Group Testing With DNA Chips: Generating Designs and Decoding Experiments

Alexander Schliep[1], David C. Torney[2], Sven Rahmann[1,3]

[1] Department of Computational Molecular Biology
Max-Planck-Institute for Molecular Genetics
Ihnestraße 63–73, D-14195 Berlin, Germany

[2] Theoretical Biology and Biophysics
Los Alamos National Laboratory
Los Alamos, NM 87544, USA

[3] Department of Mathematics and Computer Science
Freie Universität Berlin
Arnimallee 2–6, D-14195 Berlin, Germany

E-mail: {Alexander.Schliep|Sven.Rahmann}@molgen.mpg.de

Abstract

DNA microarrays are a valuable tool for massively parallel DNA-DNA hybridization experiments. Currently, most applications rely on the existence of sequence-specific oligonucleotide probes. In large families of closely related target sequences, such as different virus subtypes, the high degree of similarity often makes it impossible to find a unique probe for every target. Fortunately, this is unnecessary.

We propose a microarray design methodology based on a group testing approach. While probes might bind to multiple targets simultaneously, a properly chosen probe set can still unambiguously distinguish the presence of one target set from the presence of a different target set. Our method is the first one that explicitly takes cross-hybridization and experimental errors into account while accommodating several targets.

The approach consists of three steps: (1) Pre-selection of probe candidates, (2) Generation of a suitable group testing design, and (3) Decoding of hybridization results to infer presence or absence of individual targets.

Our results show that this approach is very promising, even for challenging data sets and experimental error rates of up to 5%. On a data set of 28S rDNA sequences we were able to identify 660 sequences, a substantial improvement over a prior approach using unique probes which only identified 408 sequences.

1. Introduction

DNA microarrays are a widely used tool for performing large numbers of DNA-DNA hybridization experiments in parallel. We distinguish two principal kinds of applications:

1. Quantitative analysis of expression levels of individual genes, measured by quantifying the hybridization levels of gene-specific oligonucleotide probes. Prominent applications are the comparison of cell samples from different tissues and computational diagnostics.

2. Qualitative analysis of an unknown sample; most notably, establishing presence or absence of target sequences in a sample by observing appropriate hybridization reactions. Examples from biology, ecology, biotechnology and medicine are identification of micro-bacterial organisms, detection of contamination of biotechnological products, or identification of viral subtypes.

To measure the expression level in the former, quantitative setting as precisely as possible, *unique probes* (also called gene-specific probes or signature oligos) are de rigeur. A probe is called unique if, under specified experimental conditions such as temperature and salt concentration, it hybridizes to its intended target and (almost certainly) does not hybridize to any other target that might be expressed in the sample. The selection of unique probes is an interesting problem, and several methods of varying speed and accuracy have been proposed (e.g., [8, 9, 11, 13]).

In large families of closely related target sequences, the high degree of similarity makes it impossible to find a unique probe for every target — given the probe length and

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

melting temperature constraints. This issue is hard to resolve in a setting where unique probes are called for. Most authors suggest either clustering or leaving out target sequences for which no probe can be found, or increasing oligo length. The latter approach is probably the best when quantitative measurements are essential.

We focus not on quantification, but on robust presence or absence calls, such as in virus subtyping. In this case, unique probes are not a necessity, and we have many more liberties in designing the chip. We propose a *statistical group testing* approach which we will elaborate upon in the sequel.

Group Testing. Group testing is a general procedure applicable whenever a large population of individuals has to be subjected to the same test. A general introduction to the field is given in [7]. The idea is to group objects and test groups instead of individuals. A group tests positive when at least one individual within the group tests positive. Every experiment involving group testing requires a *design*, i.e., a definition of the groups (every individual can belong to many groups), and an corresponding *decoding* procedure to infer the status of individuals from the status of groups. When we select the groups a priori we have a *non-adaptive* group test. In contrast, in *adaptive* group testing the groups are chosen iteratively, using the results from previous tests to guide selection of groups for the next iteration. When the test results are exact, that is error-free, we speak about *combinatorial* group testing; in the presence of errors one has the harder *statistical* variant of the problem.

Group testing is most successful in general, whenever only few individuals are expected to test positive, because we can use large groups and hence need only a few tests to infer the status of every individual. For example, when it is known that there is at most a single positive individual among n, $\lceil \log_2 n \rceil$ tests suffice to determine its identity. When the proportion of positive individuals is large, nothing is to be gained in comparison to individual testing.

Successful applications are from medical diagnostics, including screening draftees for syphilis during World War II, the first recorded application [7], and from industrial quality assurance. In molecular biology, group testing has been applied to the problem of screening DNA clone libraries for sequence tagged sites to aid in the construction of physical maps [1, 5, 6].

Group Testing Issues for Microarrays. In the microarray setting we propose to use a statistical, non-adaptive group testing scheme. The target sequences we intend to identify correspond to individuals. Potential groups, which can be freely chosen in the general setting, are specified here by a probe which hybridizes to a set of target sequences. The goal is to devise a group testing design which covers each target with a certain number of probes and allows identification of several targets simultaneously while using a reasonably small total number of probes. We encounter several novelties that are not present in other group testing settings.

- *Constrained assignment of individuals to groups.* In contrast to a medical screening setting, we cannot arbitrarily assign individuals to groups. Groups (sets of target sequences) are always defined by an oligo that occurs in all sequences of the group. This restriction is the most important difference between DNA array group testing and "classical" group testing designs.

- *Cross-hybridization.* Even though we allow non-unique probes, the cross-hybridization problem does not disappear. Assume that a probe p occurs in all target sequences in a set T, and also approximately (but not exactly) matches another target $s \notin T$. The hybridization behavior of s with respect to p depends on many parameters and will vary from experiment to experiment. To keep error rates as low as possible, it is preferable to discard probes whose hybridization behavior is unclear.

- *Comparatively high error rates.* Even if we avoid potential cross-hybridization problems, false positive (a probe giving a signal when it should not) and false negative errors (a probing not showing a signal when it should) with rates of up to 5% must be anticipated.

- *Moving Targets.* In reality, target sequences are not static objects. They undergo mutations, recombine, or are altered in other ways. Covering a target with many probes adds robustness to the target identification; allowing these probes to be non-unique helps to keep the required probe number low.

These difficulties and their practical importance in real-world applications, such as virus subtyping, make the design of group test oligo arrays an interesting challenge. While previous work has addressed the use of non-unique probes and group testing (e.g., [3, 15]), none of it seems to take cross-hybridization and error tolerance explicitly into account, or it simplifies the problem unrealistically by assuming at most one target sequence (positive individual) to be present.

Our approach to deal with the above issues is as follows (cf. Figure 1). First, we pre-select suitable probe candidates, where the usual constraints for oligo design apply (see Section 2.1). From these candidates, we generate a group testing design, i.e., we pick a subset of probes that allows discrimination between as many (small-sized) target sets as possible. Note that in theory, adding a probe to a design never decreases the design's ability to separate two

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

target sets. Therefore, using *all* candidates guarantees the best possible separation properties of the design (the design quality is naturally limited by the properties of the best candidate probes). In practice, however, the number of candidate probes is very large, and several probes might hybridize to the same target sets. Therefore adding a probe does not necessarily contribute new information, and we would waste spots on the chip with uninformative oligos. Selecting a smaller design may allow use of a smaller chip and considerable reduction of cost. Furthermore, in practice, candidates are ranked according to quality (see the previous section), and we want to include lower-quality probes only when absolutely necessary in order to keep the noise level for the decoding procedure as low as possible. In short, while a small design is preferable, it would be misleading to minimize the number of probes as the only objective. Our heuristic solution is described in Section 2.2.

Our decoding procedure to infer the presence of target sequences uses a Bayesian framework. It is based on Monte Carlo Markov Chain sampling and explicitly allows false positive and false negative experimental errors (Section 2.3). We evaluate the method on a test set of 660 28S rDNA sequences in Section 3, and a concluding discussion can be found in Section 4.

2. Methods

We use the following notation. The m target sequences are denoted by t_i ($1 \leq i \leq m$), the initial n_0 probe candidates by p'_k ($1 \leq k \leq n_0$), and the final $n \leq n_0$ probes selected for the design by p_j ($1 \leq j \leq n$). Thus, every p_j is equal to some p'_k.

We define a target-candidate-incidence matrix H by $H_{ik} := 1$ if target t_i hybridizes to probe candidate p'_k, and $H_{ik} := 0$ otherwise. The design matrix D is a sub-matrix of H, where columns that correspond to candidates not included in the final design have been removed. So $D_{ij} = 1$ if target t_i hybridizes to the selected probe p_j.

The set of probes hybridizing to target t_i, i.e., the index set of nonzero entries in row i of the incidence matrix D (or H), is denoted by $P(i)$. Similarly, $T(j)$ denotes the set of target sequences probe p_j hybridizes to, or equivalently, the index set of nonzero entries in column j of D.

2.1. Computing Probe Candidates

As mentioned above, targets cannot be arbitrarily assigned to groups. Instead, a potential target group T only exists if there is a probe that binds to all — and exclusively those — sequences in T. Additionally, not every probe can be used to define a group because all candidate probes must obey the typical restrictions also encountered in unique oligo selection. For instance, all probes should

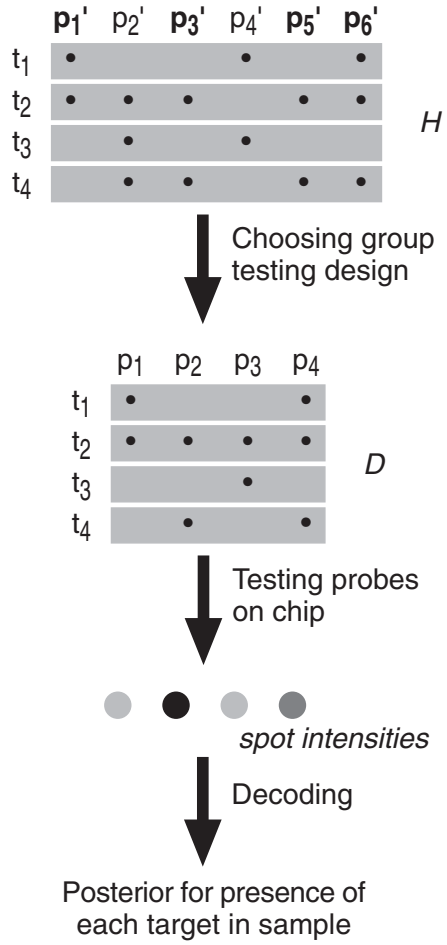

Figure 1. An overview of the method. In the target-and-probe hybridization matrix H (top) a dot in row i and column k indicates that probe candidate k binds to target sequence i. A group testing design D is a subset of columns of H (the candidate probes p'_k selected are displayed in bold face). The hybridization experiment allows us to measure spot intensities for each probe selected for D. The information in D about which targets hybridize to each probe is used to identify the targets present in the sample in the decoding step.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

ideally exhibit the same hybridization affinity, expressed as the Gibbs free energy ΔG of the oligo-target-duplex, at a given temperature and salt concentration. Furthermore, the probes should not be self-complementary and not cross-hybridize to targets outside their intended target set. The last issue is essential for robust designs.

Our candidate selection is based on an extended and modified version of the longest common factor method described in [13]. The *longest common factor* length $\mathrm{lcf}(s,t)$ of two strings s and t is the maximum length of a substring that occurs in both s and t. The length of a string o is written as $|o|$. For each candidate oligo o of gene g, and for each length $\ell \leq |o|$, we define the *longest common factor statistics* $\mathrm{lcfs}(o,\ell)$ as the number of genes $g' \neq g$ containing an oligo o' with $\mathrm{lcf}(o,o') = \ell$. When $\mathrm{lcfs}(o,\ell)$ is nonzero for some ℓ close to $|o|$ or exceeds the allowed maximal target set size for $\ell = |o|$, the oligo is at risk of unintended cross-hybridization and excluded from further consideration. The basic longest common factor approach is very fast, but it overlooks that a GC-binding is more stable than an AT-binding. Recently, the method was extended to not only compute the length difference but also to estimate the difference in Gibbs free energy between the perfect match probe-target duplex and potential secondary binding sites (approximate matches). This permits a more accurate estimation of cross-hybridization risk. Details are given in [14].

As a rule of thumb for 20-mers, a change of 1 bp in match length corresponds to a change of 1.1 kcal/mol in the duplex's free energy, but for individual oligos, these quantities vary considerably depending on sequence composition. A 20-mer oligo is considered as a suitable design candidate for the next step when it occurs as an exact match in at most 40 target sequences and the estimated Gibbs free energy difference δ to the best approximate match is higher than 5.5 kcal/mol at 45°C and 0.075 M NaCl.

We rank the list of candidates according to the value of δ (higher is better) and compute the target-candidate hybridization matrix H.

2.2. Finding a Good Design

This section describes a fast heuristic to find a good group testing design D, i.e., to select columns of the full target-candidate hybridization matrix. We want to be able to distinguish between most (ideally all) target sets whose size is not too large.

Definition 1 (d-separability of target sets). Let S be a set of target sequences. We say that a probe p *hybridizes to the set* S when p hybridizes to at least one target in S. By $P(S)$ we denote the set of all probes hybridizing to S, i.e., $P(S) := \bigcup_{t_i \in S} P(t_i)$.

Now let S and T be two different target sets. Probe p *separates* S and T if $p \in P(S)\Delta P(T)$, i.e., if p hybridizes to either S or T, but not to both (Δ denotes symmetric set difference).

The target sets S and T are d-separable if at least d probes separate them, i.e., if $|P(S)\Delta P(T)| \geq d$.

As an example, assume we have d *unique* probes for each target. Then two target sets S and T with $|S\Delta T| = c$ are $(c \cdot d)$-separable, because the signal of d different probes differs for each target in $S\Delta T$. Of course, we do not generally have unique oligos to choose from and are restricted to the available candidate probes.

A call to the following procedure $\textsc{Separate}(S,T,d)$ ensures d-separation of S and T, or produces a warning if the candidate set allows only d'-separation for some $d' < d$.

$\textsc{Separate}(S,T,d)$
Add oligos to the current partial design D to d-separate S and T
1. Let $C := P(S)\Delta P(T)$
2. Partition C into the disjoint union $C = C_D \cup C'$, where $C_D := C \cap D$, and C' contains the separating oligos not yet included in D
3. If $|C_D| \geq d$ **return** (nothing to do).
4. If $|C'| < (d - |C_D|)$ **warn** "Can only $(|C_D| + |C'|)$-separate S and T"
5. Add $d - |C_D|$ highest-quality probes from C' to D

Because of the many different considerations mentioned in the introduction, formulating the design problem as a simple optimization problem without sacrificing realism is difficult. Therefore we use the following common-sense approach to generate a reasonably good and small design.

1. We add probes until every target is covered by at least d oligos, i.e., every singleton target set $\{t_i\}$ is d-separated from the empty set $\{\}$, by calling $\textsc{Separate}(\{t_i\},\{\},d)$ for all $i = 1,\ldots,m$.

2. We ensure that all pairs of targets are separated by at least d oligos by calling $\textsc{Separate}(\{t_i\},\{t_{i'}\},d)$ for all $1 \leq i < i' \leq m$.

3. Since there are usually several hundred targets, it would take too much time to systematically ensure d-separation for all larger target sets up to a certain cardinality. Instead, we randomly pick a number N of additional pairs of target sets S and T and d-separate them by a calling $\textsc{Separate}(S,T,d)$. The size distribution of the sets we pick follows the distribution of the number of targets present in a typical sample (cf. the cardinality prior in Sec. 2.3). The parameter N can be chosen according to the time available to refine the design. As an example, this step took 5 minutes for 600 targets and 14000 candidates, using $N = 500000$.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

2.3. Decoding

Once we have performed the hybridization experiment, we are faced with the problem of inferring which targets were present in the sample using the results for the different probes. Following [10], we use a Bayesian approach for the decoding.

Formally, we are given D and the result vector $r = (r_1, \ldots, r_n)$ for the probes $p_1, \ldots, p_n$, where we assume $r_j \in \{0,1\}$.

We consider the posterior probability that a set T of targets constitutes all targets present in a sample, given the result vector r. Using Bayes' formula, we can write this probability as

$$\mathbb{P}[T|r] = \frac{\mathbb{P}[r|T] \cdot \mathbb{P}[T]}{\mathbb{P}[r]}.$$

A likelihood model. Assuming that all and only those targets from a set T are present in a sample, what is the probability of observing a result vector r? Intuitively, if probe p_j does not hybridize to any of the targets in T, the result r_j for that probe should be negative. Similarly, if another p_k hybridizes to one or several of the targets in T we expect $r_k = 1$. Given the typical error rates of DNA chip experiments, we have to assume that the probability of observing a specific result merely correlates to the number of targets from T a probe binds to. Another assumption we make is that the observed results are independent between probes. This allows us to write the likelihood as

$$\mathbb{P}[r|T] = \prod_{p_j} f\left(r_j, |T(j) \cap T|\right),$$

where the product is over all probes p_j and $|T(j) \cap T|$ denotes the number of targets the probe p_j hybridizes to which also are contained in the set T. Note that $f(0,0)$ is simply one minus the false positive error rate, denoted f_p, for a hybridization reaction in one spot and that the sum $\Sigma_{k \geq 1} f(0,k)$ is the corresponding false negative error rate, denoted f_n. We only distinguish between the presence of none respectively one or more targets a probe hybridizes to in the sample. Hence, we set $f(0,0) = 1.0 - f_p$, $f(0, \geq 1) = f_n$, $f(1,0) = f_p$ and $f(1, \geq 1) = 1.0 - f_n$. These guesses should be replaced with observed error rates, whenever possible.

Choosing A Prior. We formulate a Bayesian prior for the presence of a group of targets in the sample. Generally, one assigns a probability to every set T from $2^{\{1,\ldots,m\}}$, reflecting the prior belief that all targets in T and no others are found in the sample. We assume independence between targets and, furthermore, that there are only two main contributing factors: the prevalence of each particular target t_i in samples containing at least one target is denoted f_i, and the distribution describing the likelihood of finding a particular number of different targets in one sample. For example, HIV infections with more than three subtypes are very rare and — at least in the US — in a vast majority of infected individuals exactly one subtype is detectable [4]. The prior probability of observing k different targets in one sample will be denoted by c_k. Combining the two factors we can define a quantity proportional to the prior we want. Let T denote the set of targets, then

$$\mathbb{P}[T] \propto c_{|T|} \cdot \prod_{t_i \in T} f_i \prod_{t_i \notin T} (1 - f_i).$$

An uninformed prior on the prevalence frequencies, that is $f_1 = \cdots = f_n$, yields the the binomial distribution neglecting the c_k terms. In cases in which sufficient data is available, a more refined statistical model not limited by our assumptions should be formulated to obtain better decoding performance.

The Posteriors. The fact that $\mathbb{P}[r]$ is not readily available precludes us from computing the posterior in closed form. Moreover, what we are really interested in are marginals such as $\mathbb{P}[t \text{ present in sample}|r]$. These marginals can be expressed in terms of the posterior for a set by summing over all sets T which contain the specific target t. That is,

$$\mathbb{P}[t \text{ present in sample}|r] \propto \Sigma_{T:t \in T} \mathbb{P}[T|r].$$

Thus, an exact computation requires work exponential in the number of targets.

Markov Chain Monte Carlo (MCMC). One way to cope with this problem is to use a Monte Carlo approach. Sampling a sufficient number of sets T_k according to $\mathbb{P}[T|r]$ allows to estimate the marginal $\mathbb{P}[t \text{ present in sample}|r]$ as the relative frequency with which t is contained in the sets T_k. This requires sampling from $\mathbb{P}[T|r]$, for which Gibbs sampling suffices. Gibbs sampling allows construction of a Markov chain over the space of all sets T, that is $2^{\{1,\ldots,m\}}$, which has $\mathbb{P}[T|r]$ as its stationary distribution.

In the Gibbs sampler, the probability of changing from some T_k to $T_k \Delta \{t_i\}$ for some probe p equals

$$\frac{1}{1 + \mathbb{P}[T_k|r] / \mathbb{P}[T_k \Delta \{t_i\}|r]}.$$

Note that this suffices to obtain convergence in distribution to $\mathbb{P}[T_k|r]$ [2, pp. 353-354]. By use of Bayes' Theorem and the model distributions defined earlier, it follows that the fraction $\mathbb{P}[T_k|r] / \mathbb{P}[T_k \Delta \{t_i\}|r]$ equals

$$\frac{c_{|T_k|}(1 - f_i)}{c_{|T_k|+1} f_i} \cdot \prod_{p_j \in P(i)} \frac{f(r_j, |T(j) \cap T_k|)}{f(r_j, |T(j) \cap T_k| + 1)}$$

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

if $t_i \notin T_k$ and, for $t_i \in T_k$,

$$\frac{c_{|T_k|+1} f_i}{c_{|T_k|}(1-f_i)} \cdot \prod_{p_j \in P(i)} \frac{f(r_j, |T(j) \cap T_k|)}{f(r_j, |T(j) \cap T_k| - 1)}$$

If $f(\cdot,\cdot)$ is never 0 nor 1, the constructed Markov chain is aperiodic and irreducible, which guarantees convergence to the unique stationary distribution. Once stationarity is exhibited, the states T_k the chain visits can be used to compute the relative frequencies of the targets t_i, which are estimators for the marginals $\mathbb{P}[t_i \text{ present in sample}|r]$.

Implementation. We adapted the Markov Chain Pool decoder (MCPD) [16] to the problem at hand. As it is routine for MCMC methods, the user can select a number of warm-up steps and also the number of steps between states which are used for computation of the marginals. We checked the convergence of the MCMC method by performing 100 runs each for 100 sets of artificial results and computing the standard deviation of the marginal probabilities for a number of time points. The artificial results were sets of targets chosen according to the prior; each run was started from a randomly selected initial state. Inspection of the convergence data guided our choice of using 5,000 warm-up and 50,000 total steps. The rate of convergence of the marginals seemed not to depend on the number of steps in-between samples used (not shown), hence we simply used every state visited as a sample. On an AMD Athlon XP 2100 Linux machine, a decoding run needs about 15 seconds CPU time.

Additionally, we are able to use confirmed external information, say from an alternative testing method or infection history, about presence or absence of targets in a sample. We can compute marginals for the presence of individual targets *conditioned on this external information* — i.e., conditioned on specified targets present or absent — as follows: targets present (absent) are always included (exluded) from the state of the MCMC. Furthermore, this can be used to re-run the decoding if confirmatory tests on the targets with highest marginal probabilities are performed. Even without further tests, re-running the decoding with a selection of the targets with highest marginal probabilities gives an indicator for the sufficiency of those targets selected. If the selection is the one most likely to be correct, marginals of all the remaining targets should be negligible.

3. Evaluation on 28S rDNA Sequences

We used a set of 1230 28S rDNA sequences from different organisms present in the Meiobenthos [12]. The set contains redundancies and many close homologs. To reduce the level of redundancy we used the blastclust software from NCBI to cluster sequences in the date set which share at least 99% sequence identity over at least 99% of their length. Given the average sequence length of about 676 nucleotides this corresponds to about 7 mismatches between clustered sequences on average. The 149 clusters containing two or more sequences represent about 56% of all sequences. For each of those clusters we picked an arbitrary (first id in the blastclust output) representative. This procedure resulted in a test set consisting of 679 sequences.

For 19 of the 679 sequences, we were unable to find any suitable oligos under the prescribed experimental conditions (Temperature of 45°C, oligo length between 19 and 22 nt, ΔG range of the perfect duplex between -22700 and -21500 cal/mol, 0.075 M NaCl). The remaining 660 sequences resulted in 13112 candidate probes. A final design containing 2246 probes was computed using the greedy algorithm proposed. Each target was covered by at least five probes (five-fold coverage) if at all possible. On average each probe hybridizes to 2.55 targets.

To quantify the performance of the group testing scheme in the presence of errors, we carried out the following experiment. A set of target sequences was chosen randomly so that the number of sequences chosen was distributed according to the prior on the cardinalities c_k. We used an uninformed prior on the frequencies f_i. The selected sequences were assumed to be true positives. Given the design D we computed the set of positive probes and removed each probe from that set according to the assumed false negative rate $f_p = 0.01$ respectively $f_p = 0.05$. Probes were selected i.i.d as false positives with probability $f_n = 0.01$ respectively $f_n = 0.05$. The results were used as input for the decoder and the rank of the true positives was computed and averaged over the 592 respectively 667 repetitions of the experiment; see Table 1. Ideally, the k true positives should appear as the first k targets in the output.

4. Discussion

We have presented a DNA microarray design methodology based on non-unique oligos and group testing. Our method is the first that explicitly considers cross-hybridization issues and experimental errors within this framework. The results are very promising. Even in the case of 5% false positive and false negative errors for the hybridizations, which for smaller numbers of targets present in a sample implies more false positive than true positive spots on a microarray, 92% of the targets present can be identified if we allow for one false positive target.

The high robustness of the method is even more remarkable given the fact that in prior work only 408 sequences — compared to 660 sequences for our group testing scheme — could be identified using unique oligos [9]. It remains unclear whether multiple coverage (i.e., several unique probes for each target) could be obtained at all in the prior setting.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

#positives	n	top 1	top 2	top 3	top 4	top 5	top 10
1	285	0.905	0.965	0.989	0.993	0.993	0.996
2	190		0.926	0.976	0.995	1.000	1.000
3	80			0.921	0.958	0.983	1.000
4	34				0.846	0.919	0.971
5	3					0.800	1.000
1	309	0.909	0.974	0.990	0.997	1.000	1.000
2	202		0.884	0.946	0.973	0.993	1.000
3	108			0.873	0.951	0.969	1.000
4	36				0.910	0.951	1.000
5	12					0.867	0.967

Table 1. The fraction of true positives among the top k in the output of the decoder, ranked according to their probability. We assumed that $f_p = f_n = 0.01$ (top) and $f_p = f_n = 0.05$ (bottom). A total of 592 (top) respectively 667 (bottom) experiments were performed. The proportion of samples of each cardinality follows the prior c_k.

The framework presented can be easily extended to apply in other settings. We have demonstrated for expository reasons the decoding of hybridization experiments with a binary classification of the hybridization level. In [10] details for an arbitrary but discrete set of experimental outcomes, which translates directly to the situation described here, are given.

A quantitative analysis to provide information about the respective abundance of each target would require a more substantial modification. The aspect which mainly needs further investigation are the dynamics of hybridization reactions. A clarification of their behavior if multiple parallel or competing reactions occur is required before construction of a valid statistical model. A statistical model and a MCMC approach using essentially mixtures of probability functions is supported in the framework described.

Evolutionary information about targets, particularly when the target sequences are given as members of the families they belong to, will likely improve performance of the experimental setting. Probes are classified and selected at two hierarchy levels, family and members of each family. This requires a preprocessing step identifying probes in the hybridization matrix H which bind to (almost) all family members. The decoding should be adjusted to account for the hierarchical information, particularly conditioning the probability of the presence of individual family members upon the results of the corresponding family probes.

Detection of recombination events is another problem of relevance. For viruses like *H. influenzae*, where recombination events are frequent, a large number of probes as well as their location on the target sequence could be used for their characterization.

5. Acknowledgments

We thank Catherine Macken and Carla Kuicken for helpful discussions with regard to virus subtyping, as well as the participants of a recent Dagstuhl seminar on Bioinformatics. Special thanks to Martin Vingron for discussions and support, and also to Diethard Tautz and Melanie Markmann for their dataset and helpful discussions.

References

[1] E. Barillot, B. Lacroix, and D. Cohen. Theoretical analysis of library screening using an *n*-dimensional pooling strategy. pages 6241–6247, 1991.

[2] J. Bernado and A. Smith. *Bayesian Theory*. John Wiley, 1994.

[3] J. Borneman, M. Chrobak, G. D. Vedova, A. Figueroa, and T. Jiang. Probe selection algorithms with applications in the analysis of microbial communities. *Bioinformatics*, 17 Suppl 1:S39–48, 2001.

[4] S. K. Brodine, J. R. Mascola, P. J. Weiss, S. I. Ito, K. R. Porter, A. W. Artenstein, F. C. Garland, F. E. McCutchan, and D. S. Burke. Detection of diverse hiv-1 genetic subtypes in the usa. *Lancet*, 346(8984):1198–1199, Nov 1995.

[5] W. J. Bruno, D. J. Balding, E. H. Knill, D. Bruce, C. Whittaker, N. Doggett, R. Stallings, and D. C. Torney. Design of efficient pooling experiments. 26:21–30, 1995.

[6] N. A. Doggett, L. A. Goodwin, J. G. Tesmer, L. J. Meincke, D. C. Bruce, L. M. Clark, M. R. Altherr, A. A. Ford, H. C. Chi, B. L. Marrone, and a. l. et. An integrated physical map of human chromosome 16. *Nature*, 377(6547 Suppl):335–65, Sep 1995.

[7] D. Z. Du and F. K. Hwang. *Combinatorial Group Testing and Applications*. World Scientific Publishing, 1993.

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

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 IEEE

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

[10] E. Knill, A. Schliep, and D. C. Torney. Interpretation of pooling experiments using the markov chain monte carlo method. *J Comput Biol*, 3(3):395–406, Fall 1996.

[11] F. Li and G. Stormo. Selecting optimum dna oligos for microarrays. In *IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE)*, Key Bridge Marriott, Arlington, USA, Nov. 8-10 2000.

[12] M. Markmann. *Entwicklung und Anwendung einer 28S rDNA-Sequenzdatenbank zur Aufschlüsselung der Artenvielfalt limnischer Meiobenthosfauna im Hinblick auf den Einsatz moderner Chiptechnologie*. PhD thesis, University of Munich, 2000.

[13] S. Rahmann. Rapid large-scale oligonucleotide selection for microarrays. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB)*, pages 54–63. IEEE, 2002.

[14] S. Rahmann. Fast and sensitive probe selection for dna chips using jumps in matching statistics. In *Proceedings of the Second IEEE Computer Society Bioinformatics Conference (CSB)*. IEEE, 2003. To appear.

[15] S. Rash and D. Gusfield. String barcoding: Uncovering optimal virus signatures. In *Proceedings of RECOMB 2002*, pages 254–261, April 2002.

[16] A. Schliep. The markov chain pooling decoder. Los Alamos National Laboratories, http://algorithmics.molgen.mpg.de/MCPD/, 1998.

Proceedings of the Computational Systems Bioinformatics (CSB'03)
0-7695-2000-6/03 $17.00 © 2003 **IEEE**