Information theoretic approach for microarray based pathogen detection

Joshy George
Genome Institute of Singapore
60, Biopolis Street #02-01
Singapore 138672

Wing-Kin Sung
Department of Computer Science
National University of Singapore
3 Science Drive 2, Singapore 117543

Wah-Heng Lee
Genome Institute of Singapore
60, Biopolis Street #02-01
Singapore 138672

Vinsensius B. Vega
Genome Institute of Singapore
60, Biopolis Street #02-01
Singapore 138672

Lance D. Miller
Genome Institute of Singapore
60, Biopolis Street #02-01
Singapore 138672

*A*bstract:- The accurate and rapid detection of viral and bacterial pathogens in human patients and populations is very important. Currently used techniques for detecting pathogens are not satisfactory in terms of speed, accuracy and sensitivity. Recently oligo microarray (DNA chip) [1–3] technology has been proposed as a potential solution to this problem. However, effectiveness of the microarray for pathogen detection depends on the probes selected. Previous work used a heuristic based approach to select probes for this purpose. In this work we demonstrate the use of information theoretic approach to select optimal probes for detecting pathogens. We also demonstrate that probes can be selected in such a way that partial characterization of newly evolved pathogens is possible. Simulation results to demonstrate the correctness of the proposed solution is also included.

*K*ey-Words:- Information content, probe design, pathogen detection, microarray

1 Introduction

Determining the causative agent of a disease accurately is a prerequisite for giving proper medical care. It is imperative that this process is completed as early as possible. Not being able to do so can certainly effect the patient involved as well as the population in case the disease is contagious. The recent SARS epidemic is a case in point. In this paper we propose an oligo microarray hybridization based approach for the rapid detection and identification of both viral and bacterial pathogens. All pathogens contain unique molecular regions within their genome. Oligomers derived form this area can be used to detect them. Selecting probes from highly conserved regions of different family and genus members will enable the partial characterization of some novel pathogens as explained later.

1.1 Previous Works

Probe selection algorithms such as those proposed by Li and Stormo [12], Kaderali and Schliep [13] and Rahmann [10] find unique probes for each gene, resulting in a minimal set of probes used for measurement of transcript level of the genes in a given sample. Precise and exact measurement of transcript level of each gene is the key requirement. Pathogen detection problem that we are considering has a different requirement which will be clear in the following discussion.

Recently, microarray based detection of pathogens was suggested and demonstrated by a group in University of California at San Francisco for a subset of viral genomes [5]. The same group later extended their work to include all known sequenced human, animal and plant viruses [6]. Their approach to select oligos is based on heuristics and the optimality of their probe selection method was not demonstrated mathematically.

Others attempted to model this problem as string barcoding problem [4, 7]. In 2003, Schliep *et. al.* [11] recognized the potential usefulness of cross hybridization and proposed the idea of group testing to find a small set of probes and analyze hybridization outcome for robust detection of presence of target sequences. In these approaches, the goal is mainly to identify the possible presence of known sequences. They do not have the capability to characterize newly evolved pathogens.

Unique to our probe design and subsequent post hybridization analysis approaches is the additional consideration for *a priori* classifications of genomes, which could be introduced based on the existing taxonomy or formulated by any other means. Partial characterization of previously unknown pathogens (with completely unknown genomic sequences) is made possible because of this.

1.2 Our Result

We have developed a program which will optimally select a set of probes that can be used to discover pathogens that are present in a biological sample. We have also developed a program that will predict the most probable pathogen that is present based on the microarray data after hybridization. The downstream analysis also includes the possibility of incorporating additional information about the probable class of genomes present.

We evaluated our probe selection and analysis technique by simulating the hybridization process. We randomly selected various percentages of know pathogens and simulated the hybridization process. In every case we were able to identify the pathogen if we had at least 70% of the pathogen genome in our sample.

We were also able to demonstrate the potential for partial characterization of the unknown pathogen as follows. We designed the chip without including the SARS genome and later tried to hybridize the SARS genome to the chip set. We were able to correctly predict that the SARS genome was a potentially new virus that may belong to the coronavirus genus and coronaviridae family.

2 Probe Design Problem

This section introduces the motivation behind the construction of pathogen chip, formalizes the associated probe selection problem, and outlines our solution.

2.1 Problem Description

The goal of devising a microarray-based pathogen detection platform is to provide a rapid and accurate mechanism for determining the existence of viral or bacterial pathogen in a given sample. Such pathogen chips should ideally be able to determine whether pathogens with known genomes are present, i.e. given a sample containing genomic fragments of a known pathogen, its associated probes are “lighted up”. However, since only a very minute fraction of pathogens is known, we will inevitably encounter the previously uncharacterized ones in the samples. In such instances, we would like to discover as much information as possible about the unknown pathogen. One important piece of information is to figure out where in a given groupings (e.g. taxonomy) of the known pathogens that the new pathogen is most likely to be categorized into. This information could be of great value to the scientist trying to find a defence against the pathogen.

Thus, the challenge is to select oligonucleotide probes to be put in the pathogen chip, such that hybridization with a given pathogen would reveal (1) the highly probable genome(s) of the pathogen and/or (2) the likely group(s)/class(es) that the pathogen might belong to.

2.2 Problem Definition

2.2.1 Probe selection for known pathogens

Given a set of genomes $G = \{g_1, ..., g_N\}$, the task is, for each genome g_i, to select a set of M_i length-t probes $P_i = \{p_1, ..., p_{M_i}\}$ that (as much as possible) satisfies the (1) Homogeneity, (2) Sensitivity, and (3) Specificity criteria as described below.

Homogeneity: Temperature is one of the important experiment conditions to ensure a probe can hybridize. We select probes whose melting temperature is close to the experiment temperature. GC rich sequences are susceptible to non-specific interactions that may reduce reaction efficiency. Thus, the GC content of good probes should not be too high or too low [9].

Sensitivity: Sensitivity, the ability to detect low abundance mRNAs, is a key performance feature of microarrays. Probes that form significant secondary structures jeopardize sensitivity. Thus it is important to reject probes with high self complementariness and select probes with minimal secondary structure. To do this, the free energy for each probe is computed based on the nearest neighbor model [8]. The free energy for each probe should be as high as possible.

Specificity: The specificity criterion refers to the uniqueness of the probes in each probe set P_i to each genome g_i. Bear in mind that this criterion may not be able to be met completely as two distinct genomes g_i and g_j might be highly similar, e.g. g_j could very well be a strain variant of g_i. Nevertheless, the probe sets P_i and P_j might be able to effectively distinguish the two genomes. As such, we define specificity of a probe set P_i to a genome as the total information carried by the probe set P_i about the existence of genome g_i. A more involved discussion is presented in Section 2.2.3. Note also that due to space constraint of the microarray, only a portion of potential probes for g_i might be included in P_i, making some sort of probe ranking mechanism necessary.

2.2.2 Probe selection for assessing unknown pathogens

The input given in this problem is similar to the previous problem, i.e. the set of genomes G, with an additional set S of disjoint subsets of G, i.e. $S = \{S_1, ..., S_K | S_i \subseteq G \bigwedge \bigcup_{i=1}^{K} S_i = G \bigwedge \forall x, y, x \neq y : S_x \bigcap S_y = \emptyset\}$. Each subset S_i represents a grouping of genomes, which could be based on the present biological taxonomy or sequence similarity among the genomes or other classifications/partitions. For a given grouping S, the pathogen chip should suggest the highly probable groups that an unknown pathogen might be classified into. Hence, the aim is to select, for each S_i, a set of length-t probes $P_{S_i} = \{p_1, ..., p_{M_{S_i}}\}$ that as much as possible satisfies the (1) Homogeneity, (2) Sensitivity, and (3) Specificity criteria. These criteria are similar to those described earlier in Section 2.2.1 with the exception of the specificity criteria. Here specificity refers to the total information capacity of a probe set P_{S_i} regarding the conserved regions among the genomes in set S_i. Note that the groups here are given a priori, as part of the user's input. Clearly, the groupings here are different from that discussed in [11], which uses randomized groupings to closely estimate the posterior probability of the presence of each target sequence. Note also that the problem previously described in Section 2.2.1 is actually a special case of this problem where each S_i contains exactly one genome, e.g. $\forall i : S_i = \{g_i\}$.

2.2.3 Total information capacity of a probe set

One could model the probe design problem as the selection of the most informative questions regarding the existence of a pathogen. Each probe is a question asked about the pathogen and we get a yes/no answer to our question from the microarray hybridization. We get some information from every answer we get. Under the settings of Information Theory, the goodness or the capacity of a message is defined as $I = -\log_2(p)$ where p is the probability of that event to occur. To exemplify, suppose that we have two probes p_1, which occurs in genome g_1, and p_2, which is found in genomes g_1 and g_2, and assume that all N genomes have an equal probability to appear. The information content of p_1 and p_2 are $I(p_1) = -\log_2\left(\frac{1}{N}\right)$ and $I(p_2) = -\log_2\left(\frac{2}{N}\right)$ respectively, since when p_1 "lights up" it could only means g_1 while for p_2 it could mean either g_1 or g_2.

Hence, the specificity criterion in pathogen chip probe design can be stated as trying to maximize the information content of a probe set P_i regarding its associated genome g_i among the genomes in G or, in a more general framework, the information capacity of a probe set P_{S_i} regarding the group of genomes S_i among other groups in S.

2.3 Comparison with probe design problem for genes

Cautious readers might question the necessity of finding the probe set P_i (as compared to just finding one probe) for each genome g_i and argue that the probe selection problem for pathogen chip is similar to probe selection problem for genes in a genome. While this problem can in fact be modeled as a probe design problem for genes in a genome by simply treating each g_i as an unusually large gene, there is an added dimension to probe design for pathogen chip. In the design of probes for genes in a genome, the number of genes is usually huge, to the order of tens of thousands and that the goal is to measure the relative amount of mRNA expressed in the sample. As such, ideally we only want one probe per gene. Whereas in the pathogen chip, we only wish to measure the presence of pathogens and may initially ignore the amount of pathogen present. Further, pathogens are likely to have a much higher mutation rate than genes. It is thus only natural to rely on a set of probes to detect the presence of pathogens. The number of known pathogens is also small, in order of hundreds or at most a couple of thousands, which then allows for multiple probes per genome to be fit in one chip. Doing so would also allow the possibility of partial characterization of unknown pathogen.

2.4 Probe design algorithm

This paper proposes a multi step approach to the pathogen chip design. The first step in designing the pathogen chip involved downloading from NCBI all of the taxonomic annotation and sequence identifiers for the hundreds of thousands of viral sequences available. The taxonomic information was then manually curated and all viral genomes which are known to infect animals (from insects to humans) were included in the list.

The next step in the algorithm is based on probe elimination. Initially, we assume that for every genome g_i, every length-m substring of g_i is a feasible probe. "Bad" probes are filtered out using the following steps:

1. Filter out redundant probes in every genome.

2. Filter out probes which fail to satisfy homogeneity criterion.

3. Filter probes which fail to satisfy sensitivity criterion.

4. Filter probes that can hybridize to human genome.

Once we have the list of good probes for every genome we can use the information carrying capacity of the probes to select the best among them so as to identify the unique pathogen present. We also include probes that are able to predict the family, genus and species of the pathogen present. We will demonstrate how these probes can help us to partially characterize newly evolved pathogens. Partial characterization involves predicting the family and genus of the newly evolved pathogen. This information can be of immense value for the scientist trying to find a defense against the disease.

We discuss Steps 1 to 4 in Sections 3. In Section 4 we introduce the notion of information capacity of the probe and explain how this can be used to select the most optimal set of probes. We then present the strategy for downstream analysis of the array data in Section 5. Simulation results are given in Section 6.

3 Probe Filtering

3.1 Non-redundant probe filtering

The goal of this step is to obtain the complete set of possible probes Q_i for each genome g_i. We started with the set of all possible oligos of the specified length for every pathogen of interest. We can save a lot of computation time in later stage by throwing away redundant probes from every genome. These probes will not give any additional information about anything so throw them away as the first step. Algorithm to remove redundant probes for each genome is described in Figure **??**.

3.2 Homogeneity filtering

Homogeneity criterion requires the melting temperature for every probe should be within some pre-defined range. This is important because probes in a good probe set need to hybridize with their intended target at about the same temperature. Homogeneity criteria also demands that the GC content of the probe should be within bounds.

Computation of melting temperature, hybridization temperate and GC content can be done very efficiently as described in [14]. We remove all probes in each Q_i that do not satisfy the homogeneity criteria.

3.3 Sensitivity filtering

Sensitivity filter eliminates probes that form secondary structures. We use a simplified secondary structure prediction algorithm to determine whether a probe can form secondary structures. In this filter, we want to eliminate probes which are able to fold back itself.

Computation of the sensitivity of the probe also can be done very efficiently as described in [14]. We remove all the probes in each Q_i that do not satisfy the sensitivity criteria.

3.4 Host genome probe filtering

We need to remove probes that can hybridize against the host genome. This will reduce the overall noise in the array. This can be achieved easily by blasting the probes against the human genome. We remove all probes in each Q_i with E scores greater than a specific threshold, when balsted against the human genome.

4 Probe Selection Using Information Content

Based on the previous steps, we have a set of good and non-redundant probes Q_i for every genome g_i in the list. Due to the limited size of the microarray, we cannot include all those probes into it. We suggest selecting a set of probes based on their information content. The next two sections describe the proposed procedures in details, following which we prove the optimality of our algorithms.

4.1 Probes for detecting known pathogens

Recall from Section 2.2.1 that for each known genome g_i we want to find a good set of probes P_i to be put on the chip. The previous steps have provided us with a list of potentially good probes, denoted as Q_i, for each genome g_i. To select the best sets of probes that can uniquely identify the individual genomes, we propose to rank the probes in each Q_i by their information content about the genome g_i. Information content of a probe p_j is defined as $I(G, p_j) = -\log_2(\frac{m}{|G|})$, where m is the number of distinct genomes (out of all genomes in G) to which probe p_j can hybridize to. P_i is then formed by selecting M_i probes with the highest information content.

More formally, let:

- G be the set of genomes and Q_i be the set of good probes for genome $g_i \in G$ as described earlier, and

- $count(G, p_j) = |\{g_k | g_k \in G \bigwedge p_j \in Q_k\}|$ be the number of genomes in G that probe p_j can hybridize to.

Then, for each genome g_i:

1. For each probe $p_j \in Q_i$, calculate its information capacity $I(G, p_j) = -\log_2\left(\frac{count(G,p_j)}{N}\right)$.

2. Rank the probes of Q_i in descending information capacity score and output the top M_i probes as the set P_i.

4.2 Assessing unknown pathogens

For unknown pathogens, we wish to discover the probable groups or classes of genomes that the unknown one might fall into (see Section 2.2.2). Here, on top of the sets genomes G and the associated good probes Q_i, an additional set S of disjoint subsets of G, as defined earlier, is also given. Given these, we can construct the set of good probes $Q_{S_i} = \{p_j | \exists g_k \in S_i : p_j \in Q_k\}$ for each $S_i \in S$. The challenge is to select probes that (1) hybridize with the conserved regions among the genomes in S_i and at the same (2) be able to uniquely distinguish group S_i from the rest.

To ensure that conserved probes are selected, the probes in Q_{S_i} are scored by the information content that they carry in distinguishing individual genomes in S_i, i.e. $\forall p_j \in Q_{S_i} : I(S_i, p_j) = -\log_2\left(\frac{count(S_i,p_j)}{|S_i|}\right)$, and only H lowest scoring probes are retained in Q_{S_i} for further ranking. The idea of selecting the probes with least information content is to select probes that are conserved within the genomes of S_i. By definition, conserved probes are unable to distinguish the individual genome of S_i, hence their information capacity, with regard to set S_i, are low. In our experiment, $H = 500$. Note also that this should be done only if the group contains more than one genome, i.e. $|S_i| > 1$.

The H probes retained in the set Q_{S_i} are further ranked by their ability to distinguish group S_i from the rest. The information capacity of probe $p_j \in Q_{S_i}$ is defined as $I(S, p_j) = -\log_2\left(\frac{m}{|S|}\right)$, where m is the number of distinct groups that probe p_j belongs to. For each group S_i, the M_{S_i} top scoring probes are selected and outputted as P_{S_i}.

In other words, let:

- G be the set of genome, Q_i be the set of good probes of genome g_i, $count(X, p_j)$ be the number of genomes in X that p_j can hybridize into, and S be the set of disjoint subsets of G as defined earlier,
- $Q_{S_i} = \{p_j | \exists g_k \in S_i : p_j \in Q_k\}$ be the set of good probes for S_i, and
- $countset(S, p_j) = |\{S_i | S_i \in S \bigwedge p_j \in Q_{S_i}\}|$ be the number of groups of genomes that p_j belongs to.

Then:

1. For each group S_i with $|S_i| > 1$, compute $I(S_i, p_j) = -\log_2\left(\frac{count(S_i,p_j)}{|S_i|}\right)$ for all $p_j \in Q_{S_i}$, rank the probes of Q_{S_i} based on decreasing $I(S_i, p_j)$, and retain only the H lowest scoring probes.
2. For all $S_i \in S$, calculate $I(S, p_j) = -\log_2\left(\frac{countset(S,p_j)}{|S|}\right)$ for each $p_j \in Q_{S_i}$ and output M_{S_i} probes with the highest information capacity as P_{S_i}.

Note that the previous algorithm is in fact a special case of this algorithm where $\forall S_i \in S : |S_i| = 1$.

4.3 The optimality of the selected probes

All probes in the set after redundancy removal, homogeneity, sensitivity, and host-genome filtering are independent. Note that the specificity criterion requires us to maximize the information content of each probe set P_i (or in a more general framework, P_{S_i}). Since the probes are independent, the total information capacity of a probe set P_i is $\sum_{p_j \in P_i} I(G, p_j)$. Our selection method, by taking probes with the largest information capacities, thus ensures that the optimal sets of probes are selected and maximizes the total information about the pathogen present.

5 Analysis of Pathogen Chip's Outcome

Having designed the microarray, we need to make inferences from the array data so as to identify the pathogen that was present in the sample. The analysis can be made by computing the total information present about all the set S_i of interest. We report the presence of group S_i if we obtain a certain fraction of information that we were expecting. Note that we can in fact decouple the analysis stage from the probe design. Although a given pathogen chip might not have been constructed using our proposed algorithm, this analysis would still be applicable nonetheless.

5.1 General framework

For the purpose of our analysis, let:

- $G = \{g_1, ..., g_N\}$ be the set of genomes,
- $P = \{p_1, ..., p_M\}$ be the set of probes that are present in the pathogen chip,
- $P_i \subseteq P$ be the set of probes which are present in genome g_i,
- $P_L \subseteq P$ be the set of lighted-up probes after the hybridization with a pathogen,
- $S = \{S_1, ..., S_K\}$ be the set of disjoint subsets of genomes, such that $S_i \subseteq G$, $G = \bigcup_i S_i$, and $\forall i, j : i \neq j, S_i \bigcap S_j = \emptyset$,
- $P_{S_i} = \bigcup_{g_j \in S_i} P_j$ be the set of probes associated with S_i, and
- $countset(S, p_j) = |\{S_i | S_i \in S \bigwedge p_j \in P_{S_i}\}|$ be the number of groups of genomes that p_j belongs to.

Since the probes are independent, we can sum the total information about each group S_i from all probes that are lighted up. The set of probes P_L which are lighted up, carry the information about the presence of each S_i. The more probes that belong to P_{S_i} are lighted up, the more likely that S_i is truly present. This can be roughly measured by calculating the total information contributed by the probes to each S_i, which is done by taking the sum of the information capacity of all the lighted up probes that belong to S_i. Ranking S_i based on the bare total information provided by the lighted up probes might be a little bit unfair towards short genomes as they have lesser number of potential probes and potentially have lesser total information. To solve this, we propose to normalize the total information score of each S_i by dividing it with the maximum score possible for the group S_i. Only groups having a normalized score greater than certain threshold T, which in our case $T = 0.7$, are considered as potential groups. They are further ranked by their raw total information score. The analysis can summarized as follows:

1. Calculate $\forall i, j : I(S, p_j) = -\log_2\left(\frac{countset(S,p_j)}{|S|}\right)$
2. For each S_i, compute $TotInf(S_i) = \sum_{p_j \in P_L \bigcap S_i} I(S, p_j)$ and $NormTotInf(S_i) = \frac{\sum_{p_j \in P_L \bigcap S_i} I(S,p_j)}{\sum_{p_j \in P_{Si}} I(S,p_j)}$
3. Exclude all S_i with $NormTotInf(S_i) \leq T$, output the remaining ones sorted first by $NormTotInf(S_i)$ in decreasing order, and breaking any ties by $TotInf(S_i)$.

5.2 Integrating prior probabilities

In some situations, we might be able to obtain the prior probability distribution $\Pr(S_i)$. Prior probality changes the information content of probes. of the event that probe p_j is lighted up due to the presence of S_i. We can refine the information content of probe p_j for group S_i as:

$$I(S, S_i, p_j) = -\log_2 \left(\frac{countset(S, p_j) \times Prob(S_i)}{|S|} \right)$$

and the total information score and the normalized total information can be calculated accordingly:

$$TotInf(S_i) = \sum_{p_j \in P_L \bigcap S_i} I(S, S_i, p_j)$$

$$NormTotInf(S_i) = \frac{\sum_{p_j \in P_L \bigcap S_i} I(S, S_i, p_j)}{\sum_{p_j \in P_{Si}} I(S, S_i, p_j)}$$

The rest of the analyses can be performed as before.

6 Simulation Results

We ran some simulation to check the efficacy of the chip. In the first step we included all the 478 fully sequenced viruses that can infect animals and gave as input to the probe design system. We took 30 probes from each genome that has maximal information carrying capacity as explained above. We also selected 30 probes for every genus, species and family known.

The next step was to test the ability of the chip to identify genomes already present in the set. We took genome sequences and randomly selected various percentage of the sequences and hybridized to the array. Information about the pathogen was obtained from the probes lighted up. As shown in Table 1, we were able to identify the pathogen present in all the experiments done if we had at least 70% of the genome of the pathogen in the sample. Even when the percentage of genome present was only 60%, we were able to identify the genus and family correctly.

We also tested the performance of pathogen chip for mutations in the pathogen genome. As shown in Table 2, the detection procedure is still quite robust for some reasonable big mutation rate.

We then tested the ability of the chip to identify novel pathogens. We designed the chip after removing the SARS genome from the list. To test this chip, we hybridized the SARS genome. We were not able to make a unique call for any genome. However as shown in Table 3, we were able to make a decision that the new virus belonged to the corona virus genus and to the coronaviridae family.

7 Discussion

Microarray technology promises to revolutionize the way pathogens are detected and characterized. The most crucial factor determining the effectiveness of this procedure is the probes selected. This determines the information we obtain about the pathogens present. More informative the probes are, more information we obtain from the microarray hybridization experiment.

We presented a new algorithm to select probes for pathogen detection. Our algorithm makes use of several smart filtering techniques to reduce the search space for probes. Then by using information capacity as a measure we are able to select the most optimal set of probes for the purpose of detection.

Further research includes making use of the prior knowledge about the kind of pathogen we expect. Our current approach can take this information as the prior probability of a pathogen being present in the sample. We are currently working on a validation strategy to include this scenario as well.

On a linux cluster consisting of 16 nodes each running at $2.6GHz$ the probe design algorithm took about a week to run. Simulating a microarray hybridization took about 2 days to run. It is quite important to further reduce the time complexity of the program while maintaining its accuracy. We are currently working on this.

	Amount of genome hybridized to the chip			
Input genome	**90%**	**80%**	**70%**	**60%**
NC_000856	NC_000856 Papillomavirus Papillomaviridae	NC_000856 Papillomavirus Papillomaviridae	NC_000856 Papillomavirus Papillomaviridae	Papillomavirus Papillomaviridae
NC_004004	NC_004004 Apthovirus Picornaviridae	NC_004004 Apthovirus Picornaviridae	NC_004004 Apthovirus Picornaviridae	Apthovirus Picornaviridae

Table 1. Pathogen identified by the algorithm from the microarray data.

	Probability of mutation rate per base pair			
Input genome	**0.001**	**0.01**	**0.05**	**0.1**
NC_000856	NC_000856	NC_000856	NC_000856	none

Table 2. Suggested pathogen from the array data when hybridized with mutated genome sequence

	Amount of SARS genome hybridized to the chip			
Input genome	**90%**	**80%**	**70%**	**60%**
Predicted groups	Coronavirus Coronaviridae	Coronavirus Coronaviridae	Coronavirus Coronaviridae	Coronavirus Coronaviridae

Table 3. SARS genome hybridized to a chip, without including SARS genome while designing the chip.

References

[1] Debouck C. and Goodfellow P.N. DNA microarrays in drug discovery and development. *Nat Genet*, 21(1):48–50, 1999.

[2] Epstein Charles B. and Butow Ronald A. Microarray technology - enhanced versatility, persistent challenge. *Biotechnology*, 11:3641, 2000.

[3] Gerhold D., Rushmore T. and Caskey C. T. DNA chips: promising toys have become powerful tools. In *Trends in biochemical sciences*, pages 168–173, 1999.

[4] James Borneman, Marek Chrobak, Gianluca Della Vedova, Andres Figueroa, Tao Jiang Probe selection algorithms with applications in the analysis of microbial communities. *Proceedings of ISMB*, 2001.

[5] David Wang, Joseph L. Derisi et al. Microarray-based detection and genotyping of viral pathogens. *PNAS* , pages 15687—15692, 2002.

[6] David Wang, Joseph L. Derisi et al. Viral Discovery and Sequence Recovery Using DNA Microarrays. *PLoS Biology* , Volume 1, Issue 2, page 257, 2003.

[7] Sam Rash and Dan Gusfield. String Barcoding: Uncovering optimal virus signatures. In *Proceedings of RECOMB 2002*, pages 254–261, April 2002

[8] Bailey W.F. and Mohanan A.S. Statistical effects and the evaluation of entropy differences in equilibrium processes. Symmetry corrections and entropy of mixing. In *J. Chem. Ed.*,55, pages 489–493, 1978.

[9] Raddatz G., Dehio M., Meyer T. F. and Dehio C. Primearray: genome-scale primer design for dna-microarray construction. *Bioinformatics*, 17:98–99, 2001.

[10] Rahmann S., Rapid large-scale oligonucleotide selection for microarrays. *Proc. of the Second Workshop on Algorithms in Bioinformatics (WABI)*, 302–311 (2002).

[11] Schliep A., Torney D.C. and Rahmann S., Group Testing With DNA Chips: Generating Designs and Decoding Experiments *IEEE Computer Society Bioinformatics Conference (CSB)*, 2003.

[12] Li F. and Stormo G. Selection of optimal DNA oligos for gene expression analysis. *Bioinformatics*, 17(11):1067–1076, 2001.

[13] Kaderali L. and Schliep A. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, 18(10):1340–1349, 2002.

[14] Wing-Kin Sung and Wah-Heng Lee. Fast and Accurate Probe Selection Algorithm for Large Genomes. *IEEE Computer Society Bioinformatics Conference (CSB)*, 65–74, 2003.

[15] Claude E. Shannon. A Mathematical Theory of Communication *Bell System Technical Journal*, July 1948

[16] SantaLucia J. J., Allawi H. T. and Seneviratne P. A. Improved Nearest-Neighbor Parameters for Predicting DNA Duplex Stability. *Biochemistry*, 35:3555–3562, 1996.