# Selecting Genotyping Oligo Probes Via Logical Analysis of Data*,**

Kwangsoo Kim and Hong Seo Ryoo***

Division of Information Management Engineering, Korea University
1, 5-Ka, Anam-Dong, Seongbuk-Ku, Seoul, 136-713, Korea
Phone: +82-2-3290-3394, Fax: +82-2-929-5888
{kksoo,hsryoo}@korea.ac.kr

**Abstract.** Based on the general framework of logical analysis of data, we develop a probe design method for selecting short oligo probes for genotyping applications in this paper. When extensively tested on genomic sequences downloaded from the Lost Alamos National Laboratory and the National Center of Biotechnology Information websites in various monospecific and polyspecific *in silico* experimental settings, the proposed probe design method selected a small number of oligo probes of length 7 or 8 nucleotides that perfectly classified all unseen testing sequences. These results well illustrate the utility of the proposed method in genotyping applications.

**Keywords:** oligo probes, microarrays, LAD, set covering, learning theory, optimization, viral pathogens.

## 1 Introduction

A microarray or a DNA chip is a small glass or silica surface bearing DNA probes. Probes are single stranded reverse transcribed mRNAs, each located at a specific spot of the chip for hybridization with its Watson-Crick complementary sequence in a target to form the double helix [1]. Microarrays currently use two forms of probes, namely, oligonucleotide (shortly, oligo) and cDNA, and have prevalently been used in the analysis of gene expression levels, which measures the amount of gene expression in a cell by observing the hybridization of mRNA to different probes, each targeting a specific gene. With the ability to identify a specific target in a biological sample, microarrays are also well-suited for detecting biological agents for genetic and chronic disease [2,3,4,5]. Furthermore, as viral pathogens can be detected at the molecular and genomic level much before the onset of physical symptoms in a patient, the microarray technology can be used for an early detection of patients infected with viral pathogens [6,7,8].

* This paper is dedicated to the life and memory of Dr. Peter L. Hammer (December 23, 1936 - December 27, 2006), the inventor of LAD and an OR giant.

** This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2005-003-D00445).

*** Corresponding author.

Z. Kobti and D. Wu (Eds.): Canadian AI 2007, LNAI 4509, pp. 86–97, 2007.

The success of microarrays depends on the quality of probes that are tethered on the chip. Having an optimized set of probes is beneficial for two reasons. One, the background hybridization is minimized, hence true gene expression levels can be more accurately determined [9]. The other, as the number of oligos needed per gene is minimized, the cost of each microarray is minimized or the number of genes on each chip is increased, yielding oligo fingerprinting a much faster and more cost-efficient technique [9,10]. Short probes consisting of 15-25 nucleotides (nt) are used in genotyping applications [1]. Having short optimal probes means a high genotyping accuracy in terms of both sensitivity and specificity [6,9] and can play a key role in genotyping applications.

From the perspective of numerical optimization, genomic data present an unprecedented challenge for supervised learning approaches for a number of reasons. First, genomic data are long sequences over the nucleic acid alphabet $\Sigma = \{\text{A,C,G,T}\}$. Second, for example, the complexity of viral flora, owing to constantly evolving viral serotypes, requires a supervised learning theory to be trained on a large collection of target and non-target samples. That is, a typical training set contains a large number of large-scale samples. Third, a supervised learning framework usually requires a systematic pairing or differencing between each target and non-target samples during the course of training a decision rule [10,11,12,13]. Adding to these, the nature of data classification is difficult [14].

Based on the general framework of logical analysis of data (LAD) from [15], we develop in this paper a probe design method for selecting short oligo probes of length $l$ nt, where $l \in [6, 10]$. To list some advantages of selecting oligo probes by the proposed method, first, the method selects probes via sequential solution of a small number of compact set covering (SC) instances, which offers a great advantage from computational point of view. To be more specific, consider classification of two types of data and suppose that a training set is comprised of $m^+$ target and $m^-$ non-target sequences. The size of the SC training instances solved by the proposed method is minimum of $m^+$ and $m^-$ orders of magnitude smaller than optimization learning models used in [10,11,12]. Second, the method uses the sequence information only and selects probes via optimization based on principles of probability and statistics. That is, the probability of an $l-$mer (oligo of length $l$) appearing in a single sequence by chance is $(0.25)^l$, hence the probability of an $l-$mer appearing in multiple samples of one type but in none or only a few of the sequences of the other type by chance alone is extremely small. Third, the proposed method does not rely on any extra tool, such as BLASTn [16], a local sequence alignment search tool that is commonly used for probe selection [6,8,17], or the existence of pre-selected representative probes [6]. This makes the method truly stand-alone and free of problems that may possibly be caused by limitations associated with external factors. Last, with an array of efficient (meta-)heuristic solution procedures for SC, the proposed method is readily implementable for an efficient selection of oligo probes.

As for the organization of this paper, we develop an effective method for selecting short oligo probes in Section 2 (for reasons of space, we omit proofs for the mathematical results in this section) and extensively test the proposed

probe design method in various *in silico* genotyping experiments in Section 3 with using viral genomic sequences from the Los Alamos National Laboratory and the National Center of Biotechnology Information websites.

## 2 Proposed Probe Selection Method

The task of classifying more than two types of data can be accomplished by sequential classifications of two types of + and − data (see [18,19,20] and Section 3 below). Without loss of generality, therefore, we present the material below in the context of binary classification.

The backbone of the proposed procedure is LAD. A typical implementation of LAD analyzes data on hand via four sequential stages of data binarization, support feature selection, pattern generation and classification rule formation. As a Boolean logic-based, LAD first converts all non-binary data into equivalent binary observations. A + (−) 'pattern' in LAD is defined as a conjunction of one or more binary attributes or their negations that distinguishes one or more + (−) type observations from all − (+) observations. The number of attributes used in a pattern is called the 'degree' of the pattern. As seen from the definition, patterns hold the structural information hidden in data. After patterns are generated, they are aggregated into a partially-defined Boolean discriminant function/rule to generalize the discovered knowledge to classify new observations.

Referring readers to [13,15,21] for more background in LAD, we design a LAD-based method below for efficiently analyzing large-scale genomic data.

### 2.1 Data Binarization

Let there be $m^+$ and $m^-$ sample observations of type + (target) and − (non-target), respectively. For $\bullet \in \{+,-\}$, let us use $\bar{\bullet}$ to denote the complementary element of $\bullet$ with respect to the set $\{+,-\}$. Let $S^\bullet$ denote the index set of $m^\bullet$ sample sequences for $\bullet \in \{+,-\}$.

A DNA sequence is a sequence of nucleic acids A, C, G and T, and the training sequences need to be converted into Boolean sequences of 0 and 1 before LAD can be applied. Toward this end, we first choose an integer value for $l$, usually $l \in [6,10]$ (see Section 3), generate all $4^l$ possible $l-$mers over the four nucleic acid letters and then number them consecutively from 1 to $4^l$ by a mapping scheme. Next, each $l-$mer is selected in turn and every training sample is fingerprinted with the oligo for its presence or absence. That is, with oligo $j$, we scan each sequence $p_i$, $i \in S^+ \cup S^-$, from the beginning of the sequence and shifting to the right by a base and stamp

$$p_{ij} = \begin{cases} 1, & \text{if oligo } j \text{ is present in sequence } i; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

After this, the oligos that appear in all or none of the training sequences can be deleted from further consideration. We re-number the surviving $l-$mers consecutively from 1 to $n$ and replace the original training sequences described in the nucleic acid alphabets by their Boolean representations. Let $N = \{1,\ldots,n\}$.

### 2.2 Pattern Generation

The data are now described by $n$ attributes $a_j \in \{0,1\}$, $j \in N$. For observation $p_i$, $i \in S^\bullet$, $\bullet \in \{+,-\}$, let $p_{ij}$ denote the binary value the $j-$th attribute takes in this observation. Denote by $l_j$ the literal of binary attribute $a_j$. Then, $l_j = a_j$ ($l_j = \overline{a}_j$) instructs to take (negate) the value of $a_j$ in all sequences. A term $t$ is a conjunction of literals. Given a term $t$, let $N_t \subseteq N$ denote the index of literals included in the term. Then, we have $t = \bigwedge_{j \in N_t} l_j$. A $\bullet$ pattern is a term that satisfies $t(p_i) := \prod_{\substack{l_j = a_j, \\ j \in N_t}} p_{ij} \prod_{\substack{l_j = \bar{a}_j, \\ j \in N_t}} \bar{p}_{ij} = 1$ for at least one $p_i$, $i \in S^\bullet$, and $t(p_k) = 0$ for all $p_k$, $k \in S^{\bar{\bullet}}$. Note here that $N_t$ of a $\bullet$ pattern identifies probes that collectively distinguish one or more $\bullet$ sequences from the sequences of the other type.

Let us introduce $n$ additional features $a_{n+j}$, $j \in N$, and use $a_{n+j}$ to negate $a_j$. Let $N' = \{1, \ldots, 2n\}$ and let us introduce a binary decision variable $x_j$ for $a_j$, $j \in N'$, to determine whether to include $l_j$ in a pattern. [15] formulated a compact mixed integer and linear programming (MILP) model below with respect to a reference sample $p_i$, $i \in S^\bullet$, $\bullet \in \{+,-\}$:

$$\text{(MILP-2.i}^\bullet\text{)} \quad \left| \begin{array}{lll} z_{2.i} = \min\limits_{\mathbf{x},\mathbf{y},d} \sum\limits_{l \in S^\bullet \setminus \{i\}} y_l & & \\ \quad \text{s. t.} & \sum\limits_{j \in J_i} x_j = d & \\ & \sum\limits_{j \in J_i} p_{lj} x_j + y_l \geq d, & l \in S^\bullet \setminus \{i\} \\ & \sum\limits_{j \in J_i} p_{lj} x_j \leq d - 1, & l \in S^{\overline{\bullet}} \\ & 1 \leq d \leq n & \\ & \mathbf{x} \in \{0,1\}^n & \\ & \mathbf{0} \leq \mathbf{y} \leq \mathbf{n}, & \end{array} \right.$$

where $J_i := \{j \in N' : p_{ij} = 1\}$ for $p_i$, $i \in S^\bullet$. Consider the following.

**Lemma 1.** *Let* $(\mathbf{x}, \mathbf{y}, d)$ *denote a feasible solution of (MILP-2.i*$^\bullet$*). Let* $N_t = \{j \in J_i : x_j = 1\}$. *Then,* $\mathcal{P} := \bigwedge_{j \in J_i, x_j = 1} a_j$ *forms a* $\bullet$ *pattern.*

We note here that genomic data are large-scale in nature. Furthermore, owing to constantly evolving viral serotypes, the complexity of viral flora is high, and this requires large numbers of target and non-target viral samples to be used for selecting optimal genotyping probes. Adding to these the difficulties associated with numerical solution of MILP, we see that (MILP-2.i$^\bullet$) above presents no practical way of selecting genotyping probes.

With the need to develop a more efficient pattern generation scheme, we select a reference sequence $p_i$, $i \in S^\bullet$, $\bullet \in \{+,-\}$, and set

$$a_j^{(i,k)} = \begin{cases} 1, & \text{if } p_{ij} \neq p_{kj}; \text{ and} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

for $k \in S^{\bar{\bullet}}$ and $j \in N$. Next, we set

$$a_j^{(i,l)} = \begin{cases} 1, & \text{if } p_{ij} = p_{lj}; \text{ and} \\ 0, & \text{otherwise,} \end{cases}$$

for $l \in S^\bullet$ and $j \in N$. Now, consider the set covering model

$$(\mathrm{SC}_i^\bullet) \quad \left| \begin{array}{lll} \min\limits_{\mathbf{x},\mathbf{y}} & \sum\limits_{j\in N} c_j x_j + \sum\limits_{l\in S^\bullet\setminus\{i\}} y_l & \\ \text{s.t.} & \sum\limits_{j\in N} a_j^{(i,l)} x_j + y_l \geq 1, & l \in S^\bullet \setminus \{i\} \\ & \sum\limits_{j\in N} a_j^{(i,k)} x_j \geq 1, & k \in S^{\bar{\bullet}} \\ & x_j \in \{0,1\}, & j \in N \\ & y_l \in \{0,1\}, & l \in S^\bullet \setminus \{i\}, \end{array} \right.$$

where $c_j$ $(j \in N)$ are positive real numbers.

**Theorem 1.** *Let* $(\mathbf{x}, \mathbf{y})$ *denote a feasible solution of* $(SC_i^\bullet)$*. Then,*

$$\mathcal{P} := \bigwedge_{\substack{x_l=1,\\ p_{il}^\bullet=1}} a_l \bigwedge_{\substack{x_l=1,\\ p_{il}^\bullet=0}} \bar{a}_l \tag{2}$$

*forms a* $\bullet$ *LAD pattern.*

**Lemma 2.** *With a feasible solution* $(\mathbf{x}, \mathbf{y})$ *of* $(SC_i^\bullet)$*, let* $N_t = \{j \in N : x_j = 1\}$*. Then,* $y_l = 0$ *for* $l \in S^\bullet \setminus \{i\}$ *if and only if* $p_{lk} = p_{ik}$ *for all* $k \in N_t$*.*

Although smaller than the MILP counterpart by only one constraint and one integer variable, $(\mathrm{SC}_i^\bullet)$ has a much simpler structure and is defined only in terms of 0-1 variables. In addition, it can exploit any of SC heuristic procedures developed so far (see, for example, [22] and references therein) for its efficient solution, hence is much preferred.

Note that $(\mathrm{SC}_i^\bullet)$ is defined by $m^+ + m^- - 1$ cover inequalities and $n + m^\bullet - 1$ binary variables. Also, recall that $n$ is large for genomic sequences and the analysis of viral sequences requires large numbers of target and non-target sequences, that is, $m^+$ and $m^-$ are also large numbers. To develop a more compact SC-based probe selection model, we select a reference sequence $p_i$, $i \in S^\bullet$, $\bullet \in \{+,-\}$, and

set the values of $a_j^{(i,k)}$ for $k \in S^{\bar{\bullet}}$ and $j \in N$ via (1). Consider the following SC model:

$$(\text{SC-pg}_i^{\bullet}) \quad \left| \begin{array}{ll} \min\limits_{\mathbf{x}} \sum\limits_{j \in N} c_j x_j & \\ \text{s.t.} \sum\limits_{j \in N} a_j^{(i,k)} x_j \geq 1, & k = 1, \ldots, m^{\bar{\bullet}} \\ x_j \in \{0,1\}, & j \in N, \end{array} \right.$$

where $c_j$'s are positive reals.

**Theorem 2.** *Let* $\mathbf{x}$ *denote a feasible solution of (SC-pg*$_i^{\bullet}$*). Then,* $\mathcal{P}$ *generated on* $\mathbf{x}$ *via* (2) *forms a* $\bullet$ *LAD pattern.*

**Lemma 3.** *With a feasible solution* $\mathbf{x}$ *of (SC*$_i^{\bullet}$*), generate a* $\bullet$ *pattern* $\mathcal{P}$ *via* (2)*. Then,* $\mathcal{P}$ *distinguishes every* $\bullet$ *sequence* $p_l$*,* $l \in S^{\bullet}$*, with* $p_{lk} = p_{ik}$ *for all* $k \in N_t$ *from the* $\bar{\bullet}$ *observations, where* $N_t = \{j \in N : x_j = 1\}$*.*

Below, we use (SC-pg$_i^{\bullet}$) to design one simple oligo probe selection procedure. Let $P^{\bullet}$ denote the set of $\bullet$ patterns generated so far.

**procedure** SC-pg
**begin**
  **for** $\bullet \in \{+, -\}$ **do**
    set $P^{\bullet} = \emptyset$ and $S \leftarrow S^{\bullet}$.
    **while** $S \neq \emptyset$ **do**
      - randomly choose $p_i$, $i \in S$, and formulate (SC-pg$_i^{\bullet}$).
      - solve (SC-pg$_i^{\bullet}$).
      - generate a $\bullet$ pattern $\mathcal{P}$ via (2).
      - set $P^{\bullet} = P^{\bullet} \cup \{\mathcal{P}\}$ and $S = S \setminus \{i\} \setminus \{j \in S, j \neq i : p_{jk} = p_{ik}, \forall k \in N_t\}$.
    **end while**
  **end for**
**end**

**Theorem 3.** ***procedure*** *SC-pg terminates finitely.*

## 3 Experiments and Discussions

In this section, we extensively test the proposed probe design for the classification of viral disease-agents in *in silico* setting with using genomic sequences obtained from the Los Alamos National Laboratory (LANL) and the National Center for Biotechnology Information (NCBI). Table 1 summarizes the number and the length (the minimum, average$\pm$1 standard deviation and maximum lengths) of each type of the genomic data that were used in our experiments.

In analyzing data in an experiment, we first decided on a length of oligos to use by calculating the smallest integer value $l$ such that $4^l$ became larger than or equal to the average of the lengths of target and non-target sequences of the experiment. Then, $4^l$ candidate oligos were generated to fingerprint and binarize

the data. Note here that if a constraint in (SC-pg$_i^\bullet$) has all zero coefficients, then the SC instance has no feasible solution, and this case arises when the reference sequence $p_i$, $i \in S^\bullet$, and the sequence $p_j$, $j \in S^{\bar\bullet}$ have identical 0-1 fingerprints, which is a contradiction. Supervised learning methodologies, including LAD, presume for the existence of a classification function that each unique sequence in the training set belongs to exactly one of the two classes. When data under analysis are indeed contradiction-free, then contradiction-free 0-1 clones of the data can always be obtained by using oligos of longer length for data fingerprinting and binarization. Therefore, when we generated the identical fingerprint for data of different types, we incremented the value of $l$ by 1 and repeated the data binarization stage until the binary representations of the data became contradiction free. Next, **procedure** *SC-pg* was applied to generate patterns, hence probes. In applying **procedure** *SC-pg* in these *in silico* experiments, we selected a minimal set of oligo probes by setting $c_j = 1$ for all $j \in N$. For solving the unicost (SC-pg$_i^\bullet$)'s generated, we used the textbook greedy heuristic [23] for ease of implementation.

Denote by $P_1^+, \ldots, P_{n_+}^+$ and $P_1^-, \ldots, P_{n_-}^-$ the positive and negative patterns, respectively, generated via **procedure** *SC-pg*. In classifying unseen + (target) and − (non-target) sequences, we use three decision rules. Specifically, for the polyspecific genotyping experiments (in Section 3.1 and Experiments 2 and 3 in Section 3.2), we form the standard LAD classification rule [13]

$$\Delta := \sum_{i=1}^{n_+} \frac{\omega_i^+}{|S^+|} P_i^+ - \sum_{i=1}^{n_-} \frac{\omega_i^-}{|S^-|} P_i^-, \tag{3}$$

where $\omega_i^\bullet$ denotes the number of $\bullet$ training sequences covered by $P_i^\bullet$. We assign class + (−) to new sequence $p$ if $\Delta(p) > 0$ ($\Delta(p) < 0$). We fail to classify sequence $p$ if $\Delta(p) = 0$.

For monospecific genotyping in Experiment 1 in Section 3.2, we form a decision rule by

$$\Delta^+ := \sum_{i=1}^{n_+} P_i^+ \text{ and } \Delta^- := \sum_{i=1}^{n_-} P_i^- \tag{4}$$

and assign $p$ to class $\bullet$ if $\Delta^\bullet(p) > 0$ while $\Delta^{\bar\bullet}(p) = 0$. When $\Delta^\bullet(p) > 0$ and $\Delta^{\bar\bullet}(p) > 0$ or when $\Delta^\bullet(p) = 0$ and $\Delta^{\bar\bullet}(p) = 0$, we fail in classifying the sequence.

For monospecific classification of more than two viral (sub-)types $k = 1, \ldots, m$ in Experiment 4 in Section 3.2, we use the decision rule

$$\Delta^k := \sum_{i=1}^{n_k} P_i^k, \tag{5}$$

where $P_1^k, \ldots, P_{n_k}^k$ are the probe(s) selected to for virus (sub-)type $k$, and assign $p$ to class $k$ if $\Delta^k(p) > 0$ while $\Delta^i(p) = 0$ for all $i = 1, \ldots, m, i \neq k$. When $\Delta(p) > 0$ for more than two virus types or $\Delta^k(p) = 0$ for all $k$, then we fail to assign a class to sequence $p$.

In each of the experiments in this section, we tested the proposed oligo probe selection method in 20 independent hold-out experiments, each with randomly selected 90% of the target and of the non-target data forming a training set of sequences and the remaining 10 % of the target and of the non-target sequences forming the testing data. More specifically, after a training set of data was formed, we binarized the training data and selected optimal oligo probes on them via **procedure** *SC-pg*. Next, a classification rule was formed by one of (3), (4) and (5) above and then used for classifying the corresponding testing sequences. These steps were repeated 20 times to obtain the average testing performance and other relevant information of the experiment.

The computational platform used for these experiments was an Intel 2.66GHz Pentium Linux PC with 512Mb of memory.

**Table 1.** Viral sequences used in experiments

| viral sequence | number | length | | |
|---|---|---|---|---|
| | | min. | avg.±1 std. dev. | max. |
| human papillomavirus (HPV): | | | | |
| - high risk HPV | 18 | 449 | 7365±1730 | 7989 |
| - low risk HPV | 54 | 455 | 7198±1683 | 8027 |
| SARS coronavirus | 105 | 29350 | 29692±91 | 29765 |
| coronavirus | 39 | 9203 | 29013±3569 | 31526 |
| other virus: | | | | |
| - human respiratory syncytial virus | 10 | 13933 | 15091±386 | 15226 |
| - human adenovirus | 32 | 34125 | 35215±618 | 36015 |
| - human parainfluenza virus | 4 | 15646 | 15652±3 | 15654 |
| - human rhinovirus (A, B) | 8 | 7102 | 7157±36 | 7212 |
| - influenza virus (A, B, C) | 53 | 838 | 1701±527 | 2368 |
| influenza virus hemagglutinin (H) subtype: | | | | |
| - H1 | 137 | 1698 | 1749±24 | 1778 |
| - H3 | 660 | 1695 | 1735±21 | 1768 |
| - H5 | 148 | 1677 | 1721±25 | 1779 |
| - H7 | 77 | 1659 | 1690±27 | 1792 |
| - H9 | 93 | 1683 | 1704±26 | 1742 |
| - H else (2, 4, 6, 8, 11, 12, 13, 16) | 65 | 1689 | 1742±29 | 1773 |
| influenza virus neuraminidase (N) subtype: | | | | |
| - N1 | 218 | 1344 | 1410±39 | 1463 |
| - N2 | 1050 | 1341 | 1434±28 | 1467 |
| - N3 | 44 | 1326 | 1411±29 | 1460 |
| - N else (4, 5, 6, 7, 8, 9) | 64 | 1341 | 1434±25 | 1467 |

### 3.1 Classification of High and Low Risk HPV: A Comparative Experiment

The infection with HPV is the main cause of cervical cancer, the second most common cancer in women worldwide [24,25]. There are more than 80 identified

types of HPV and the genital HPV types are subdivided into high and low risk types: low risk HPV types are responsible for most common sexually transmitted viral infections while high risk HPV types are a crucial etiological factor for the development of cervical cancer [26].

We applied the proposed probed design method on the 72 HPV sequences downloaded from LANL with their classification found in Table 3 of [27]. The selected probes were used to form a decision rule by (3) and tested for their classification capability.

Results from this polyspecific probe selection experiment are provided in Table 2. In this table and also in the table found in the following subsection, the target (+) and the non-target (−) virus types of the experiments are first specified. Then, the tables provide two bits of information on the candidate oligos, namely, the length $l$ and the average and the standard deviation of the number of features generated and used in the 20 runs of each experiment for data binarization and for pattern generation. Provided next in the tables is the information on the number of probes selected in the format 'the average ± 1 standard deviation' and information on the LAD patterns generated. Finally, the testing performance of the probes selected is provided in the last column of the tables, summarized in format 'the average ± 1 standard deviation' of the percentage of the correct classifications of the unseen sequences.

**Table 2.** Polyspecific classification of high and low risk HPV by the proposed method

| **Experiment** | $l$−mers used | | probes selected | | testing |
|---|---|---|---|---|---|
| | $l$ | number* | number* | patterns | accuracy*† |
| high risk HPV (+) vs. HPV low (−) | 8 | 58359.9±130.4 | 18.7±1.7<br>22.8±1.6 | degree 1 & 2 patterns<br>degree 1 & 2 patterns | 90.6±9.8 |

*: in format average ± standard deviation

†: percentage of correct classifications of testing/unseen data

Briefly summarizing, the proposed probe design method selected probes on the HPV data in a few CPU seconds that tested 90.6% accurate in classifying the unseen HPV samples. For comparison, the same HPV dataset was used in [2] and [27] for the classification of HPV by high and low risk types. In brief, the probe design methods of [2] and [27] required several CPU hours of computation and selected probes that obtained 85.6% and 81.1% correct classification rates, respectively.

Before moving on, we note that the sequences belonging to the target and the non-target groups in this experiment all have different HPV subtypes (see Table 3 in [27]). The combination of all target and non-target sequences being different from one another and the presence of noise in the data (the classification errors) gave rise to selecting a relatively large number of polyspecific probes in this experiment.

### 3.2 Genotyping Experiments with Viral Pathogens

The proposed probe design method was tested on genomic viral sequences from NCBI for selecting monospecific and polyspecific probes for screening for SARS and AI in a number of different binary and multicategory experimental setting and performed superbly on all counts. We describe individual experiments below and summarize results from these experiments in Table 3.

**Table 3.** Genotyping viral pathogens by the proposed method

| Experiment | viruses | $l-$mers used | | probes selected | | testing |
|---|---|---|---|---|---|---|
| | distinguished | $l$ | number* | number* | patterns generated | accuracy*† |
| **1** | SARS virus (+) | 8 | 57745.3±306.1 | 1±0 | degree 1 | 100±0 |
| | coronavirus (−) | | | 1±0 | degree 1 | |
| **2** | SARS virus (+) | 8 | 64141.5±36.5 | 1±0 | degree 1 | 100±0 |
| | influenza virus (−) | | | 10.1±0.8 | degree 1 only | |
| **3** | H5 & H9 (+) | 8 | 39056±398.3 | 6.7±0.5 | degree 1 only | 100±0 |
| | other H strains (−) | | | 21.6±1.3 | degree 1 only | |
| **4** | N1 | 7 | 13151±39.3 | 3±0 | degree 1 | 100±0 |
| | N2 | | | 3.7±0.5 | degree 1 only | |
| | N3 | | | 1±0 | degree 1 | |

*: in format average ± standard deviation

†: percentage of correct classifications of testing/unseen data

**Experiment 1.** SARS virus vs. coronavirus
SARS virus is phylogenetically most closely related to group 2 coronavirus [28]. 105 SARS sequences and 39 coronavirus samples were used to select 1 monospecific probe for screening for SARS. Used in a classification rule (4), the SARS probe and one probe selected for coronavirus together perfectly classified all testing sequences.

**Experiment 2.** SARS virus vs. influenza virus
This experiment simulates a SARS pandemic where suspected patients with SARS-like symptoms are screened for the disease. We used the 105 SARS virus sequences and 108 samples of other influenza virus types (the 'other virus' in Table 1) in this experiment and selected polyspecific probes. Used in a classification rule (3), these probes collectively gave the perfect classification of all testing sequences.

**Experiment 3.** Classification of lethal AI virus H5 & H9 and other influenza virus H subtypes
AI virus H5 and H9 subtypes cause a most fatal form of the disease [29], and they were separated from the other H subtypes of influenza virus in this experiment. 241 H5 and H9 target sequences and 1010 other H subtype sequences were used to select polyspecific probes for detecting AI virus H5 and H9 subtypes from the rest. In a classification rule (3), the selected probes collectively classified all testing sequences correctly.

**Experiment 4.** Monospecific Classification of N1, N2 and N3 influenza virus
The statement "monospecific neuraminidase (NA) subtype probes were insufficiently divers to allow confident NA subtype assignment" from [6] motivated us to design this experiment on multicategory and monospecific classification of influenza virus by N subtypes. We used the three influenza virus N subtypes with 30 or more samples in Table 1 and selected monospecific probes for their classification. Tested in a classification rule (5), the selected probes performed perfectly in classifying all testing sequences. Note that only a small number of monospecific probes were selected and proved 'needed' in this experiment.

## References

1. Stears, R., Martinsky, T., Schena, M.: Trends in microarray analysis. Nature Medicine **9**(1) (2003) 140–145
2. Eom, J.H., Park, S.B., Zhang, B.T.: Genetic mining of dna sequence structures for effective classification of the risk types of human papillomavirus (hpv). In Pal, N., Kasabov, N., Mudi, R., Pal, S., Parui, S., eds.: Lecture Notes in Computer Science. Volume 3316. Springer-Verlag, Berlin Heidelberg (2004) 1334–1343
3. Heller, R., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D., Davis, R.: Discovery and analysis of inflammatory disease-related genes using cdna microarrays. Proceedings of the National Academy of Sciences **94** (1997) 2150–2155
4. Liu, C.H., Ma, W.L., Shi, R., Ou, Y.Q., Zhang, B., Zheng, W.L.: Possibility of using dna chip technology for diagnosis of human papillomavirus. Journal of Biochemistry and Molecular Biology **36**(4) (2003) 349–353
5. Lee, Y., Lee, C.K.: Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics **19**(9) (2003) 1132–1139
6. Sengupta, S., Onodera, K., Lai, A., Melcher, U.: Molecular detection and identification of influenza viruses by oligonucleotide microarray hybridization. Journal of Clinical Microbiology **41**(10) (2003) 4542–4550
7. Vernet, G.: Dna-chip technology and infectious diseases. Virus Research **82** (2002) 65–71
8. Wang, D., Coscoy, L., Zylberberg, M., Avila, P., Boushey, H., Ganem, D., DeRisi, J.: Microarray-based detection and genotyping of viral pathogens. PNAS **99**(24) (2002) 15687–15692
9. Li, F., Stormo, G.: Selection of optimal dna oligos for gene expression arrays. Bioinformatics **17**(11) (2001) 1067–1076
10. Borneman, J., Chrobak, M., Vedova, G., Figueroa, A., Jiang, T.: Probe selection algorithms with applications in the analysis of microbial communities. Bioinformatics **17**(Suppl. 1) (2001) S39–S48
11. Rahmann, S.: Fast large scale oligonucleotide selection using the longest common factor approach. Journal of Bioinformatics and Computational Biology **1**(2) (2003) 343–361
12. Klau, G., Rahmann, S., Schliep, A., Vingron, M., Reinert, K.: Optimal robust non-unique probe selection using integer linear programming. Bioinformatics **20** (Suppl. 1) (2004) i186–i193
13. Boros, E., Hammer, P., Ibaraki, T., Kogan, A., Mayoraz, E., Muchnik, I.: An implementation of logical analysis of data. IEEE Transactions on Knowledge and Data Engineering **12** (2000) 292–306

14. Megiddo, N.: On the complexity of polyhedral separability. Discrete and Computational Geometry **3** (1988) 325–337
15. Ryoo, H., Jang, I.Y.: Milp approach to pattern generation in logical analysis of data. Machine Learning (2005) submitted.
16. Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D.: Basic local alignment search tool. Journal of Molecular Biology **215** (1990) 403–410
17. Wang, X., Seed, B.: Selection of oligonucleotide probes for protein coding sequences. Bioinformatics **19**(7) (2003) 796–802
18. Cortes, C., Vapnik, V.: Support vector networks. Machine Learning **20** (1995) 273–297
19. Ullman, J.: Pattern Recognition Techniques. Crane, London (1973)
20. Vapnik, V.: Statistical Learning Theory. Wiley-Interscience (1998)
21. Hammer, P.: Partially defined boolean functions and cause-effect relationships. Proceedings of the International Conference on Multi-Attribute Decision Making Via OR-Based Expert Systems (1986)
22. Caprara, A., Fischetti, M., Toth, P.: A heuristic method for the set covering problem. Operations Research **47**(5) (1999) 730–743
23. Nemhauser, G.L., Wolsey, L.A.: Integer and Combinatorial Optimization. Wiley-Interscience Series I Discrete Mathematics and Optimization. Wiley, New York (1988)
24. Muñoz, N., Bosch, F., de Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K., Snijders, P., C.J.L.M. Meijer, for the International Agency for Research on Cancer Multicenter Cervical Cancer Study Group: Epidemiologic classification of human papillomavirus types associated with cervical cancer. The New England Journal of Medicine **348**(6) (2003) 518–527
25. Bosch, F., Lorincz, A., Muñoz, N., Meijer, C., Shah, K.: The causal relation between human papillomavirus and cervical cancer. Journal of Clinical Pathology **55** (2002) 244–265
26. McFadden, S., Schumann, L.: The role of human papillomavirus in screening for cervical cancer. Journal of the American Academy of Nurse Practitioners **13** (2001) 116–125
27. Park, S.B., Hwang, S.H., Zhang, B.T.: Classification of the risk types of human papillomavirus by decision trees. In: Proceedings of the 4th International Conference on Intelligent Data Engineering and Automated Learning. (2003) 540–544
28. Snijder, E., Bredenbeek, P., Dobbe, J., Thiel, V., Ziebuhr, J., Poon, L., Guan, Y., Rozanov, M., Spaan, W., Gorbalenya, A.: Unique and conserved features of genome and proteome of sars-coronavirus, an early split-off from the coronavirus group 2 lineage. Journal of Molecular Biology **331** (2003) 991–1004
29. Koopmans, M., Wilbrink, B., Conyn, M., Natrop, G., van der Nat, H., Vennema, H., Meijer, A., van Steenbergen, J., Fouchier, R., Osterhaus, A., Bosman, A.: Transmission of h7n7 avian influenza a virus to human beings during a large outbreak in commercial poultry farms in the netherlands. Lancet **363** (2004) 587–593 www.thelancet.com.