability to prime separate reverse transcription-PCRs with RNA extracted from orchid leaves showing symptoms of viral infection. Although these programs are applied to very limited data sets and are not used for microarray applications, they allowed introduction of the idea of the use of a more systematic approach to the selection of capture oligonucleotides for diagnostic applications.

The method for the efficient identification of capture and label pairs described herein is similar to the method used with the GPRIME program, in that it begins with a set of aligned sequences. In contrast to the limited data sets used by the GPRIME program, however, the individual gene-specific databases in this study contained up to 1,000 sequences or more. Conserved regions of a minimum length and a Shannon entropy not exceeding 0.2 for each nucleotide were found by using a "majority consensus" (21). Importantly, the method described here can be used to design microarray probes as well as primers for PCR experiments.

MATERIALS AND METHODS

Implementation and programs used. The BioEdit software package (version 7.0.4.1) was used to visualize sequences (6). Wherever possible, other programs were run as accessory applications within the BioEdit interface. Multiple-sequence alignment was performed by using the Clustal W program (version 1.4) (24). DNADIST (version 3.5c in PHYLIP, version 3.6) was used to create phylogenetic trees. DNADIST was chosen because it uses a fast algorithm that allows phylogenetic trees to be constructed from large data sets in a reasonable amount of time. TreeView (Win32, version 1.6.6) (13) and MEGA3 (version 3.0) (7) were used to display and manipulate the phylogenetic trees. In addition to these existing programs, a number of Python scripts were written and implemented as follows: the label_tree program labels each node in a .dnd file (phylogenetic tree) with a unique integer to facilitate the visualization and subdivision of phylogenetic trees. The dnd2fa program converts the information in a .dnd (or a Newick .nwk) file back to a FASTA file containing sequence information. The fa2fa program allows the contents of one FASTA file to be subtracted from another, outputting a file containing the remaining sequences. The ConFind program identifies conserved regions in a specified data set (ConFind has been described in detail and published elsewhere [21]). The find_oligos program chooses all appropriate capture and label sequences by iteratively walking along the conserved region until the minimum G+C content, melting temperature, and Shannon entropy requirements are met. The pick_oligos program ranks the potential capture and label sequence output from the find_oligos program on the basis of length, Shannon entropy, and melting temperature and chooses the oligonucleotide pairs with the lowest penalty without allowing the nucleotide positions of the oligonucleotides to overlap with other capture-label pairs.

Databases. Sequence information for a large number of influenza viruses was available from the large publicly available database at the Los Alamos National Laboratory (http://www.flu.lanl.gov/) (11) and a smaller database at the Centers for Disease Control and Prevention.

In order to detect sequence similarities to non-influenza virus sequences, i.e., to confirm the specificities of the sequences identified, a BLAST (Basic Local Alignment Search Tool) analysis was performed. The database created for BLAST analysis of the identified sequences contained human genome sequence information from the EST (Expressed Sequence Tags) database and sequence information for several organisms that cause influenza-like illnesses (specifically, influenza B and C viruses, paramyxovirus, rhinovirus, respiratory syncytial virus, *Bacillus anthracis*, coronaviruses, adenoviruses, *Legionella* spp., *Chlamydia pneumoniae*, *Mycoplasma pneumoniae*, and *Streptococcus pneumoniae* [http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5044a5.htm]) from the NCBI nonredundant database (ftp://ftp.ncbi.nlm.nih.gov/BLAST/db/). By default, BLAST uses the top and bottom strands, i.e., the sequence and its reverse complement, to search for sequence similarities in the database; but only the top strand of each capture and label sequence was analyzed by use of BLAST against the sequences in this database. Individual sequences with E values less than 10,000 were considered a "hit"; i.e., they were considered to potentially hybridize to a non-influenza virus sequence.

RESULTS AND DISCUSSION

Overview. The goal of this study was to develop an algorithm for use in the mining of large databases to find potential capture and label sequences that would enable the typing and subtyping of a wide range of different influenza viruses on a microarray.

The microarray assay consisted of short (~25-mer) "capture" DNA oligonucleotides immobilized on a microarray surface, hybridization of influenza virus RNA to the capture sequence, and detection by the hybridization of a fluorophore-conjugated "label" DNA oligonucleotide (~25-mer) to a second region on the target RNA. In addition, several positive control spots in which a capture sequence annealed directly to a complementary label sequence were included in the microarray design for ease of viewing (Fig. 1).

In order to achieve these goals, the capture and label sequences were required to meet a set of defined criteria, as follows: (i) the sequences were specific for a targeted gene segment and showed no cross-reactivity with other capture and label sequences, (ii) the sequences were conserved over a wide range of influenza viruses in order to allow the typing and subtyping of as many different influenza viruses as possible, and (iii) each capture and label sequence was between 16 and 25 nucleotides (nt) in length (these lengths result in a sufficiently high melting temperature and sufficient specificity) and were separated by only 1 nucleotide. As described by Chandler et al. (2), the separation of capture and label sequences by more than 1 nucleotide results in a significant decrease in the fluorescence signal detected. A conserved region of at least 45 nt in length allowed for capture and label sequences within these limits.

Method development. (i) Finding conserved regions. The flowchart shown in Fig. 2 describes the overall process of finding conserved regions for a specific database of interest. From all available sequences, gene-specific databases containing only the sequences of a specific gene and subtype (e.g., influenza A virus, hemagglutinin [HA gene], H1 subtype) were created and converted to the FASTA (a sequence alignment package) format (Fig. 2, step 1). In certain cases, the gene-specific database created was limited by specification of a starting year, especially for viral subtypes or types that predominated during the time period from 2000 to the present, and, as a result, were frequently sequenced. Once the gene-specific database was created, a multiple-sequence alignment was performed with the data set by using ClustalW (Fig. 2, step 2) (24) A multiple-sequence alignment was performed by using the fast algorithm with bootstrap values of 1,000 and a *k*-tuple value of 4. Additionally, a neighbor-joining phylogenetic tree was created. A more rigorous phylogenetic tree prepared by using a maximum-likelihood or parsimony method is possible; however, the neighbor-joining algorithm was chosen due to the large size of the databases and the computational time involved in applying a more rigorous method. The nodes of the phylogenetic tree were arbitrarily numbered to assist with the later division of the tree.

The conserved regions finder (called "ConFind"; Fig. 2, step 4) was written in-house and modeled after the Find Conserved Regions option in BioEdit. A full description of this software can be found elsewhere (21). Briefly, ConFind identifies conserved regions found even when only a fraction of the sequences included contain sequence information at certain positions. The default values were set to a minimum length of 45 nt, 0.2 allowed bits of Shannon entropy per base (with 2 exceptions allowed), and a minimum of 10 sequences. Note, however, that the stringency of these requirements (step 3) was often changed to enable the selection of more or less conserved regions, depending on the particular situation.

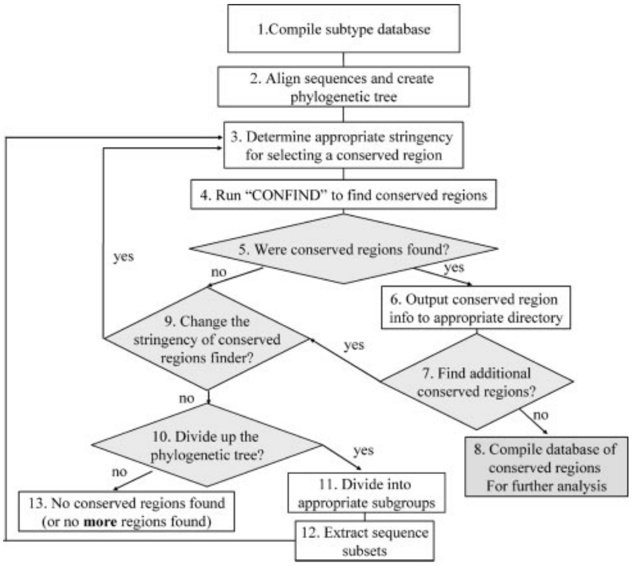
1.Compile subtype database
2. Align sequences and create phylogenetic tree
3. Determine appropriate stringency for selecting a conserved region
4. Run "CONFIND" to find conserved regions
5. Were conserved regions found?
yes
no
yes
6. Output conserved region info to appropriate directory
9. Change the stringency of conserved regions finder?
yes
7. Find additional conserved regions?
no
no
10. Divide up the phylogenetic tree?
yes
no
8. Compile database of conserved regions For further analysis
11. Divide into appropriate subgroups
13. No conserved regions found (or no **more** regions found)
12. Extract sequence subsets

FIG. 2. Flowchart outlining the overall process for finding conserved regions.

ConFind was applied to a gene-specific database by using the default stringency requirements, as noted in step 4 of Fig. 2. If conserved regions were found, information regarding the original sequence information, the positions of the conserved region, and the positional Shannon entropies were output to file, noted in Fig. 2, step 6. If conserved regions were not found, the stringency was loosened and the procedure was repeated.

Often, even when very loose stringency requirements were applied, the genetic variability of influenza virus prevented the identification of conserved regions over an entire gene-specific database (sometimes containing 1,000-plus sequences). The phylogenetic tree was then examined and divided into smaller subtrees, shown as steps 10 and 11 in Fig. 2, in an effort to find additional regions of conservation. This process was not automated, as a number of different criteria could potentially be chosen to determine sequence "difference" or "similarity," such as virus age, geographic region, and host organism. The power of this analysis lies in the fact that the process is very goal specific, and a different desired end goal may result in a different breakdown of the phylogenetic tree.

The subtrees (in the Newick tree format with no sequence information) were extracted from the main tree and converted back to the FASTA format (Fig. 2, step 12) to be used as the subsequent input in step 3. As one of the goals was to capture the largest number of "different" influenza viruses with a limited set of capture and label sequences, the phylogenetic trees were originally broken down into as few subtrees as necessary.

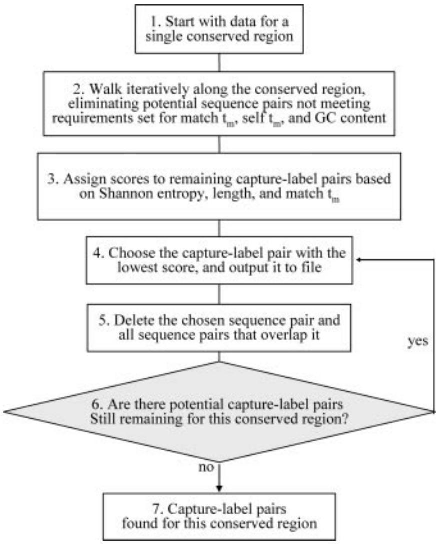
1. Start with data for a single conserved region
2. Walk iteratively along the conserved region, eliminating potential sequence pairs not meeting requirements set for match t_m, self t_m, and GC content
3. Assign scores to remaining capture-label pairs based on Shannon entropy, length, and match t_m
4. Choose the capture-label pair with the lowest score, and output it to file
5. Delete the chosen sequence pair and all sequence pairs that overlap it
yes
6. Are there potential capture-label pairs Still remaining for this conserved region?
no
7. Capture-label pairs found for this conserved region

FIG. 3. Flowchart describing the process of choosing appropriate capture-label pairs from a single conserved region. t_m, melting temperature.

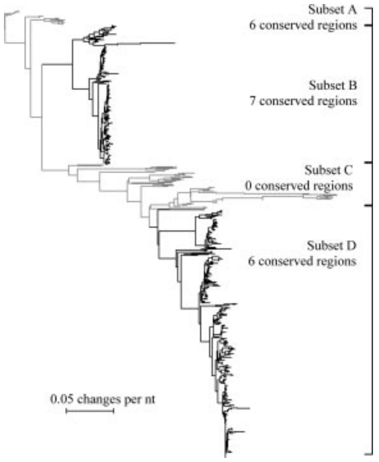

FIG. 4. Neighbor-joining phylogenetic tree for 499 influenza virus A NA (subtype N1) gene segment sequences. The brackets at the right show the initial division of the tree, together with the initial number of conserved regions found for each particular subset.

Once conserved regions that adequately represented the sequences in the gene-specific database examined were found, capture and label sequences were selected.

(ii) Selection of capture and label sequences from conserved regions. While "conservation" of a sequence within a large number of influenza viruses is an important criterion, several other criteria were used in order to optimize selection of capture-label pairs, including secondary structure melting temperatures, G+C content, and length. Initially, 28 capture-label sequences representing influenza A virus HA genes from the H1 and H3 subtypes, influenza A virus neuraminidase (NA) genes of the N1 and N2 subtypes, and influenza A virus M genes were manually selected based on a "score" (described below) that reflected these criteria. The selection routine was then automated for the selection of a much larger pair set.

For automated sequence selection, an additional program (the find_oligos program) that allowed the identification of all possible capture-label pairs within a single conserved region was written. As outlined in Fig. 3, the algorithm walks iteratively, starting at position 1, along the conserved region and searches for pairs of sequences separated by 1 nucleotide. Additional requirements are a length for each sequence between 16 and 25 nt; a minimum melting temperature for the annealing to the reverse complement (matching melting temperatures) for both the label and the capture sequences of 50°C; a maximum melting temperature of 35°C for the most probable secondary structure, as determined by MFOLD (29); and a G+C content of between 30 and 70%. Because of the length range of 16 to 25 nt for each sequence, several pairs with different lengths could be found for each starting position. If several pairs were found, the pair with the highest degree of conservation, i.e., the pair with the lowest maximum Shannon entropy score, was chosen. If several potential capture-label pairs still remained for this start position, the longest one was chosen (Fig. 4, step 2). An additional program, pick_oligos, was written to rank the possible capture-label pairs identified (Fig. 4, step 3) according to the following rules: (i) "good" capture and label pairs should be highly conserved (have a low Shannon entropy), and any highly mutable positions present should be located on separate oligonucleotides (for stability, it is preferable to have two potential mismatches on two separate sequences rather than to have two potential mismatches on a single sequence); and (ii) to improve the stability of the hybridization, longer oligonucleotides with higher melting temperatures are preferred.

The ranking was performed by defining a set of penalties, as outlined in Table 1. To our knowledge, there is no detailed understanding of the combined effects that numerous variables such as the criteria shown in Table 1 have on hybridization to surface-bound oligonucleotides. As a result the penalty values were chosen empirically so that the ranking results of the pick_oligos program on a test data set matched the results of a manual ranking performed by a skilled researcher. The pick_oligos program chose the capture-label pair with the lowest penalty and removed capture-label pairs that had a sequential overlap with the chosen pair (Fig. 4, steps 4 and 5). This process was repeated with all possible capture-label pairs.

Method implementation. A total of 4,917 influenza virus sequences were divided into 15 different smaller gene-specific databases, as shown in Table 2, representing different gene specific subtypes (e.g., subtypes H1, N1, and N3). Databases containing very large numbers of sequences (>1,000) were generally reduced by investigating viruses recovered only relatively recently, which is reasonable, considering the rapid evolutionary nature of influenza virus. ConFind found conserved regions by use of the gene-specific database; but if none were found, the database was divided into smaller subtrees, as discussed later. The total numbers of conserved regions for each gene-specific database are shown in Table 2.

An integral and unique aspect of the method used to find capture and label pairs was the breakdown of the original gene-specific database into several smaller subsets. Depending on the research objectives, the breakdown can be conducted according to the use of a large number of different criteria,

TABLE 1. Empirical penalties assigned to potential capture-label pairs for final sequence selection

Criterion[a]	Assigned penalty value	Explanation, notes
Total Shannon entropy penalty	10 (E1 + E2)	
E1 > 0.1	15	Extra penalty for high mismatch probability
E2 > 0.1	15	
Both E1 and E2 on the same oligonucleotide	10	The presence of E1 and E2 on separate oligonucleotides is preferred to minimize potential mismatches
E1 and E2 > 0.1 and both E1 and E2 on the same oligonucleotide	20	
T_m	$1/T_m$ (°C)	Higher T_m preferred
Length	1/length (nt)	Longer sequence preferred

[a] E1 and E2, the two highest Shannon entropies within the capture-label pair examined; T_m, melting temperature.

such as phylogenetic lineage, virus age, the geographic region of origin, the host species, or sample pretreatment. For the influenza virus microarray, each gene-specific database was subdivided according to phylogenetic information, as there is likely a connection between phylogenetic information and antigenicity (22). As an example, the breakdown of the tree for the N1 subtype of the NA gene of influenza A virus is shown in Fig. 4. In this example, by using the parameters described in the "Finding conserved regions" section, no conserved regions were found for the complete set of 499 subtype N1 sequences. A visual inspection of the phylogenetic tree suggested a logical breakdown into four smaller subsets, which were analyzed separately. Subset A consisted of 16 H1N1 sequences, most of which (14) were the sequences of strains circulating in humans between 1933 and 1947. The two exceptions were A/swine/Korea/S175/2004 and A/swine/Korea/S10/2004, respectively. Six conserved regions were found for this subset. Subset B (156 sequences) contained, with only a few exceptions, sequences from H1N1 viruses that infected humans within the last 10 years. The few exceptions observed were H1N1 viruses recovered from 1977 to 1992 (A/Hokkaido/2/92, A/Yamagata/32/89, A/Swine/Obihiro/5/92, A/Hokkaido/11/88, A/Yamagata/120/86, A/Chile/1/83, A/Fiji/15899/83, A/camel/Mongolia/82, A/USSR/90/77A, A/USSR/90/77B, and A/Kiev/59/79). Seven conserved regions were found for this subset. Subset C (51 sequences) contained 40 H1N1 sequences collected from swine and avian viruses. The remaining 11 sequences were from viruses of other subtypes (subtypes H5N1, H6N1 H7N1, and H9N1). Due to the large genetic diversity, no conserved regions were initially found for subset C. Subset D contained 276 sequences for viruses collected in the last 8 years and consisted entirely of species from Eurasia. It contained 258 H5N1, 13 H6N1, 2 H7N1, and 2 H9N1 sequences and 1 H11N1 sequence. While the H5N1 strains were mostly circulating in avian species, subset D also contained 31 avian H5N1 strains that had been contracted by humans. Both Eurasian and North American avian lineages were represented. A total of six conserved regions were found for subset D. As subsets B and D both contained sequence information from viruses that recently infected humans, these subsets were further evaluated in a manner similar to that described for the initial breakdown. Subset C was also further analyzed, as no conserved regions were found initially.

TABLE 2. Description of original influenza virus sequence databases and results from application of conserved region and sequence selection methods described

Database		Yr included[a]	Species included	Total no. of sequences in database	No. of conserved regions found	No. of capture-label pairs found
Influenza virus type	Gene segment (subtype[b])					
A	HA (H1)	2000 or later	Swine, bird, human, camel	230	10	7
A	HA (H2)	All	Human, bird	110	19	15
A	HA (H3)	2000 or later	Swine, bird, human, equine	850	107	65
A	HA (H5)	2000 or later	Swine, bird, human, leopard, tiger, equine	248	45	27
A	HA (H7)	1998 or later	Bird	156	15	15
A	HA (H9)	All	Bird	326	17	13
A	NA (N1)	All	Swine, bird, human, leopard, tiger	499	133	106
A	NA (N2)	All	Swine, bird, human	1012	40	28
A	NA (N3)	All	Bird	44	15	25
A	NA (N7)	All	Bird, equine	9	9	8
A	NP	All	Swine, bird, human	487	53	43
A	MP	2000 or later	Swine, human, bird	540	77	41
B	HA	All	Human	343	66	39
B	MP	All	Human	31	11	7
B	NP	All	Human	32	12	8
Total				4,917	629	447

[a] The year indicated is the earliest year of virus isolation, whereas "all" indicates that sequences from all available years were included in the analysis.
[b] If applicable.

Evaluation of potential interferences. The final step in selecting capture and label sequences for the identification of influenza virus was to search for potential cross-hybridizations by using the BLAST program (1). This required an additional database that contained sequences from potentially interfering species that might be present in the target RNA hybridization mixture and that might also hybridize to the identified capture and label pairs, resulting in false-positive signals. Since it was impractical to analyze all available genomes with the BLAST program, a smaller database was created to include human mRNA and genomes from other microorganisms that cause influenza-like illnesses, as well as the genomes for influenza B and C viruses (as described in the Materials and Methods section). Because of the two-step hybridization, false-positive signals from nontarget organisms can be observed on a microarray only if one of the capture sequences together with any of the label sequences hybridizes to the same gene. Thus, if a capture sequence was found to "hit" a gene within the database (as described in the "Databases" section), a second level of comparison was conducted to check whether a label sequence also hit the same gene. If both capture and label sequences were found to hit the same gene, the sequence was discarded as a possible source of false-positive signals on the microarray.

From the 629 conserved regions identified from all of the influenza virus sequence databases accessed, a total of 447 potential capture-label pairs (Table 1) were selected after application of the find_oligos and pick_oligos programs. From these 447 capture-label pairs, 75 pairs with the best scores that represented different types and subtypes were chosen for initial experimental evaluation, as follows: influenza A virus HA genes of the H1, H3, and H5 subtypes; influenza NA genes of the N1 and N2 subtypes; the M genes of both influenza type A and type B viruses; and influenza B virus HA and NP genes. Together with the 28 manually chosen sequences (as described in the "Selection of capture and label sequences from conserved regions" section), a total of 103 capture-label pairs were experimentally evaluated. The sequences identified by this method and refined experimentally are shown in Table 1 of reference 26.

ACKNOWLEDGMENTS

We gratefully acknowledge funding from the National Institute of Allergy and Infectious Diseases, NIH (grant U01 AI056528-03).

We thank Rebecca Garten at CDC for her assistance.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.

1a. **Andersson A., R. Bernander, and P. Nilsson.** 2005. Dual-genome primer design for construction of DNA microarrays. Bioinformatics **21:**325–332.

2. **Chandler, D. P., G. J. Newton, J. A. Small, and D. S. Daly.** 2003. Sequence versus structure for the direct detection of 16S rRNA on planar oligonucleotide microarrays. Appl. Environ. Microbiol. **69:**2950–2958.

3. **Clewley, J. P.** 2004. A role for arrays in clinical virology: fact or fiction? J. Clin. Virol. **29:**2–12.

4. **Gibbs, A., J. Armstrong, A. M. Mackenzie, and G. F. Weiller.** 1998. The GPRIME package: computer programs for identifying the best regions of aligned genes to target in nucleic acid hybridization-based diagnostic tests, and their use with plant viruses. J. Virol. Methods **74:**67–76.

5. **Haas, S., M. Vingron, A. Poustka, and S. Wiemann.** 1998. Primer design for large scale sequencing. Nucleic Acids Res. **26:**3006–3012.

6. **Hall, T. A.** 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41:**95–98.

7. **Kumar, S., K. Tamura, and M. Nei.** 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings Bioinformatics **5:**150–163.

8. **Li, F., and G. D. Stormo.** 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics **17:**1067–1076.

9. **Li, J., S. Chen, and D. H. Evans.** 2001. Typing and subtyping influenza virus using DNA microarrays and multiplex reverse transcriptase PCR. J. Clin. Microbiol. **39:**696–704.

10. **Li, P., K. C. Kupfer, C. J. Davies, D. Burbee, G. A. Evans, and H. R. Garner.** 1997. PRIMO: a primer design program that applied base quality statistics for automated large-scale DNA sequencing. Genomics **40:**476–485.

11. **Macken, C., H. Lu, G. J. Goodman, and L. Boykin.** 2001. The value of a database in surveillance and vaccine selection, p. 103–106. *In* A. D. M. E. Osterhaus, N. Cox, and A. W. Hampson (ed.), Options for the control of influenza IV. Elsevier Science, Amsterdam, The Netherlands.

12. **Nielsen, H. B., R. Wernersson, and S. Knudsen.** 2003. Design of oligonucleotides for microarrays and perspectives for design of multi-transcriptome arrays. Nucleic Acids Res. **31:**3491–3496.

13. **Page, R. D.** 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:**357–358.

14. **Raddatz, G., M. Dehio, T. F. Meyer, and C. Dehio.** 2001. PrimeArray: genome-scale primer design for DNA-microarray construction. Bioinformatics **17:**98–99.

15. **Rimour, S., D. Hill, C. Militon, and P. Peyret.** 2005. GoArrays: highly dynamic and efficient microarray probe design. Bioinformatics **21:**1094–1103.

16. **Rodriguez, A., E. Martinez-Salas, J. Dopazo, M. Davila, J. C. Saiz, and F. Sobrino.** 1992. Primer design for specific diagnosis by PCR of highly variable RNA viruses: typing of foot-and-mouth disease virus. Virology **189:**363–367.

17. **Rouillard, J.-M., C. J. Herbert, and M. Zuker.** 2002. OligoArray: genome-scale oligonucleotide design for microarrays. Bioinformatics **18:**486–487.

18. **Rouillard, J.-M., M. Zuker, and E. Gulari.** 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. **31:**3057–3062.

19. **Russell, R.** 2003. Designing microarray oligonucleotide probes. Briefings Bioinformatics **4:**361–367.

20. **Sengupta, S., K. Onodera, A. Lai, and U. Melcher.** 2003. Molecular detection and identification of influenza viruses by oligonucleotide microarray hybridization. J. Clin. Microbiol. **41:**4542–4550.

21. **Smagala, J. A., M. Mehlmann, M. B. Townsend, E. D. Dawson, R. D. Kuchta, and K. L. Rowlen.** 2005. ConFind: a robust tool for conserved sequence identification. Bioinformatics **21:**4420–4422.

22. **Smith, D. J., A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. M. E. Osterhaus, and R. A. M. Fouchier.** 2004. Mapping the antigenic and genetic evolution of influenza virus. Science **305:**371–376.

23. **Striebel, H.-M., E. Birch-Hirschfeld, R. Egerer, and Z. Foldes-Papp.** 2003. Virus diagnostics on microarrays. Curr. Pharm. Biotechnol. **4:**401–415.

24. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.

25. **Tomiuk, S., and K. Hofmann.** 2001. Microarray probe selection strategies. Briefings Bioinformatics **2:**329–340.

26. **Townsend, M. B., E. D. Dawson, M. Mehlmann, J. A. Smagala, D. Dankbar, C. L. Moore, C. B. Smith, N. J. Cox, R. D. Kuchta, and K. L. Rowlen.** 2006. Experimental evaluation of the FluChip diagnostic microarray for influenza virus surveillance. **44:**2863–2871.

27. **Wernersson, R., and H. B. Nielsen.** 2005. OligoWiz 2.0—integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Res. **33:**W611–W615.

28. **Zammatteo, N., S. Hamels, F. de Longueville, I. Alexandre, J.-l. Gala, F. Brasseur, and J. Remacle.** 2002. New chips for molecular biology and diagnostics. Biotechnol. Annu. Rev. **8:**85–101.

29. **Zuker, M., D. H. Mathews, and D. H. Turner.** 1999. Presented at the NATO Science Series, 3: High Technology.