Published online 13 December 2007

Nucleic Acids Research, 2008, Vol. 36, No. 1 ***e3***
doi:10.1093/nar/gkm1106

Comprehensive viral oligonucleotide probe design using conserved protein regions

Omar J. Jabado[1], Yang Liu[2], Sean Conlan[1], P. Lan Quan[1], Hédi Hegyi[3], Yves Lussier[4], Thomas Briese[1], Gustavo Palacios[1] and W. I. Lipkin[1,*]

[1]Center for Infection and Immunity, Mailman School of Public Health, Columbia University, 722 West 168th Street, Room 1801, New York, NY 10032, [2]Sigma-Aldrich, Research Biotech, 2909 Laclede Ave, St. Louis, MO 63103, USA, [3]Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, H-1518 Budapest, PO Box 7, Hungary and [4]Center for Biomedical Informatics, Department of Medicine, Section of Genetic Medicine, University of Chicago, 5841 South Maryland Ave, AMB N660B, Chicago, IL 60637, USA

Received October 15, 2007; Revised November 26, 2007; Accepted November 27, 2007

ABSTRACT

Oligonucleotide microarrays have been applied to microbial surveillance and discovery where highly multiplexed assays are required to address a wide range of genetic targets. Although printing density continues to increase, the design of comprehensive microbial probe sets remains a daunting challenge, particularly in virology where rapid sequence evolution and database expansion confound static solutions. Here, we present a strategy for probe design based on protein sequences that is responsive to the unique problems posed in virus detection and discovery. The method uses the Protein Families database (Pfam) and motif finding algorithms to identify oligonucleotide probes in conserved amino acid regions and untranslated sequences. *In silico* testing using an experimentally derived thermodynamic model indicated near complete coverage of the viral sequence database.

INTRODUCTION

The capacity of DNA microarrays to simultaneously screen for hundreds of viral agents makes them an attractive supplement to traditional methods in microbiology. Their utility has been demonstrated through detection of papilloma virus in cervical lesions (1), SARS coronavirus in tissue culture (2), parainfluenza virus 4 in nasopharyngeal aspirates (3), influenza from nasal wash and throat swabs (4,5), gammaretrovirus in prostate tumors (6), coronaviruses and rhinoviruses from nasal lavage (7), metapneumovirus from bronchoalveolar lavage (8), filoviruses and malarial parasites in blood in hemorrhagic fever (9), and a wide variety of respiratory pathogens in nasal swabs and lung tissue (10).

Viral microarrays have increased in density and strain coverage as fabrication technologies have improved. cDNA pathogen arrays derived from reference strain nucleic acids (11,12) have been replaced by oligonucleotide arrays due to their increased flexibility. Oligonucleotide design strategies have focused on pairwise sequence comparisons to identify conserved regions within a variety of viral pathogens (13–15). Multiple alignments have been used to design probes for clinically important virus genera, e.g. rotaviruses (16), orthopoxviruses (17) or influenzaviruses (18). Viral resequencing arrays have recently been introduced that allow single nucleotide resolution (4,19–21). Although such tiling arrays enable accurate typing, the number of probes required to build a resequencing array for all viral sequences exceeds current art.

A comprehensive viral microarray should address the entire viral sequence database. Pairwise nucleic acid comparisons, while rapid, do not scale well with sequence number and ignore valuable coding information. Non-overlapping segments, heterogeneous sizes and the large number of sequences preclude automated multiple alignments of nucleic acids for probe design. Protein–protein comparisons are more sensitive for detecting conserved regions due to the power of substitution matrices (22); however, at the time of writing, no reported oligonucleotide design algorithm leverages this information.

The Protein Families Database (Pfam) (23) is a repository of hand curated protein multiple alignments and Hidden Markov Models (HMMs) across all phylogenetic kingdoms. HMMs are probabilistic representations of protein alignments that are well suited to identifying homologies (24,25). Beginning with the Pfam database as a foundation, we established a tiered method for creating viral probes that uses all sequence information available for viruses. Our method for probe design employs protein alignment information, discovered protein motifs, nucleic acid motifs and finally, sliding windows to ensure near complete coverage of the database.

*To whom correspondence should be addressed. Tel: +1 212 342 9031; Fax: +1 212 342 9044; Email: wil2001@columbia.edu

MATERIALS AND METHODS

Exploratory array design and hybridization

We pursued experiments to determine the effects of probe-target mismatch and background nucleic acid concentration on array sensitivity and specificity; results were used to derive parameters for probe design. West Nile virus RNA (WNV, strain New York 1999, AF202541) was used as template in hybridization experiments on an Agilent oligonucleotide array with 1131 complementary probes of length 60 nucleotides (nt). Approximately one third of the probes had between 1 and 20 randomly introduced mismatches. The plus and minus (reverse complement) strands of each sequence were deposited, in duplicate. In addition to the flaviviral specific probes, the array contained nearly 36 500 probes for other viral families, negative and positive controls. A volume containing 10^6 copies of WNV and 200 ng of background nucleic acid (human lung tissue RNA) was amplified using random primers and hybridized in four replicate experiments as previously described (10).

Analysis of impact of probe-template mismatches on fluorescence

Hybridizing a WNV isolate of known sequence allowed prediction of probe-viral hybrid strength and correlation to fluorescence data. To predict hybrids with high accuracy, Smith–Waterman alignments of the virus sequence against microarray probes were generated using the EMBOSS bioinformatics suite (26). The number of mismatches was calculated for each expected probe-target pair. The change in Gibbs free energy at 65°C (hybridization temperature) was calculated using PairFold version 1.7 (27) as a separate measure of probe-template binding strength. PairFold employs a dynamic programming algorithm to compute the minimum free energy structure (excluding pseudo-knots); the standard free energy model is used (28) with empirical nearest neighbor energies (29). The arrays were visualized with an Agilent slide scanner, then processed with the quantile normalization technique (30). SPSS version 14 was used for statistics and data plots (http://www.spss.com/), fluorescence data is available as supplementary material.

Viral sequence database

The EMBL nucleotide sequence database [July 2007, Release 91; 461,353 nucleic acid sequences (31)] was chosen as the reference for this study because it is tightly integrated with the Pfam protein family database (23,32). We incorporated all viral genomes in the NCBI Reference Sequence (RefSeq) project on 21 May 2007 (1701 unique viruses; 2790 genome segments). EMBL Release 91 included 183 924 HIV-1 sequences (39.8% of the database, classified by NCBI Taxid 11676). Of these, 151 878 HIV-1 sequences were ≤1 kb, (median length 440 nt). To simplify computation, we selected only those 1915 HIV-1 sequences >7000 nt in length (full genome is 9200 nt). The final dataset contained 281 632 nucleic acid sequences.

Taxon growth was estimated using a standard least squares method, with the SPSS statistical package. A non-redundant database comprising 74 044 sequences was generated with CD-Hit (33), using a similarity cutoff of 98% to define sequences as identical. Bacteriophages were not included in the analysis; however, data were retained to allow probe design using the EMBL phage database.

Extraction of conserved regions and nucleic acid sequence from Pfam-A alignments

The Pfam database is comprised of hand curated seed protein alignments that are converted to a probabilistic representation using HMMs. These HMMs are used to search the protein database for homologues that can be added to the seed to create a comprehensive alignment (23,24). Pfam domains were analyzed to identify short, conserved protein regions and corresponding nucleic acid sequences. In the first step, the log-odds score for each position of the HMM built from the seed alignment was summed; lower scores were considered to indicate conservancy. The most conserved, non-overlapping 20 amino acid (aa) regions were identified. In the second step, protein alignments of all Pfam-A families were extracted and mapped to their underlying nucleotide sequences by cross reference to the EMBL records. HMM parsing modules from the BioPerl package were used. In the third step, the underlying nucleotide sequences were extracted and stored. In cases where the region contained gaps, flanking nucleotides were brought together to yield sequences of length 60. These sequences formed the basis for downstream probe design. Domain alignments in the Pfam-B were not used in probe design because they are of lower quality; also, as domain quality improves these alignments will be integrated into Pfam-A (23).

Motif finding for non-Pfam coding sequences

All coding nucleic acid sequences that were not part of a Pfam-A alignment were extracted. In this step, the most conserved regions within homologous genes were identified for probe design. Sequences were clustered at the protein level with CD-Hit, using a similarity threshold of 80%. All sequence clusters were subjected to a MEME motif search (34) using the following parameters: motif width of 20, zero or one motif allowed per sequence, a minimum of two sequences per motif. Three motifs were selected for each sequence cluster. The underlying nucleic acid sequence extracted for each protein motif was used for probe design.

A sliding window approach was used for highly divergent sequences that did not share any motifs. Using the PAM250 matrix (35) a summed log odds scores for every 20 aa subsequence in the protein was calculated; the three least likely to vary (lowest log odds score) were selected as regions for probe design.

Motif finding for non-coding regions and unannotated sequences

Viruses often have highly conserved non-coding regions at the termini of their genomes or genome segments that serve critical roles in replication, transcription, and packaging. We reasoned that probes based in these

regions may be useful in microarray design. We identified conserved probes across homologous regions in sequences annotated as 5′ UTR, 3′ UTR, LTR, and those without annotation. Sequences were first clustered at the 80% threshold. Clustered sequences were then subjected to a motif search using the same parameters employed for proteins, except that a length of 60 nt per motif was specified. We addressed sequences that did not contain a shared motif separately; three non-overlapping 60 nt subsequences were chosen as probes.

Probe selection and minimization with set cover algorithm

An algorithm was designed to automate identification of the minimum set of probes required to address a repertoire of potential viral targets (36). In the first step of analysis, the number of mismatches between a probe and its viral target was computed; the algorithm considered a probe to be 'covering' if it had $\leq$5 mismatches to the template. Coverage data were converted to a matrix of binary values. A greedy algorithm was implemented to choose a probe combination from the matrix, minimizing the number required probes. Candidate probes were further screened to ensure a T_m >60°C, no repeats exceeding a length of 10 nt, no hairpins with stem lengths exceeding 11 nt, and <33% overall sequence identity to non-viral genomes.

Analysis of coverage by Gibbs free energy

Because it is not feasible to test all probes with all known viruses, we tested probe validity using a Gibbs free energy model of hybridization. All probe sequences were compared to the non-redundant set of viral sequences by BLASTN (37). Probe-target pairs were aligned by Smith–Waterman to ensure accuracy; mismatches and change in Gibbs free energy at 65°C (hybridization temperature) were then calculated.

Design of viral genome tiling probes

To gauge the performance of our probe selection algorithm, another comprehensive method was devised that used only nucleic acid sequence. Sequences in the Reference Sequence Viral Genomes project (38) are evenly distributed among viral families; therefore, we reasoned that probes derived from these sequences would provide broad coverage. To contrast with our method, we selected 60 nt oligonucleotides end-to-end along all viral genomic RefSeq sequences (1701 viruses). This resulted in a tiling probe-set where the length of a sequence was proportional to the number of interrogating probes.

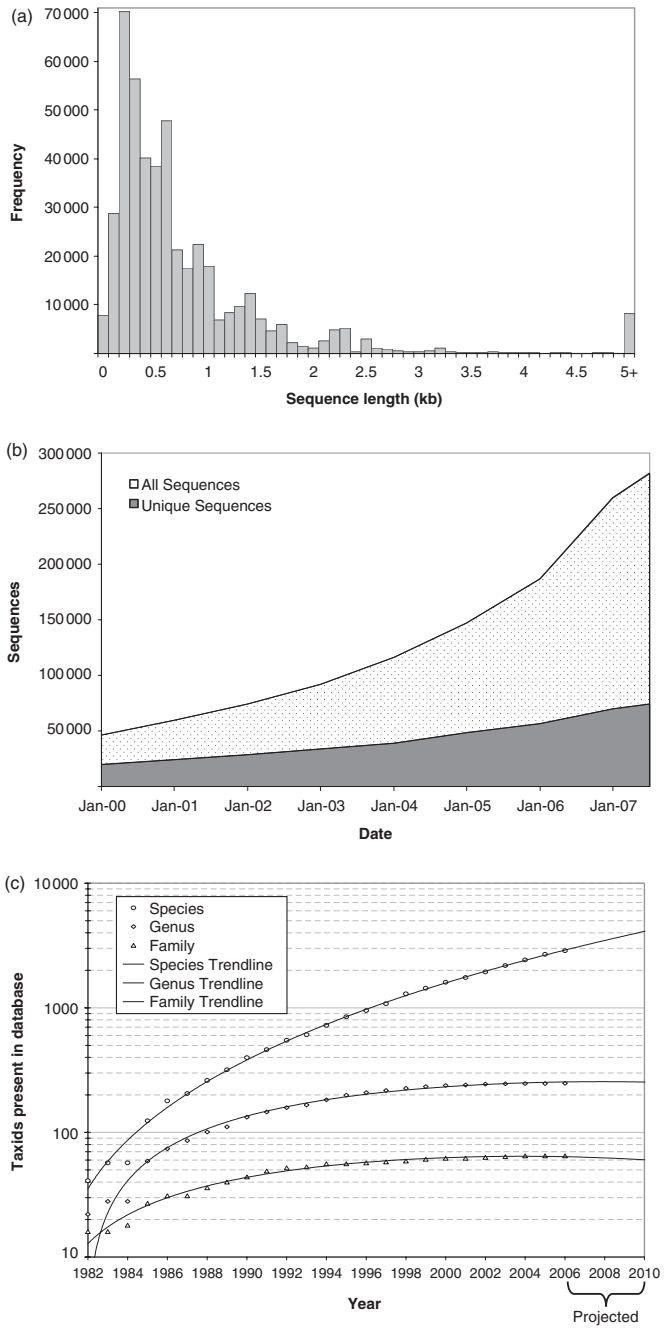

Figure 1. Characteristics of the viral database. (**a**) Distribution of viral sequence length in EMBL nucleotide database July 2007, all sequence ≥5 kb were grouped together. (**b**) Growth of sequence diversity over time in the HIV-1 filtered database; unique sequences were identified by sequence clustering the 98% similarity level. (**c**) Growth of taxonomic classes over time; linear regression was used to project growth from year 2006 to 2010.

RESULTS

The viral sequence database is dominated by gene fragments

Sequence information has grown rapidly since the advent of modern capillary sequencing techniques; 461 353 viral sequences are currently available [EMBL release 91 (31)]. Research into human pathogens has driven database growth, as indicated by the high number of retroviral (204 668, 44%), flaviviral (71 531, 16%) and orthomyxoviral (47 248, 10%) sequences. The high number of retroviral sequences reflects the importance of HIV-1 *gag* and *pol* gene sequences in clinical management of HIV infection (39).

We queried the EMBL viral database to determine the frequencies of coding sequences and full genomic sequences. The majority of viral sequences were <1 kb

fragments (Figure 1a). Approximately 84% of sequences (385 951, 83.7%) were annotated as coding for a single gene; 7.2% coded for two or more genes. Of all coding sequences, ~80% (306 933, 79.5%) were annotated as partial gene sequences (EMBL Release 91, July 2007). As of May 2007, the Reference Sequence Viral Genomes project (38) has catalogued 1701 viral strain representatives (2790 total genome segments). Overall, 1.6% (7662) of viral sequences are annotated as complete genomes. Of the 275 known viral genera, nearly all had at least one complete genomic sequence. All genera containing vertebrate pathogens had at least one fully sequenced genome.

The viral sequence database oversamples HIV-1 and is highly redundant

HIV-1 sequences represent 44% of the current EMBL viral database; 83% of these sequences are less than 1 kb in length. A viral microarray that mirrored the database would oversample HIV and undersample other viruses. Thus, we elected use only the near complete genome sequences of HIV-1, for which exist representatives of each known group and subtype (40). The filtered sequence database comprising 281 632 sequences was generated from EMBL Release 91 by excluding HIV-1 sequences <7000 nt in length. This filtered database was subsequently used for probe design.

A commonly used method to reduce sequence complexity is generation of a non-redundant sequence set by clustering (33). We grouped sequences at the 98% identity level and selected the longest sequences as unique representatives of each group. This method was used to assess the growth of sequence diversity between January 2000 and the current release of July 2007. The database grew 600% in the 7-year period; doubling every three years. Unique sequences decreased as a proportion of the database, from 42% to 27%; overall growth of unique sequences was 378% (Figure 1b). The current database comprised 74 044 unique sequence representatives at the 98% similarity level. Thus, the growth in the number of sequences in the viral database has been rapid, while growth in diversity has been more modest.

One hypothesis to explain this slower growth of sequence diversity is that many of the existing viruses infecting humans have already been discovered and new isolates deposited are variants of well studied viruses. We charted the growth of viral taxonomic groups as a function of time to visualize trends in viral discovery (Figure 1c). The number of families and genera has remained stable since 1996; however, the number of sequences that have been classified as a new species has steadily risen. A least squares fit of this growth indicates that the steady increase in new species characterization is likely to continue, while the discovery of new viral families will be less common.

A tiered, protein-motif-based approach to probe design addresses all viral sequences

Nucleotide sequences were divided into four subtypes: (i) coding sequences that corresponded to Pfam-A alignments (cPf), (ii) coding sequences not in the Pfam-A (cNPf), (iii) sequences that were annotated as untranslated regions (UTR) or long terminal repeats (LTR) and (iv) sequences that were unannotated (UA). We sought to match the quality of Pfam-A alignments in the non-Pfam coding sequences by clustering them into groups of related sequences, approximating homologous genes. These were then subjected to a protein motif finding program to identify the conserved regions within each cluster. The untranslated and unannotated sequences were subjected to a similar clustering analysis, but at the nucleotide level.

All four subtypes were subjected to the same three step design method: identification of conserved regions, extraction of nucleotide probe sequences, and minimization of covering probes. By allowing a limited number of mismatches to cognate templates, the number of probes required can be reduced. The mismatch threshold was determined based on experiments with West Nile virus (strain New York 1999, AF202541) that indicated high, homogenous fluorescence signal was observed if probes had five or fewer mismatches to the viral template (Figure 2). The probe minimization technique serves to lower microarray printing costs and simplify analysis while maintaining sequence coverage. A flowchart of the design method is depicted in Figure 3.

The most recent Pfam-A release (Version 22) comprised 9318 families, of which 1540 had viral members. Of 405 543 annotated protein sequences with length >20 aa, 278 119 (68.6%) belonged to a Pfam-A family, while 127 424 (31.4%) did not. Three probes were chosen for each gene, yielding a total of 104 467 cPf and 133 513 cNPf probes. Of sequences not contained in Pfam-A, only 5.6% (6956) were found in Pfam-B alignments. Thus, due to the lower quality of alignments (23) and poor viral representation, the Pfam-B was not used for probe design. The 12 428 untranslated regions processed yielded 4616 probes. For the 24 841 unannotated sequences processed, 13 740 probes were designed. Sequences that were not covered due to high/low GC%, low complexity, repetitive sequence or a preponderance of ambiguous nucleotides (4244) were processed with a sliding window strategy; 14 530 probes were designed. Overall, the number of probes required to address all viral sequences was 270 866. Sequence counts and probe counts for the most recent EMBL/Pfam release are detailed in Figure 4. An example of typical probe distribution is shown with respect to the Dengue virus 1 genome (NC_001477; Figure 5).

Validation of probes by predicted affinity to targets

Probe sequence composition is a major determinant of hybridization signal and is responsible for much of the variance between probes that target the same nucleic acid strand. Probe-target thermodynamics have been successfully modeled to predict fluorescence (41,42), control for variance (43) and even estimate concentrations of target detected in samples (44). Observing that some probes with more than five mismatches to their targets showed strong fluorescent signal, we concluded that sequence composition is a major factor in our array platform.

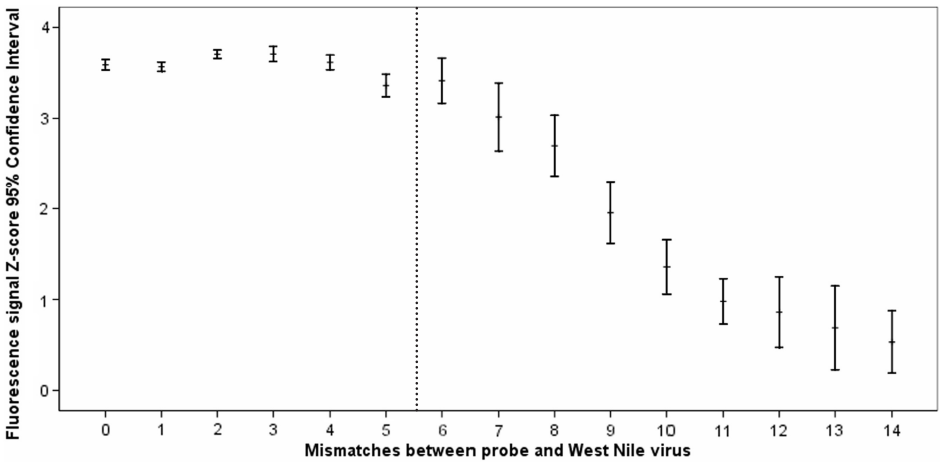

Figure 2. Impact of mismatches on fluorescence signal in microarray hybridization. West Nile virus (New York 1999 strain RNA) at 10^6 copies was spiked into 200 ng of human lung (background) RNA. Total nucleic acid was amplified, labeled and hybridized. After normalization of replicate arrays, $\log_2$ fluorescence was converted to Z-Scores. 95% confidence intervals of the mean for probes with the same number of mismatches were plotted. Dotted line indicates maximum number of mismatches yielding an acceptable fluorescence signal.

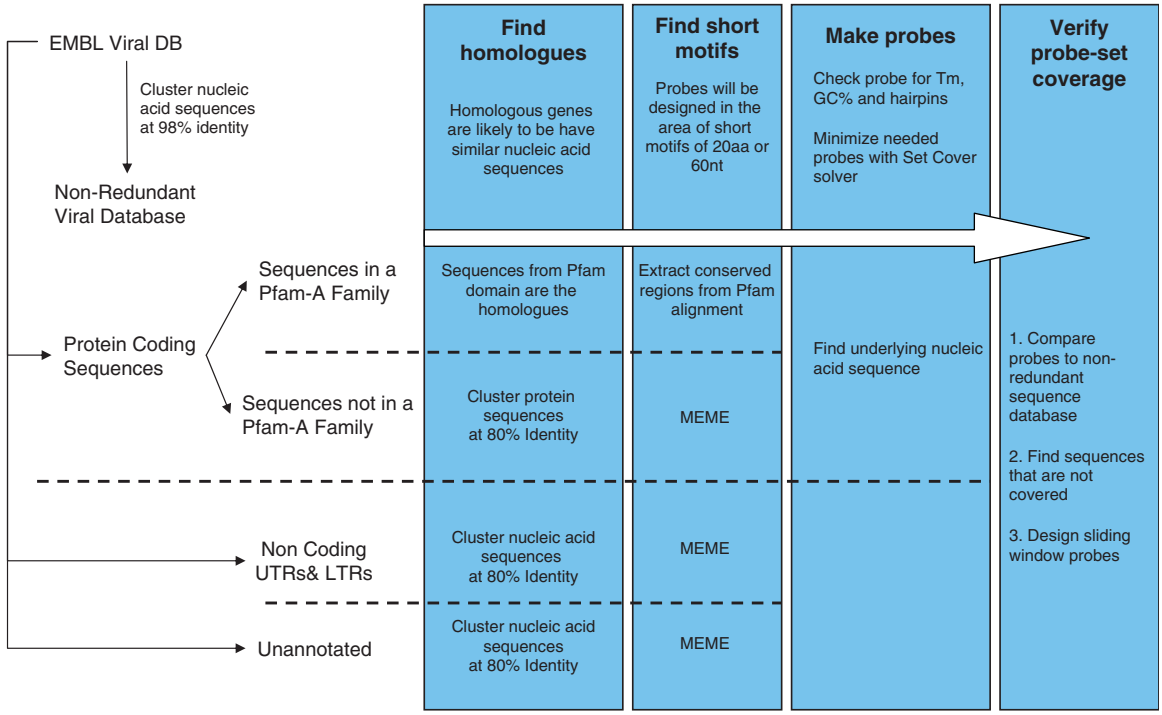

Figure 3. Comprehensive motif-based probe design. The EMBL viral database is clustered with a threshold of 98% nucleotide identity to create a non-redundant sequence database. Coding sequences are subjected to an amino acid motif search, and then probes are made from the underlying nucleic acid sequences. Similarly, nucleic acid motifs are found in non-coding sequences and used to make probes. Database coverage is checked; supplementary probes for highly divergent sequences are designed as necessary. Acronyms: Pfam—Protein Families database, MEME—Multiple Expectation maximization for Motif Elicitation, UTR—untranslated region, LTR—long terminal repeat.

We sought to validate the probe design method by generating a simple thermodynamic model to predict hybridization signal based on sequence composition. We computed the change in Gibbs free energy (ΔG) for all expected probe-viral nucleic acid pairs in the West Nile virus hybridization experiments described above. The calculation method employed finds the most thermodynamically stable structure (minimum free energy) (28) based on empirically established nearest neighbor energies (29). Strong signal was observed from probe-virus hybrids with ΔG of -32.5 kJ or less. Thus, this value was chosen as the threshold to classify a probe as likely to generate high signal when the cognate viral target is present (Figure 6).

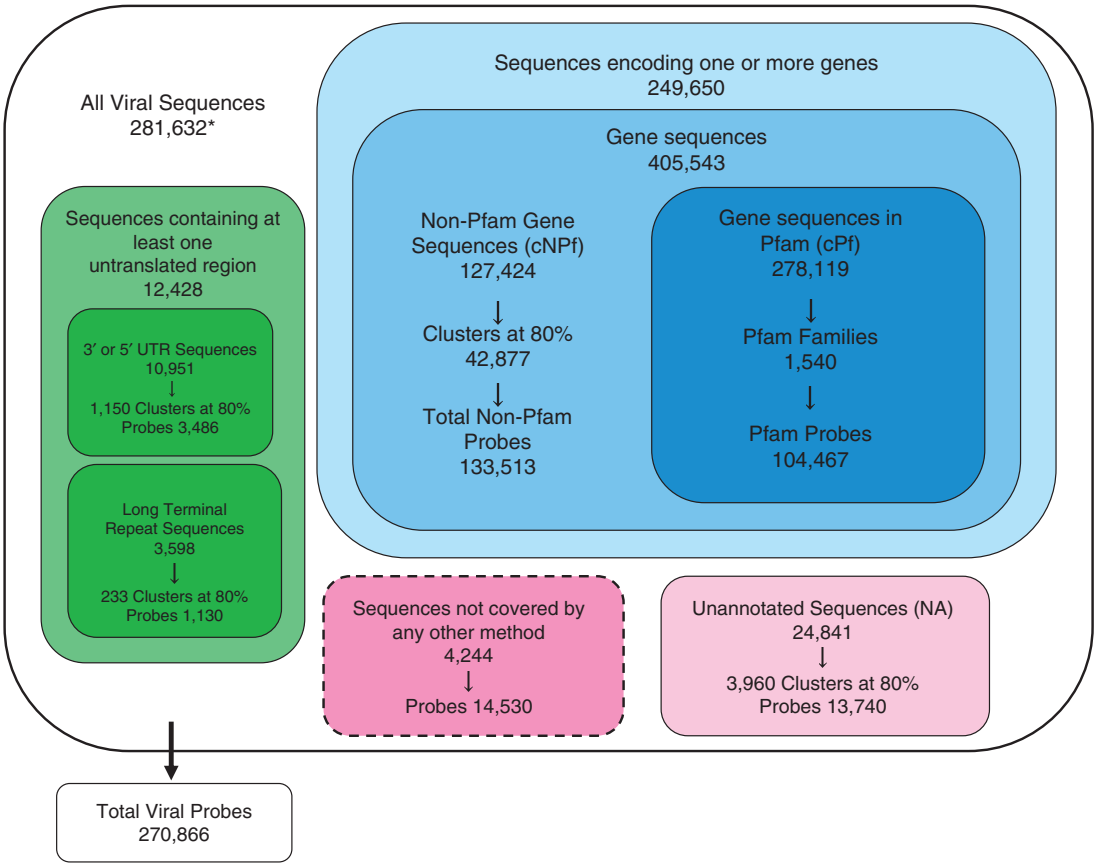

Figure 4. Sequence counts for the July 2007 EMBL release (Pfam 22.0). *Only complete genomes of HIV-1 were included in this database.

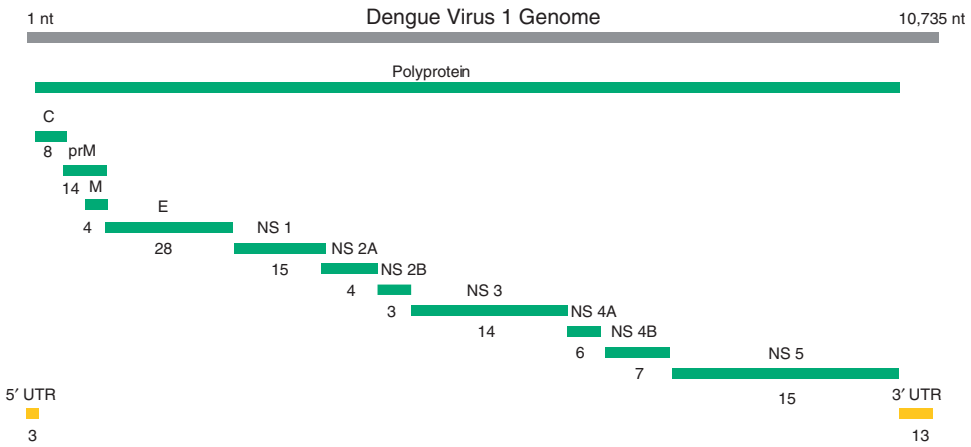

Figure 5. Probe distribution for Dengue virus 1. The number of probes targeting each region of the Dengue genome (NC_001477) appears below the protein segment.

Motif-based probe design provides higher coverage than virus genome tiling

Use of motif finding and set cover minimization markedly increases the computational resources needed to generate probe sets. To determine whether increased complexity results in a more comprehensive probe set, we compared our method to a genome tiling strategy. Probes of 60 nt were designed end-to-end along the entire genome for all 1701 Reference Sequence viral strains available as of May 2007. The tiling probe set served as a contrast to our design method since it was based on nucleic acid sequence, had more probes per gene, required less computation, and included viruses from all genera.

In comparison of the methods, the following rules were used to compute database coverage: sequences >400 nt in length were considered covered if six or more probes

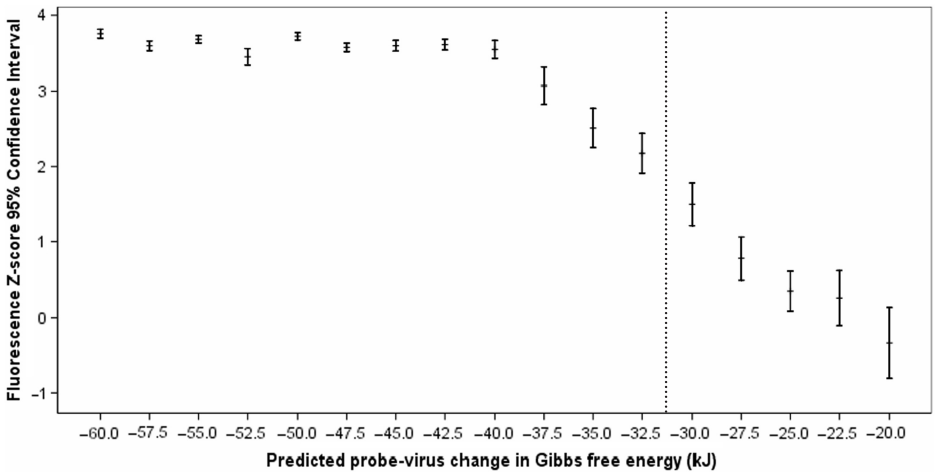

Figure 6. Gibbs free energy model of hybridization signal. The change in Gibbs free energy of probe-West Nile virus hybrids was computed. Aliquots of West Nile virus (New York 1999 strain RNA) at 10^6 copies were spiked into 200 ng of human lung (background) RNA. The fluorescent signal values of replicate arrays were $\log_2$ transformed, normalized, and converted to *Z*-scores. 95% confidence intervals of the mean for fluorescence versus Gibbs energy is plotted. Probe-virus hybrids with free energy $\leq$–32.5 kJ had high fluorescence; this value was chosen as the threshold for considering a probe likely to generate a strong signal when the target virus is present (dotted line).

met hybridization criteria; sequences <400 nt in length were considered covered if two probes met hybridization criteria; sequences <200 nt in length were considered covered with a single probe meeting hybridization criteria. Coverage of the entire database was gauged by computing probe-template ΔG for all 74 044 unique sequence representatives. Database coverage using the tiling method was 47.8% and required 850 136 probes; coverage using the motif-based method was 99.7% and required 270 866 probes (Table 1).

Whereas probe design in motif-based arrays can exploit partial genome sequences, probes in tiling arrays are based on full length genome sequences. Complete Reference Sequence genomes represent 1% of EMBL sequence entries. Although at least one full length genome sequence is described for all viral genera, only 49% (1701 of 3441) of viral species have a fully sequenced representative genome. The impact of differences in the motif and tiling-based strategies for probe design is reflected in differences in coverage. For the tiling-based probe-set, 40 of 44 families with <80% sequence coverage included species lacking representative genomes. Coverage with motif-based probe-sets for these same species was ≥93%.

DISCUSSION

Application of motif-based probe design to viral surveillance and genomics

There is an increasing appreciation for the power of microarray technology in clinical microbiology, public health and environmental surveillance. Viral microarray probe design poses unique challenges due to the rapid increase in sequence data and the high propensity for sequence divergence within viral taxa. To ensure coverage of the newest isolates it is essential to consider partial as well as complete genomic sequences in probe design. Probe design based on multiple alignments or pairwise comparisons of nucleic acids for all known sequences is computationally intensive and scales poorly with database size. Protein sequence comparisons are rapid and incorporate rich evolutionary models, but require a cumbersome mapping step to extract underlying nucleic acid sequence. We have described a method that capitalizes on the Pfam protein alignment database and a motif finding algorithm to automate the extraction of nucleic acid sequence for probes from conserved protein regions. The protein motif-centric method has several advantages: (i) the majority of viral nucleic acid sequences encode proteins; thus, using this information leverages knowledge about function; (ii) protein sequence comparison and the resulting probesets are independent of viral taxonomy; this may enable incorporation of misclassified sequences; (iii) the Pfam is a well established and highly annotated database that will provide a basis for future design efforts; and (iv) probes designed in conserved regions may be able to detect novel viruses.

Application to viral transcript profiling

A second application of this design method is viral expression profiling. Insights into the replication cycle, host evasion and virulence factors may be obtained by monitoring viral transcript levels during infection. To this end, arrays could be synthesized that combine probes for a single viral family and all host genes. Because the viral probe sets generated by our method account for known variants across all genes, a variety of strains could be profiled with a single array. This would provide a unique experimental platform for investigating virus biology, while minimizing fabrication cost and simplifying analysis.

Probe selection for other assay platforms

The thresholds used to design and validate probes were experimentally determined for the Agilent Technologies array platform and the types of clinical samples our

Table 1. Database sequence coverage of probe design methods

Virus family	Total sequences	Full genome sequences	Motif-based		Tiling-based	
			Covered sequences	Percentage	Covered sequences	Percentage
Adenoviridae	385	43	385	100.0	255	66.2
Arenaviridae	263	26	262	99.6	137	52.1
Arteriviridae	1891	4	1891	100.0	1328	70.2
Ascoviridae	38	3	38	100.0	26	68.4
Asfarviridae	92	1	86	93.5	64	69.6
Astroviridae	312	6	312	100.0	168	53.8
Avsunviroidae	124	4	124	100.0	122	98.4
Baculoviridae	414	40	408	98.6	249	60.1
Barnaviridae	1	1	1	100.0	1	100.0
Bicaudaviridae	2	2	2	100.0	2	100.0
Birnaviridae	284	12	284	100.0	255	89.8
Bornaviridae	22	1	22	100.0	22	100.0
Bromoviridae	475	72	472	99.4	409	86.1
Bunyaviridae	1204	66	1197	99.4	623	51.7
Caliciviridae	1644	16	1643	99.9	311	18.9
Caulimoviridae	227	27	224	98.7	95	41.9
Chrysoviridae	28	8	28	100.0	8	28.6
Circoviridae	174	16	174	100.0	170	97.7
Closteroviridae	386	27	386	100.0	240	62.2
Comoviridae	339	36	337	99.4	249	73.5
Coronaviridae	1040	18	1029	98.9	537	51.6
Corticoviridae	1	1	1	100.0	1	100.0
Cystoviridae	12	12	12	100.0	12	100.0
Dicistroviridae	67	14	67	100.0	54	80.6
Filoviridae	19	4	19	100.0	17	89.5
Flaviviridae	22 034	42	22 016	99.9	6105	27.7
Flexiviridae	916	61	910	99.3	464	50.7
Fuselloviridae	7	4	7	100.0	7	100.0
Geminiviridae	1406	236	1404	99.9	1353	96.2
Globuloviridae	3	2	3	100.0	3	100.0
Hepadnaviridae	2427	10	2427	100.0	2203	90.8
Hepeviridae	754	1	754	100.0	149	19.8
Herpesviridae	2164	47	2122	98.1	1147	53.0
Hypoviridae	34	4	34	100.0	21	61.8
Inoviridae	23	23	23	100.0	23	100.0
Iridoviridae	104	8	103	99.0	61	58.7
Leviviridae	9	9	9	100.0	9	100.0
Lipothrixviridae	5	2	5	100.0	2	40.0
Luteoviridae	172	18	172	100.0	154	89.5
Marnaviridae	1	1	1	100.0	1	100.0
Microviridae	42	42	42	100.0	42	100.0
Myoviridae	69	63	69	100.0	65	94.2
Nanoviridae	127	36	127	100.0	114	89.8
Narnaviridae	20	8	20	100.0	11	55.0
Nimaviridae	13	1	13	100.0	12	92.3
Nodaviridae	57	18	57	100.0	50	87.7
Orthomyxoviridae	4596	69	4589	99.8	3639	79.2
Papillomaviridae	803	60	792	98.6	248	30.9
Paramyxoviridae	1813	31	1807	99.7	1334	73.6
Partitiviridae	66	36	66	100.0	35	53.0
Parvoviridae	286	45	286	100.0	224	78.3
Phycodnaviridae	336	5	336	100.0	246	73.2
Picornaviridae	6672	35	6666	99.9	1792	26.9
Plasmaviridae	1	1	1	100.0	1	100.0
Podoviridae	70	52	70	100.0	53	75.7
Polydnaviridae	400	230	398	99.5	284	71.0
Polyomaviridae	353	14	353	100.0	187	53.0
Pospiviroidae	156	25	156	100.0	147	94.2
Potyviridae	1828	64	1827	99.9	1385	75.8
Poxviridae	364	24	342	94.0	242	66.5
Reoviridae	2717	363	2694	99.2	908	33.4
Retroviridae	7795	54	7789	99.9	4335	55.6
Rhabdoviridae	1334	19	1331	99.8	588	44.1
Roniviridae	7	0	7	100.0	0	0.0
Rudiviridae	13	3	12	92.3	6	46.2

(continued)

Table 1. Continued

Virus family	Total sequences	Full genome sequences	Motif-based		Tiling-based	
			Covered sequences	Percentage	Covered sequences	Percentage
Sequiviridae	59	3	59	100.0	52	88.1
Siphoviridae	149	149	149	100.0	148	99.3
Tectiviridae	6	6	6	100.0	6	100.0
Tetraviridae	10	6	10	100.0	7	70.0
Togaviridae	374	15	373	99.7	262	70.1
Tombusviridae	161	43	161	100.0	141	87.6
Totiviridae	95	27	95	100.0	68	71.6
Tymoviridae	78	13	78	100.0	40	51.3
No Family Designation	3671	302	3666	99.9	1647	44.9
Grand total	**74 044**	**2790**	**73 841**	**99.7**	**35 376**	**47.8**
			Total probes	270 866		850 136

laboratory encounters. Probe length can be selected to emphasize efficient coverage of higher order taxa or speciation. The goal of this project is to cover all known viral sequences and optimize potential for detecting related viral sequences. Thus, we designed 60 nt probes because they can better tolerate mismatched templates than 25 nt oligonucleotide probes (45). Using an empirical approach, appropriate thresholds can be determined for other array platforms, hybridization conditions, and probe lengths. The method of probe design and set-cover minimization is flexible and agnostic of platform; application to bead, solution, or surface-based hybridization technology should be straightforward.

Importance of continuous updates

Although the growth of the public sequence databases has been rapid, sequence diversity has not grown as quickly. If this trend continues, we anticipate that only incremental updates to a core set of probes will be needed to maintain array integrity. An update strategy would require periodic testing of probe sets against newly deposited sequences and fresh design only in the cases of high sequence divergence.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The work presented here was supported by National Institutes of Health awards (AI070411, Northeast Biodefense Center U54-AI057158-Lipkin, AI056118, HL083850 EY017404 and T32GM008224). We thank Carolyn Morrison for excellent technical assistance. Funding to pay the Open Access publication charges for this article was provided by NIH U54-AI057158-Lipkin.

Conflict of interest statement. None declared.

REFERENCES

1. An,H.J., Cho,N.H., Lee,S.Y., Kim,I.H., Lee,C., Kim,S.J., Mun,M.S., Kim,S.H. and Jeong,J.K. (2003) Correlation of cervical carcinoma and precancerous lesions with human papillomavirus (HPV) genotypes detected with the HPV DNA chip microarray method. *Cancer*, **97**, 1672–1680.
2. Wang,D., Urisman,A., Liu,Y.T., Springer,M., Ksiazek,T.G., Erdman,D.D., Mardis,E.R., Hickenbotham,M., Magrini,V. *et al.* (2003) Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol.*, **1**, E2.
3. Chiu,C.Y., Rouskin,S., Koshy,A., Urisman,A., Fischer,K., Yagi,S., Schnurr,D., Eckburg,P.B., Tompkins,L.S. *et al.* (2006) Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. *Clin. Infect. Dis.*, **43**, e71–76.
4. Lin,B., Wang,Z., Vora,G.J., Thornton,J.A., Schnur,J.M., Thach,D.C., Blaney,K.M., Ligler,A.G., Malanoski,A.P. *et al.* (2006) Broad-spectrum respiratory tract pathogen identification using resequencing DNA microarrays. *Genome Res.*, **16**, 527–535.
5. Townsend,M.B., Dawson,E.D., Mehlmann,M., Smagala,J.A., Dankbar,D.M., Moore,C.L., Smith,C.B., Cox,N.J., Kuchta,R.D. *et al.* (2006) Experimental evaluation of the FluChip diagnostic microarray for influenza virus surveillance. *J. Clin. Microbiol.*, **44**, 2863–2871.
6. Urisman,A., Molinaro,R.J., Fischer,N., Plummer,S.J., Casey,G., Klein,E.A., Malathi,K., Magi-Galluzzi,C., Tubbs,R.R. *et al.* (2006) Identification of a novel Gammaretrovirus in prostate tumors of patients homozygous for R462Q RNASEL variant. *PLoS Pathog.*, **2**, e25.
7. Kistler,A., Avila,P.C., Rouskin,S., Wang,D., Ward,T., Yagi,S., Schnurr,D., Ganem,D., DeRisi,J.L. *et al.* (2007) Pan-viral screening of respiratory tract infections in adults with and without asthma reveals unexpected human coronavirus and human rhinovirus diversity. *J. Infect. Dis.*, **196**, 817–825.
8. Chiu,C.Y., Alizadeh,A.A., Rouskin,S., Merker,J.D., Yeh,E., Yagi,S., Schnurr,D., Patterson,B.K., Ganem,D. *et al.* (2007) Diagnosis of a critical respiratory illness caused by human metapneumovirus by use of a pan-virus microarray. *J. Clin. Microbiol.*, **45**, 2340–2343.
9. Palacios,G., Quan,P.L., Jabado,O.J., Conlan,S., Hirschberg,D.L., Liu,Y., Zhai,J., Renwick,N., Hui,J. *et al.* (2007) Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg. Infect. Dis.*, **13**, 73–81.
10. Quan,P.L., Palacios,G., Jabado,O.J., Conlan,S., Hirschberg,D.L., Pozo,F., Jack,P.J., Cisterna,D., Renwick,N. *et al.* (2007) Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp Oligonucleotide Microarray. *J. Clin. Microbiol.*, **45**, 2359–2364.
11. Boriskin,Y.S., Rice,P.S., Stabler,R.A., Hinds,J., Al-Ghusein,H., Vass,K. and Butcher,P.D. (2004) DNA microarrays for virus

detection in cases of central nervous system infection. *J. Clin. Microbiol.*, **42**, 5811–5818.
12. Boonham,N., Walsh,K., Smith,P., Madagan,K., Graham,I. and Barker,I. (2003) Detection of potato viruses using microarray technology: towards a generic method for plant viral disease diagnosis. *J. Virol. Methods*, **108**, 181–187.
13. Wang,D., Coscoy,L., Zylberberg,M., Avila,P.C., Boushey,H.A., Ganem,D. and DeRisi,J.L. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc. Natl Acad. Sci. USA*, **99**, 15687–15692.
14. Lin,F.M., Huang,H.D., Chang,Y.C., Tsou,A.P., Chan,P.L., Wu,L.C., Tsai,M.F. and Horng,J.T. (2006) Database to dynamically aid probe design for virus identification. *IEEE Trans. Inf. Technol. Biomed.*, **10**, 705–713.
15. Chou,C.C., Lee,T.T., Chen,C.H., Hsiao,H.Y., Lin,Y.L., Ho,M.S., Yang,P.C. and Peck,K. (2006) Design of microarray probes for virus identification and detection of emerging viruses at the genus level. *BMC Bioinform.*, **7**, 232.
16. Chizhikov,V., Wagner,M., Ivshina,A., Hoshino,Y., Kapikian,A.Z. and Chumakov,K. (2002) Detection and genotyping of human group A rotaviruses by oligonucleotide microarray hybridization. *J. Clin. Microbiol.*, **40**, 2398–2407.
17. Laassri,M., Chizhikov,V., Mikheev,M., Shchelkunov,S. and Chumakov,K. (2003) Detection and discrimination of orthopoxviruses using microarrays of immobilized oligonucleotides. *J. Virol. Methods*, **112**, 67–78.
18. Mehlmann,M., Dawson,E.D., Townsend,M.B., Smagala,J.A., Moore,C.L., Smith,C.B., Cox,N.J., Kuchta,R.D. and Rowlen,K.L. (2006) Robust sequence selection method used to develop the FluChip diagnostic microarray for influenza virus. *J. Clin. Microbiol.*, **44**, 2857–2862.
19. Wilson,W.J., Strout,C.L., DeSantis,T.Z., Stilwell,J.L., Carrano,A.V. and Andersen,G.L. (2002) Sequence-specific identification of 18 pathogenic microorganisms using microarray technology. *Mol. Cell. Probes*, **16**, 119–127.
20. Wong,C.W., Albert,T.J., Vega,V.B., Norton,J.E., Cutler,D.J., Richmond,T.A., Stanton,L.W., Liu,E.T. and Miller,L.D. (2004) Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res.*, **14**, 398–405.
21. Sulaiman,I.M., Tang,K., Osborne,J., Sammons,S. and Wohlhueter,R.M. (2007) GeneChip resequencing of the smallpox virus genome can identify novel strains: a biodefense application. *J. Clin. Microbiol.*, **45**, 358–363.
22. Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.
23. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
24. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
25. Dunbrack,R.L., Jr. (2006) Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 374–384.
26. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
27. Andronescu,M., Aguirre-Hernandez,R., Condon,A. and Hoos,H.H. (2003) RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, **31**, 3416–3422.
28. Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.
29. SantaLucia,J., Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.
30. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
31. Cochrane,G., Aldebert,P., Althorpe,N., Andersson,M., Baker,W., Baldwin,A., Bates,K., Bhattacharyya,S., Browne,P. *et al.* (2006) EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res.*, **34**, D10–D15.
32. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
33. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
34. Bailey,T.L., Baker,M.E. and Elkan,C.P. (1997) An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. *J. Steroid Biochem. Mol. Biol.*, **62**, 29–44.
35. Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) In Dayhoff,M. O. (ed.), *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation., Washington, DC, Vol. 5, pp. 345–352.
36. Jabado,O.J., Palacios,G., Kapoor,V., Hui,J., Renwick,N., Zhai,J., Briese,T. and Lipkin,W.I. (2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Res.*, **34**, 6605–6611.
37. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
38. Bao,Y., Federhen,S., Leipe,D., Pham,V., Resenchuk,S., Rozanov,M., Tatusov,R. and Tatusova,T. (2004) National center for biotechnology information viral genomes project. *J. Virol.*, **78**, 7291–7298.
39. Shafer,R.W. (2006) Rationale and uses of a public HIV drug-resistance database. *J Infect Dis*, **194**(Suppl. 1), S51–S58.
40. McCutchan,F.E. (2006) Global epidemiology of HIV. *J. Med. Virol.*, **78**(Suppl. 1), S7–S12.
41. Held,G.A., Grinstein,G. and Tu,Y. (2003) Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl Acad. Sci. USA*, **100**, 7575–7580.
42. Matveeva,O.V., Shabalina,S.A., Nemtsov,V.A., Tsodikov,A.D., Gesteland,R.F. and Atkins,J.F. (2003) Thermodynamic calculations and statistical correlations for oligo-probes design. *Nucleic Acids Res.*, **31**, 4211–4217.
43. Bruun,G.M., Wernersson,R., Juncker,A.S., Willenbrock,H. and Nielsen,H.B. (2007) Improving comparability between microarray probe signals by thermodynamic intensity correction. *Nucleic Acids Res.*, **35**, e48.
44. Hekstra,D., Taussig,A.R., Magnasco,M. and Naef,F. (2003) Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.*, **31**, 1962–1968.
45. Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.