OR Spectrum (2008) 30:249–268
DOI 10.1007/s00291-007-0089-0

REGULAR ARTICLE

A LAD-based method for selecting short oligo probes for genotyping applications

Kwangsoo Kim · Hong Seo Ryoo

Published online: 1 June 2007

Abstract Specializing a general framework of logical analysis of data for efficiently handling large-scale genomic data, we develop in this paper a probe design method for selecting short oligo probes for genotyping applications. When tested on genomic sequences obtained from the National Center of Biotechnology Information in various monospecific and polyspecific in silico experiments, the proposed probe design method was able to select a small number of oligo probes of length 7 or 8 nucleotides that perfectly classified all unseen testing sequences. These results demonstrate the efficacy of the proposed probe design method and illustrate the usefulness and potential a well-designed optimization-based probe selection method has in genotyping applications.

Keywords Oligo probes · Microarrays · LAD · Set covering · Classification · Optimization · SARS · AI

1 Introduction

Between November 1, 2002 and July 31, 2003, Severe Acute Respiratory Syndrome (SARS) virus infected 8,096 people and proved fatal to 774 worldwide.[1] The avian influenza (AI) virus subtype H5N1 alone infected 152 people worldwide between 2003

[1] From http://www.who.int/csr/sars/country/table2004_04_21/en/index.html, accessed on January 30, 2006.

K. Kim · H. S. Ryoo (✉)
Division of Information Management Engineering, Graduate School of Information Management and Security, Korea University, 1, 5-ka, Anam-dong, Seongbuk-ku, Seoul 136-713, South Korea
e-mail: hsryoo@korea.ac.kr

and January 2006, and 83 died of the disease.[2] Luckily, none of the outbreaks of SARS and AI infections at the beginning of the new millennium brought about the worst-case scenario. Alarmingly, influenza experts seem to agree that another pandemic may be imminent (Webby and Webster 2003) and, as of this writing, a fearful AI H5N1 virus continues to spread in part of Asia and Europe.

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 (e.g., Schena 1999; Stears et al. 2003). 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 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 (e.g., Eom et al. 2004; Heller et al. 1997; Lee and Lee 2003; Liu et al. 2003). 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 (e.g., Sengupta et al. 2003; Vernet 2002; Wang et al. 2002; Zhou et al. 2005).

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 obvious reasons. One, the background hybridization is minimized; hence true gene expression levels can be more accurately determined (e.g., Li and Stormo 2001). 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 (e.g., Borneman et al. 2001; Li and Stormo 2001). Short probes consisting of 15–25 nucleotides (nt) are used in genotyping applications (e.g., Stears et al. 2003). Having short optimal probes means a high genotyping accuracy in terms of both sensitivity and specificity (e.g., Li and Stormo 2001; Sengupta et al. 2003), hence can play a key role in genotyping applications. For example, in a pandemic, an effective method for selecting short optimal probes may be used in the mass production of a cost-efficient device for screening for the disease in suspected or susceptible hosts. Reverse genetics would be the most rapid means by which to produce an antigenically matched vaccine in a pandemic (Webby and Webster 2003). An effective probe selection methodology can identify conserved regions of a viral family and hence may prove useful in the preparation of a vaccine via reverse genetics. Furthermore, the methodology can promote the availability of affordable home testing kits for accurate and confidential diagnosis of genetic and infectious disease and allow advanced and adequate medical treatment planning for patients.

A well-studied problem in machine learning and data mining deals with the discovery of a classification rule for different types of data. The probe design, say, for genotyping applications, can be roughly stated as selecting oligo probes for detecting

[2] From http://www.who.int/csr/disease/avian_influenza/country/cases_table_2006_01_25/en/index.html, accessed on January 30, 2006.

a specific disease-agent in genomic sequences, hence falls into the realm of classical classification. Thus far, this interesting problem at the intersection of molecular biology and optimization has received relatively little attention from the optimization community, and systematic oligo design methods proposed so far are based on a simple greedy procedure (Herwig et al. 2000), the set covering-based classification methodology (Borneman et al. 2001), support vector machines (Lee and Lee 2003), an evolutionary algorithm (Eom et al. 2004; Lee et al. 2004), and mixed integer and linear programming (Klau et al. 2004), briefly summarizing.

From the perspective of numerical optimization, genomic data present an unprecedented challenge for supervised learning approaches for a number of reasons. To name a few, first, genomic data are long sequences over the nucleic acid alphabet $\Sigma = \{A,C,G,T\}$. 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. Furthermore, 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 (e.g., Borneman et al. 2001; Boros et al. 2000; Klau et al. 2004; Rahmann 2003). Owing to these and the nature of general data analysis and classification (Megiddo 1988), a supervised learning approach to classification of genomic data without specialized features for efficiently handling large-scale data is confronted by a formidable challenge.

Based on a general framework of logical analysis of data (LAD) from Ryoo and Jang (2005), 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 Borneman et al. (2001) and Klau et al. (2004), for instance. 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 $(0.25)^l$. Unless statistically significant, an $l-$mer appearing in multiple samples of one type and none or only a few of the sequences of the other type by chance is extremely small. Third, the proposed method does not rely on any extra tool, such as BLASTn (Altschul et al. 1990), a local sequence alignment search tool that is commonly used for probe selection (e.g., Sengupta et al. 2003; Wang et al. 2002; Wang and Seed 2003), or the existence of pre-selected representative probes (e.g., Sengupta et al. 2003). This makes the method truly stand-alone and free of problems that may possibly be caused by limitations associated with external factors. As mentioned earlier, the proposed probe design selects optimal probes via sequential solution of SC instances. Although SC is $\mathcal{NP}$-complete (Garey and Johnson 1979), its wide practical applications have invited an array of efficient (meta-)heuristic solution procedures to be developed.

Therefore, last, the proposed method is readily implementable for efficient selection of oligo probes.

This paper is organized as follows. In Sect. 2, we specialize a LAD framework from Ryoo and Jang (2005) for efficiently analyzing genomic sequences and develop an effective method for selecting short oligo probes. In Sect. 3, we test the proposed probe design algorithm in various in silico genotyping experiments using viral genomic sequences and report superb experimental results. To summarize, in all monospecific and polyspecific genotyping experiments on classification of viral pathogens using genomic sequences obtained from the National Center of Biotechnology Information website, the proposed probe design method selected a small number of probes of length 7 or 8 nt that perfectly classified all unseen testing sequences. Classifying the "noisy" human papillomavirus (HPV) sequences from the Los Alamos Laboratory by high and low risk types, the proposed probe design method selected optimal probes in a few CPU seconds that classified the testing sequences with 90.6% accuracy. For comparison, Eom et al. (2004) and Park et al. (2003) experimented with the same HPV dataset and reported the classification accuracy of 85.6 and 81.1%, respectively. These in silico results demonstrate efficacy and efficiency of the proposed oligo design method and further illustrate the usefulness and potential of a well-designed optimization-based probe design method in the forthcoming era of biotechnology. Finally, Sect. 4 concludes the paper with a few remarks.

Before proceeding, we refer interested readers to Schena (1999), Stears et al. (2003) and Vernet (2002) for background in microarray analysis and its usage in the diagnosis of infectious disease. Furthermore, as classification of more than two types of data can be accomplished by sequential classification of two types of data (see, for example, Cortes and Vapnik 1995; Ullman 1973; Vapnik 1998 and Sect. 3), we present the material below in the context of the classification of $+$ and $-$ types of data for convenience and without loss of generality.

2 Proposed probe selection method

The backbone of the proposed procedure is LAD. LAD is a relatively new supervised learning methodology that is based on Boolean logic, combinatorics and optimization. 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 Boros et al. (2000), Hammer (1986) and Ryoo and Jang (2005) for more background in LAD, we design a LAD-based method below for efficiently

handling and analyzing large-scale genomic data and selecting optimal oligo probes for genotyping applications.

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 Sect. 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, $l_j = a_j$ ($l_j = \bar{a}_j$) 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 $t(p_i) := \prod_{l_j = a_j, j \in N_t} p_{ij} \prod_{l_j = \bar{a}_j, j \in N_t} \bar{p}_{ij} = 1$ 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.

To aid in presentation, let us temporarily 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. Ryoo and Jang (2005) formulated a compact mixed integer and linear programming

(MILP) model below with respect to a reference sample $p_i, i \in S^\bullet, \bullet \in \{+, -\}$:

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

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.*

Proof First, via the first constraint of (MILP-2.i$^\bullet$) and the definition of J_i, we trivially have

$$\mathcal{P}(p_i) = \prod_{j \in N_t} p_{ij} = 1$$

for the reference observation $p_i, i \in S^\bullet$. Next, the second set of hard constraints yields that at least one of $p_{lj} = 0$ for $j \in N_t$ for each $p_l, l \in S^{\bar{\bullet}}$. This gives

$$\mathcal{P}(p_l) = \prod_{j \in N_t} p_{lj} = 0$$

for all $p_l, l \in S^{\bar{\bullet}}$, and completes the proof. □

Lemma 1 shows that any feasible solution of (MILP-2.i$^\bullet$) can be used to form a $\bullet$ pattern. Now, note that if $y_l = 0$ for $l \in S^\bullet \setminus \{i\}$ in the solution, then the $\bullet$ pattern $\mathcal{P}$ formed also distinguishes p_l from the $\bar{\bullet}$ observations. Therefore, with the objective of minimizing the sum of y_l's, the MILP model can be understood as a way to generate a $\bullet$ pattern that distinguishes (more or less) a maximum number of $\bullet$ observations from the $\bar{\bullet}$ observations. As easily seen, the number of 1's in the (optimal) solution determines the degree of the pattern generated.

As demonstrated in Ryoo and Jang (2005), this model efficiently generates patterns of all degree with equal ease, provided that the number of training samples used is moderate and that n is not a big number. Genomic data are large-scale in nature, however. 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 in general, we see that (MILP-2.i$^\bullet$) 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

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

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

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

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

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

where c_j ($j \in N$) are positive real numbers (refer to Remark 4).

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.*

Proof To show the result, we need to show that the conjunction of literals formed via (2) distinguishes at least one $\bullet$ observation from all $\bar{\bullet}$ observations. Toward the end, recall that $p_{ik} = 1(0)$ indicates the presence (absence) and the absence (presence) of probe k in the reference sequence selected p_i, $i \in S^\bullet$, and in p_k, $k \in S^{\bar{\bullet}}$, respectively. With the cover $(\mathbf{x}, \mathbf{y})$ of $(\mathrm{SC}_i^\bullet)$ on hand, let us subdivide the index set $N_t = \{j \in N : x_j = 1\}$ into two subsets $N_t^1 := \{j \in N_t : p_{ij} = 1\}$ and $N_t^0 := \{j \in N_t : p_{ij} = 0\}$.

Observe now that

$$\mathcal{P}(p_i) = \prod_{l \in N_t^1} p_{il} \prod_{l \in N_t^0} \bar{p}_{il} = 1$$

for $p_i, i \in S^\bullet$, hence $\mathcal{P}(p_i) = 1$ for at least one $\bullet$ observation.

Note in (1) that $a_j^{(i,k)} = 1$ if $p_{ij} \neq p_{kj}$ for $k \in S^{\bar{\bullet}}$. That is, $a_j^{(i,k)} = 1$ implies that exactly one of p_{ij} and p_{kj} equals 1 for p_i and $p_k, k \in S^{\bar{\bullet}}$. The cover $(\mathbf{x}, \mathbf{y})$ of $(\mathrm{SC}_i^\bullet)$ by definition satisfies all constraints of $(\mathrm{SC}_i^\bullet)$, and the hard constraints of the problem in the first set of cover inequalities require that at least one x_l in the cover is set to 1 among $l \in N$ with $a_l^{(i,k)} = 1$ for all $k \in S^{\bar{\bullet}}$. This in turn implies that at least one p_{kl} for $l \in N_t^1$ or $\overline{p}_{il}$ for $l \in N_t^0$ equals 0 for all $p_k, k \in S^{\bar{\bullet}}$ and yields

$$\mathcal{P}(p_k) = \prod_{l \in N_t^1} p_{kl} \prod_{l \in N_t^0} \bar{p}_{kl} = 0$$

for all $p_k, k \in S^{\bar{\bullet}}$, hence $\mathcal{P}(p_i) = 0$ for all $\bar{\bullet}$ observations. □

Note that $\mathcal{P}$ generated on the solution $(\mathbf{x}, \mathbf{y})$ of $(\mathrm{SC}_i^\bullet)$ via (2) also satisfies $\mathcal{P}(p_l) = 1$ for all $l \in S^\bullet \setminus \{i\}$ with $y_l = 0$. The following result is immediate.

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* $k \in N_t$*.*

As (MILP-2.i$^\bullet$), $(\mathrm{SC}_i^\bullet)$ is also formulated in reference to p_i for some $i \in S^\bullet$ and finds a cover that distinguishes most $\bullet$ observations from the $\bar{\bullet}$ observations. Therefore, although not identical, $(\mathrm{SC}_i^\bullet)$ can be seen as an SC version of (MILP-2.i$^\bullet$). Although smaller than the MILP model 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, owing to having a wide range of practical applications, SC has invited the development of an array of efficient (meta-)heuristic solution procedures (e.g. Caprara et al. 1999 and references therein) and any of these can be used for solving $(\mathrm{SC}_i^\bullet)$ (refer to Remark 1). From the computational point of view, therefore, $(\mathrm{SC}_i^\bullet)$ is much preferred over its MILP counterpart.

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

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

where c_j's are positive reals (again, refer to Remark 4).

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.*

Proof Same as the proof for Theorem 1. □

We immediately have the following result that can be used for efficiently identifying the $\bullet$ observations that are also distinguished from the $\bar{\bullet}$ observations by the pattern generated on the solution of (SC-pg$_i^\bullet$).

Lemma 3 *With a feasible solution* $\mathbf{x}$ *of* (SC-pg$_i^\bullet$), *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* $N_t = \{j \in N : x_j = 1\}$.

Note that (SC-pg$_i^\bullet$) can be considered as a relaxation of (SC$_i^\bullet$): to see this, project (SC$_i^\bullet$) onto the space of $\mathbf{x}$. Generally speaking, therefore, a feasible solution of (SC$_i^\bullet$) has more x_j's set to 1 in it than in a feasible solution of (SC$_i^\bullet$) formulated on the same data, hence tends to generate a higher degree pattern that generally explains a difference between the target and non-target sequences. As more $\bullet$ observations are distinguished from the $\bar{\bullet}$ observations at a time by a solution of (SC$_i^\bullet$), it is formulated and solved for a less number of times for generating a set of $\bullet$ patterns that collectively distinguish all $\bullet$ observations from the $\bar{\bullet}$ data in a dataset under analysis (refer to the oligo selection procedure detailed below). On the other hand, (SC-pg$_i^\bullet$) generates per solution a lower degree pattern that explains the specific difference between the reference $\bullet$ observation and the $\bar{\bullet}$ sequences and, hence, is formulated and solved for a more number of times for generating a set of $\bullet$ patterns. Overall, the two models select about the same number of probes. However, as (SC-pg$_i^\bullet$) is much smaller in size, hence, is more efficiently solved, and because a high specificity is desired in genotyping applications, we prefer (SC-pg$_i^\bullet$) for selecting genotyping oligo probes.

Using (SC-pg$_i^\bullet$), we design one simple oligo probe selection procedure below, where $P^\bullet$ denotes 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 $P^\bullet \leftarrow P^\bullet \cup \{\mathcal{P}\}$.
      - set $S \leftarrow S \setminus \{i\} \setminus \{j \in S, j \neq i : p_{jk} = p_{ik}, \forall k \in N_t\}$.
    end while
  end for
end

The following is immediate.

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

A few remarks are due now.

Remark 1 Simply put, the number of 1's in the covers generated via **procedure** `SC-pg` determines the number of probes to be used for a specific genotyping purpose. In other words, the quality of an SC solution determines the cost of genotyping applications.

SC is a well-known $\mathcal{NP}$-complete problem (Garey and Johnson 1979). Owing to having a wide range of practical applications (despite its simple structure), SC has invited an array of (meta-)heuristic solution procedures to be developed for its efficient heuristic solution (e.g., Caprara et al. 1999 and references therein) and any of these can be used for solving (SC-pg$_i^\bullet$). In fact, the genotyping accuracy is not affected at all as long as the covers found are near-optimal and "good enough" (see results in the following section) and this was the rationale behind our developing SC-based probe selection models in this paper: recall that probe selection is a large-scale combinatorial optimization problem in nature.

Furthermore, the efficiency of SC heuristic solution procedures allows one to apply **procedure** `SC-pg` or the similar directly to the binarized data to generate patterns without going through the feature selection phase. This is another benefit the SC-based pattern generation offers over its MILP counterparts from Ryoo and Jang (2005) or the standard term-enumeration-based procedure for generating patterns in the LAD literature (e.g., Boros et al. 2000).

Remark 2 If constraint j in (SC-pg$_i^\bullet$) has all zero coefficients, the SC instance is infeasible. 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 this holds, contradiction-free 0–1 clones of the original data can always be obtained by using oligos of longer length for data binarization.

Remark 3 If desired, the hybridization affinity of probes can be ensured in a number of ways, including the following. First, during data binarization, one can remove from further consideration each l-mer with the GC content less than a prescribed level or with the melting temperature calculated via, for example, the formula found in Wang and Seed (2003) that falls outside a certain prescribed range from the median melting temperature of all l-mers generated. Next, the proposed LAD-based method can be applied to select an optimal set of probes on the surviving l-mers that are "compatible" in terms of their hybridization behavior.

Remark 4 (SC-pg$_i^\bullet$) is a general-purpose model and can be specialized to select a minimal set of optimal oligo probes by any quantifiable probe selection criterion. For example, one may use the longest common factors from Rahmann (2003) or the OVL scores from Herwig et al. (2000) for c_j values in (SG-pg$_i^\bullet$) to select probes by the (dis-)similarity preference. One may use, for example, the Shannon entropy scores from Herwig et al. (2000) for c_j's and incorporate the complexity of oligos in probe selection.

2.3 Classification rules

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. First, in polyspecific genotyping applications (see, for example, Experiment 4 in Sect. 3.2), we form the standard LAD classification rule (Boros et al. 2000)

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

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

For the monospecific genotyping, we use a strict classification rule. Specifically, for classification of two viral (sub-)types (see, for example, Experiment 1 in Sect. 3.2), we form a decision rule by

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

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 the monospecific classification of more than two viral (sub-)types $k = 1, \ldots, m$ (see, for example, Experiment 7 in Sect. 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 $\Delta^k = 0$ for all k, then we fail to assign a class to sequence p.

3 In silico experiments

In this section, we extensively test the proposed probe design for classification of viral disease-agents in in silico setting. To make these experiments as "realistic" as possible, we design each of these experiments based on information from the literature and the official website of the World Health Organization (WHO) and use viral genomic sequences obtained from the National Center for Biotechnology Information (NCBI) and human papillomavirus (HPV) sequences from the Los Alamos National Laboratory. To be more specific about the data used, we obtained the HPV data from the Los Alamos National Laboratory site for illustrative and comparative purposes. These data correspond to the 72 high and low risk HPV sequences that are used in Eom et al. (2004) and Park et al. (2003). Although some of these manually classified virus sequences contain classification errors (Eom et al. 2004), we used the data with

Table 1 Viral sequences used in experiments

Viral sequence	Number	Length		
		Minimum	Average ± standard deviation	Maximum
Human papillomavirus (HPV):				
High risk HPV	18	449	7,365 ± 1,730	7,989
Low risk HPV	54	455	7,198 ± 1,683	8,027
SARS coronavirus	105	29,350	29,692 ± 91	29,765
Coronavirus	39	9,203	29,013 ± 3,569	31,526
Other virus:				
Human respiratory syncytial virus	10	13,933	15,091 ± 386	15,226
Human adenovirus	32	34,125	35,215 ± 618	36,015
Human parainfluenza virus	4	15,646	15,652 ± 3	15,654
Human rhinovirus (A, B)	8	7,102	7,157 ± 36	7,212
Influenza virus (A, B, C)	53	838	1,701 ± 527	2,368
Influenza virus hemagglutinin (H) subtype:				
H1	137	1,698	1,749 ± 24	1,778
H3	660	1,695	1,735 ± 21	1,768
H5	148	1,677	1,721 ± 25	1,779
H7	77	1,659	1,690 ± 27	1,792
H9	93	1,683	1,704 ± 26	1,742
H else (2, 4, 6, 8, 11, 12, 13, 16)	65	1,689	1,742 ± 29	1,773
Influenza virus neuraminidase (N) subtype:				
N1	218	1,344	1,410 ± 39	1,463
N2	1,050	1,341	1,434 ± 28	1,467
N3	44	1,326	1,411 ± 29	1,460
N else (4, 5, 6, 7, 8, 9)	64	1,341	1,434 ± 25	1,467

their classification from Park et al. (2003) to allow a comparison among our result and results reported in Eom et al. (2004) and Park et al. (2003). For the experiments on genotyping viral pathogens, we used genomic sequences of SARS virus, influenza virus classified by their hemagglutinin (H) and neuraminidase (N) types (influenza viruses are typed according to their H and N surface glycoproteins), coronavirus and other viral agents of disease with SARS-like symptoms. In Table 1, we provide the number and the length (the minimum, average ± 1 standard deviation and maximum length) of each type of the genomic data 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. If the length of oligos turned out to be not long enough during the pattern generation stage (see Remark 2), the data binarization stage was repeated with the value of l incremented

by 1 and this process was repeated 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 did not consider any oligo picking criterion that is non-theoretical in nature (refer to Remark 4) and selected a minimal set of oligo probes with using $c_j = 1$ for all $j \in N$. For solving the unicost (SC-pg$_i^\bullet$)'s generated, we used for ease of implementation the textbook heuristic procedure (e.g., Nemhauser and Wolsey 1988) that selects one variable at a time by the rule

$$k \leftarrow \text{argmax}\left\{j \in N,\ x_j = 0 : \left|I_j \cap M_u\right|\right\},$$

where I_j denotes the index of rows k with $a_j^{(i,k)} = 1$ and M_u denotes the set of rows that are not yet covered by the partial cover $\mathbf{x}$ on hand.

In each of the experiments in this section, in order to fairly assess the classification capabilities of oligo probes selected by the proposed probe design procedure, we

1. randomly selected 90% of the target and of the non-target data to form a training set of sequences;
2. binarized the training data;
3. selected optimal oligo probes on the training data via **procedure** `SC-pg`;
4. formed a classification rule by one of (3), (4) and (5) with the selected oligo probes;
5. used the classification rule to (sub-)type each of the reserved testing sequences, consisting of the remaining 10% of the target and the non-target sequences; and
6. repeated steps above 20 times to obtain the average testing performance and other relevant information of the experiment.

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

3.1 A comparative experiment: classification of high and low risk HPV

Infection with HPV is the main cause of cervical cancer, the second most common cancer in women worldwide (Bosch et al. 2002; Muñoz et al. 2003). 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 (e.g., McFadden and Schumann 2001).

We applied the proposed probe design method on the 72 HPV sequences downloaded from the Los Alamos National Laboratory with their classification found in Table 3 of Park et al. (2003). 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 and the other tables in this section, the target (+) and the non-target (−) virus types of the experiments are specified in the first column. Then, the tables provide two bits of information on the candidate oligos, namely, the length l and the average and

Table 2 Polyspecific classification of high and low risk HPV

Experiment	l-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns	accuracy[a,b]
High risk HPV (+) vs. Low risk HPV (−)	8	58,359.9 ± 130.4	18.7 ± 1.7 22.8 ± 1.6	Degree 1 & 2 patterns Degree 1 & 2 patterns	90.6 ± 9.8

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

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: recall that we skip the feature selection stage of LAD (see Remark 1). 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 format "the average ± 1 standard deviation" of the percentage of the correct classifications of the unseen sequences.

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 unseen HPV samples. For comparison, the same HPV dataset was used in Eom et al. (2004) and Park et al. (2003) for the classification of HPV by high and low risk types. In brief, the probe design methods of Eom et al. (2004) and Park et al. (2003) 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 Park et al. 2003). 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 Experiments on genotyping viral pathogens

The proposed probe design method was extensively 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 summarize the results from some of these experiments in this section.

Before proceeding, we briefly illustrate the benefit of probe selection via (SC-pg$_i^\bullet$) from the computational point of view with Experiment 5 below. For the purpose, let us first recall that probe selection is a combinatorial optimization problem. Therefore, for the selection of oligo probes for differentiating lethal AI virus H5 and H9 from the other AI virus H subtypes in Experiment 5, a supervised learning method based on a complete pairwise differencing of the target and non-target training sequences (e.g., Borneman et al. 2001; Boros et al. 2000; Klau et al. 2004; Rahmann 2003)

Table 3 Monospecific classification of SARS virus and coronavirus, a phylogenetically closest sibling of SARS

Experiment 1	l-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
SARS virus (+)	8	57,745.3 ± 306.1	1 ± 0	Degree 1	100 ± 0
Coronavirus (−)			1 ± 0	Degree 1	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

would require solving one or more combinatorial optimization problems with between $(148+93)\times(137+660+77+65) = 226,299$ and $137\times660\times148\times77\times93\times65 \approx 6.23\times10^{12}$ rows (refer to Table 1 above for the numbers of the target and non-target viral sequences) and with at least 39, 056 0–1 decision variables (see Table 7 for the average number of $l-$mers generated in this experiment). For Experiment 5, we note in comparison that the largest (SC-pg$_i^\bullet$) instance generated and solved by **procedure** `SC-pg` had $\max\{148+93, 137+660+77+65\} = 939$ rows and 39, 056 columns.

Experiment 1 SARS virus vs. coronavirus.

SARS virus is phylogenetically most closely related to group 2 coronavirus (Snijder et al. 2003). 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 (see Table 3).

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 107 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 (see Table 4).

Experiment 3 Classification of pathogenic AI virus H7 and other influenza virus H subtypes.

AI virus H7N7 is highly pathogenic with the capacity to pass from human-to-human, and this raised concerns for a possible viral reassortment with human influenza H1N1 and H3N2 strains during a large outbreak of H7N7 infection in the Netherlands in 2003 (Koopmans et al. 2004; Webby and Webster 2003).

Based on information from Koopmans et al. (2004), we replicated the classification of H7 and other influenza virus H subtypes in this experiment by using 77 H7 sequences and 1,103 other H subtype samples. Polyspecific probes were selected and tested in a classification rule (3) to give the perfect classification rate (see Table 5).

Experiment 4 Classification of pathogenic AI virus H5 and H7 and other influenza virus H subtypes.

H5 and H7 have an ominous capacity to pass from human to human (http://www.who.int; Webby and Webster 2003). This experiment, using 225 H5 and H7 viral samples and 955 other H subtype sequences, selected polyspecific probes for detecting the two pathogenic H subtypes of the AI virus from the other influenza virus H subtypes and vice versa. A classification rule was formed by (3) for testing the selected probes, and we obtained the perfect testing result (see Table 6).

Experiment 5 Classification of lethal AI virus H5 and H9 and other influenza virus H subtypes.

AI virus H5 and H9 subtypes cause a most fatal form of the disease (Koopmans et al. 2004), and they were separated from the other H subtypes of influenza virus in this experiment. 241 H5 and H9 target sequences and 939 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 (see Table 7).

Experiment 6 Monospecific classification of SARS, human influenza H1, human influenza H3, AI H5 and AI H7 virus.

This multicategory classification experiment selects monospecific probes for distinguishing one from another a few notorious viral pathogens. We used 103 SARS virus, 137 human influenza virus H1, 660 human influenza virus H3, 148 lethal AI virus H5 and 77 pathogenic AI virus H7 sequences and selected monospecific probes for each virus type in sequential binary classification of "one type against the rest." The selected probes were tested in a classification rule (5) to classify the testing sequences p by a strict decision rule of "assign class i to p only if one or more probes selected for virus type i is found in p while none of the probes selected for the other types are not" and gave the perfect classification result (see Table 8; Note that only a small number of monospecific probes were selected, as in Experiment 1).

Experiment 7 Monospecific Classification of N1, N2 and N3 influenza virus.

The statement "monospecific neuraminidase (NA) subtype probes were insufficiently diverse to allow confident NA subtype assignment" from Sengupta et al. (2003) 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 (see Table 9; note again that only a small number of monospecific probes were selected and proved "needed" in this experiment, as in the other two monospecific genotyping experiments, Experiments 1 and 6).

4 Concluding remarks

The problem of probe design for hybridization-based experiments is an interesting problem lying at the intersection of molecular biology and optimization but has received relatively little attention from the OR community. In this paper, we specialized a general LAD framework from Ryoo and Jang (2005) for efficiently handling

Table 4 Classification of SARS virus and influenza virus that cause disease with SARS-like symptoms

Experiment 2	*l*-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
SARS virus (+)	8	64,141.5 ± 36.5	1 ± 0	Degree 1	100 ± 0
Influenza virus (−)			10.1 ± 0.8	Degree 1 only	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

Table 5 Classification of highly pathogenic H7 AI virus (with capacity to pass from human to human) and other H subtypes of influenza virus

Experiment 3	*l*-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
H7 (+)	7	14,724.2 ± 30.9	1 ± 0	Degree 1	100 ± 0
Other H strains (−)			7 ± 1	Degree 1 only	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

Table 6 Classification of highly pathogenic H5 and H7 AI virus and other H subtypes of influenza virus

Experiment 4	*l*-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
H5 & H7 (+)	8	39,164.4 ± 333	15.7 ± 1.5	Degree 1 only	100 ± 0
Other H strains (−)			27.6 ± 1.2	Degree 1 only	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

Table 7 Classification of fatal H5 & H9 AI virus and other H subtypes of influenza virus

Experiment 5	*l*-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
H5 & H9 (+)	8	39,056 ± 398.3	6.7 ± 0.5	Degree 1 only	100 ± 0
Other H strains (−)			21.6 ± 1.3	Degree 1 only	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

Table 8 Monospecific classification of SARS virus and H1, H3, H5 and H7 subtypes of influenza virus

Experiment 6	l-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
SARS virus	8	45,259 ± 527	1 ± 0	Degree 1	100 ± 0
H1			1 ± 0	Degree 1	
H3			2.9 ± 0.3	Degree 1 only	
H5			3 ± 0	Degree 1 only	
H7			1 ± 0	Degree 1	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

Table 9 Monospecific classification of N1, N2 and N3 subtypes of influenza virus

Experiment 7	l-mers used		Probes selected		Testing
	l	Number[a]	Number[a]	Patterns generated	accuracy[a,b]
N1	7	13,151 ± 39.3	3 ± 0	Degree 1	100 ± 0
N2			3.7 ± 0.5	Degree 1 only	
N3			1 ± 0	Degree 1	

[a] In format average ± standard deviation

[b] Percentage of correct classifications of testing/unseen data

large-scale genomic data and developed a probe design method for selecting short oligo probes for genotyping applications. Extensively tested on genomic sequences obtained from the National Center of Biotechnology Information and the Los Alamos National Laboratory in various monospecific and polyspecific *in silico* experiments, the proposed probe design method was able to select a small number of oligo probes of length 7 or 8 nucleotides that performed superbly in classifying unseen testing sequences. These *in silico* results demonstrate the efficacy of the proposed oligo design method. Experimental results further illustrate a huge potential a well-designed optimization-based probe design method has in hybridization-based genotyping applications.

Collaborative research activities are planned to realize the in silico performance of the proposed probe design method on microarrays and in real hybridization experiments. Also, we plan to investigate the possibility of exploiting frequently used oligo selection criteria (e.g., Herwig et al. 2000; Lee et al. 2004; Li and Stormo 2001; Rahmann 2003) within the proposed probe design framework to further improve its effectiveness in terms of the number of probes needed.

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

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Borneman J, Chrobak M, Vedova GD, Figueroa A, Jiang T (2001) Probe selection algorithms with applications in the analysis of microbial communities. Bioinformatics 17(Suppl 1):S39–S48

Boros E, Hammer PL, Ibaraki T, Kogan A, Mayoraz E, Muchnik I (2000) An implementation of logical analysis of data. IEEE Trans Knowl Data Eng 12:292–306

Bosch FX, Lorincz A, Muñoz N, Meijer CJLM, Shah KV (2002) The causal relation between human papillomavirus and cervical cancer. J Clin Pathol 55:244–265

Caprara A, Fischetti M, Toth P (1999) A heuristic method for the set covering problem. Oper Res 47(5): 730–743

Cortes C, Vapnik VN (1995) Support vector networks. Mach Learn 20:273–297

Eom J-H, Park S-B, Zhang B-T (2004) Genetic mining of dna sequence structures for effective classification of the risk types of human papillomavirus (hpv). In: Pal NR, Kasabov N, Mudi RK, Pal S, Parui SK, (eds) Lecture notes in computer science, vol 3316. Springer, Berlin Heidelberg, pp 1334–1343

Garey MR, Johnson DS (1979) Computers and intractability: a guide to the theory of $\mathcal{NP}$–completeness. Freeman, New York

Hammer PL (1986) Partially defined boolean functions and cause-effect relationships. In: Proceedings of the international conference on multi-attribute decision making via OR-based expert systems, April 1986

Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW (1997) Discovery and analysis of inflammatory disease-related genes using cdna microarrays. Proc Nat Acad Sci 94:2150–2155

Herwig R, Schmitt AO, Steinfath M, O'Brien J, Seidel H, Meier-Ewert S, Lehrach H, Radelof U (2000) Information theoretical probe selection for hybridisation experiments. Bioinformatics 16(10):890–898

Klau GW, Rahmann S, Schliep A, Vingron M, Reinert K (2004) Optimal robust non-unique probe selection using integer linear programming. Bioinformatics 20(Suppl 1):i186–i193

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 (2004) Transmission of h7n7 avian influenza a virus to human beings during a large outbreak in commercial poultry farms in the netherlands. Lancet 363:587–593 http://www.thelancet.com.

Lee I-H, Kim S, Zhang B-T (2004) Multi-objective evolutionary probe design based on thermodynamic criteria for hpv detection. In: Zhang C, Guesgen HW, Yeap WK (eds) Lecture notes in artificial intelligence, vol 3157. Springer, Berlin Heidelberg, pp 742–750

Lee Y, Lee C-K (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics, 19(9):1132–1139

Li F, Stormo GD (2001) Selection of optimal dna oligos for gene expression arrays. Bioinformatics 17(11):1067–1076

Liu C-H, Ma W-L, Shi R, Ou Y-Q, Zhang B, Zheng W-L (2003) Possibility of using dna chip technology for diagnosis of human papillomavirus. J Biochemistry Mol Biol 36(4):349–353

McFadden SE, Schumann L (2001) The role of human papillomavirus in screening for cervical cancer. J Am Acad Nurse Pract 13:116–125

Megiddo N (1988) On the complexity of polyhedral separability. Discrete Comput Geom 3:325–337

Muñoz N, Bosch FX, de Sanjosé S, Herrero R, Castellsagué X, Shah KV, Snijders PJF, Meijer CJLM, (2003) For the International Agency for Research on Cancer Multicenter Cervical Cancer Study Group Epidemiologic classification of human papillomavirus types associated with cervical cancer. New Engl J Med 348(6):518–527

Nemhauser GL, Wolsey LA (1988) Integer and combinatorial optimization. Wiley-Interscience Series I: discrete mathematics and optimization. Wiley, New York

Park S-B, Hwang S-H, Zhang B-T (2003) 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, pp 540–544

Rahmann S (2003) Fast large scale oligonucleotide selection using the longest common factor approach. J Bioinform Comput Biol 1(2):343–361

Ryoo HS, Jang I-Y (2005) Milp approach to pattern generation in logical analysis of data (submitted)

Schena M (1999) DNA microarray: a practical approach. Oxford University Press, Oxford

Sengupta S, Onodera K, Lai A, Melcher U (2003) Molecular detection and identification of influenza viruses by oligonucleotide microarray hybridization. J Clin Microbiol 41(10):4542–4550

Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE (2003) Unique and conserved features of genome and proteome of sars-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol 331:991–1004

Stears RL, Martinsky T, Schena M (2003) Trends in microarray analysis. Nat Med 9(1):140–145

Ullman J (1973) Pattern recognition techniques. Crane, London

Vapnik VN (1998) Statistical learning theory. Wiley-Interscience, New York

Vernet G (2002) Dna-chip technology and infectious diseases. Virus Res 82:65–71

Wang X, Seed B (2003) Selection of oligonucleotide probes for protein coding sequences. Bioinformatics 19(7):796–802

Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL (2002) Microarray-based detection and genotyping of viral pathogens. PNAS 99(24):15687–15692

Webby RJ, Webster RG (2003) Are we ready for pandemic influenza? Science 302:1519–1522

Zhou YM, Yang RQ, Tao SC, Li Z, Zhang Q, Gao HF, Zhang ZW, Du JY, Zhu PX, Ren LL, Zhang L, Wang D, Guo L, Wang YB, Guo Y, Zhang Y, Zhao CZ, Wang C, Jiang D, Liu YH, Yang HW, Rong L, Zhao YJ, An S, Li Z, Fan XD, Wang JW, Cheng Y, Liu O, Zheng Z, Zuo HC, Shan QZ, Ruan L, Lu ZX, Hung T, Cheng J (2005) The design and application of dna chips for early detection of sars-cov from clinical samples. J Clin Virol 33(2):123–131