2010 IEEE International Conference on Bioinformatics and Bioengineering

# Methods for designing reliable probe arrays

Michele Lombardi, Luca Benini
DEIS, University of Bologna
Viale Risorgimento 2, 40136 Bologna, IT
{michele.lombardi2,luca.benini}@unibo.it

Abhishek Garg, Giovanni De Micheli
École Polytechnique Fédérale de Lausanne
Batiment INF, Station 14, 1015 Lausanne, CH
{abhishek.garg,Giovanni.demicheli}@epfl.ch

***Abstract*—Recent advances in biosensing technologies have led to applications of biosensor probe arrays for rapid identification of biological agents such as drugs, gene expressions, proteins, cholesterol and fats in an input sample. However, monitoring the simultaneous presence of multiple agents in a sample is still a challenging task. Multiple agents may often attach to the same probes, leading to low specificity. By using microarrays as a specific example, we introduce two methods based on conditional deduction and non-unique probes to detect multiple targets. We introduce three quality metrics, namely: effectiveness, cost and reliability to evaluate different designs of microarrays and propose two ILP/Pseudo-Boolean models for optimizing on these metrics. By applying on various synthetic and real datasets, we demonstrate the importance of these quality metrics in designing microarrays for multiple target detections.**

***Microarrays; Probe based sensors; Optimization; ILP; SAT***

## I. INTRODUCTION

We target a design problem of probe based sensors for the detection of biological compounds. Without loss of generality, we take DNA microarrays as a practical example [3]. The surface of a microarray is instrumented with a number probes (single stranded DNA/mRNA oligonucleotides). When an experiment is performed on the array, probes on the surface of the device attach (or hybridize) to complementary gene sequences (often referred to as targets). The resulting hybridization reactions can be quantified due to the presence of chemical labels marking the probes, or by exploiting electro-chemical properties of the reaction (label free techniques [1]). Other type of probe based sensors differ in the nature of the probes and targets, but retain the same principle. With the integration of sensing and readout technologies on a single substrate, biosensor arrays will find more and more applications as devices for: i) medical diagnosis, ii) real-time biological experiments and iii) environmental monitoring.

Usually, probes specific to single genes are realized on the array. In some applications (e.g., monitoring water pollution by closely-related viruses), it is difficult to find good-quality unique probes for each target. In the last few years various methods (e.g. [2], [5], [6]) have been proposed that exploit non-unique probes. These methods focus on the design of arrays with minimum number of probes. All these approaches require that no more than one biological agent can be simultaneously present. However, this assumption can be very limiting in several contexts. We are interested in applications where an input sample may contain any number of targets. Targets that cannot be distinguished with the array have to be tested using alternate more expensive methods, hence maximizing the number of distinguishable targets is of interest as well. We introduce in this paper conditional detection as an alternative way to exploit non-unique probes. We demonstrate, by applying our methods on both synthetic and real datasets, how the approach can improve the microarray design. We propose to evaluate the quality of a microarray according to three metrics: (1) the cardinality of the set of targets it is able to detect, referred to as *effectiveness*, (2) the *cost* of the device and (3) the *reliability* of the detection. The reliability of an array depends on the *quality* of selected probes, which is a measure of several physical parameters, e.g. sequence identity, binding free energy and melting temperature (see [7]).

$$
H = \begin{array}{c} \\ \\ \\ \\ \\ \end{array}
\begin{array}{cccccc|c}
p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & \\
1 & 0 & 0 & 0 & 0 & 0 & t_0 \\
0 & 0 & 1 & 0 & 1 & 0 & t_1 \\
0 & 1 & 1 & 1 & 0 & 0 & t_2 \\
0 & 0 & 1 & 0 & 1 & 1 & t_3
\end{array}
$$

## II. THE DESIGN PROBLEM WITH REMOVAL ASSUMPTION

Similarly to other works [4] we assume target-to-probe bindings to be stored in an $n \times m$ matrix $H$ (see figure) such that probe $p_j$ hybridizes to target $t_i$ iff $H_{ij} = 1$. The problem we consider consists in selecting a set of targets $T^*$ and a set of probes $P^*$ (out of the full sets $T$ and $P$) such that, for each target $t_i \in T^*$ there are at least $d$ probes in $P^*$ uniquely binding to $t_i$. Additionally, we assume the user has the ability to remove targets from the sample so as to *make* some probes unique to other targets; those latter may thus become *conditionally distinguishable*. Let $P_U(t_i)$, $P_N(t_i)$ respectively be the set of unique/non-unique probes for $t_i$. Let $R_i(p_j)$ be the set of targets to which $p_j$ binds, except for $t_i$. Then we can formulate the following ILP/Pseudo-Boolean model:

$$
\begin{array}{lll}
\max z = & F(x,w,y,u) & \\
 & d \cdot x_i \le \sum_{p_j \in P_U(t_i)} y_j + \sum_{p_j \in P_N(t_i)} u_{ij} & \forall t_i \in T \\
 & u_{ij} \le y_j & \forall t_i \in T, \forall p_j \in P_N(t_i) \\
 & u_{ij} \le w_k & \forall t_i \in T, \forall p_j \in P_N(t_i), \forall t_k \in R_i(p_j) \\
 & x_i + w_i \le 1 & \forall t_i \in T
\end{array}
$$

with $x_i, w_k, y_{j,} u_{ij} \in \{0,1\}$ ; in particular, $x_i = 1$ if $t_i$ is selected, $w_k = 1$ if $t_k$ is removed, $y_j = 1$ if $p_j$ is selected and $u_{ij} = 1$ if

978-0-7695-4083-2/10 $26.00 © 2010 IEEE
DOI 10.1109/BIBE.2010.66

$p_j$ is made unique for $t_i$ by removing targets. Effectiveness (eft) and cost (cst) are among the possible quality metrics:

$$eft = F(x,w,y,u) = \sum_{t_i \in T} x_i + \alpha \cdot \sum_{t_k \in T} w_k$$
$$cst = F(x,w,y,u) = \sum_{p_j \in P} c_j y_j$$

where $c_i$ is the cost of $p_i$, and $\alpha < 0$ to take into account a cost to remove targets. The reliability metrics can be modeled with:

$$\begin{aligned} rel = \; & F(x,w,y,u) = z \\ with: \; & z \leq \sum_{p_j \in P_U(t_i)} q_j \cdot y_j + \sum_{p_j \in P_N(t_i)} q_j \cdot u_{ij} + M(1 - x_i - w_i) \quad \forall t_i \in T \end{aligned}$$

where $q_j$ is the quality of $p_j$. The reliability is the minimum overall quality of probes used to detect targets $t_i$. Note since those probes are either *unique* or made *unique*, the reliability assess robustness vs both noisy matches and noisy mismatches.

## III. The design problem without removal assumption

In some biological settings, removing targets may not be feasible. In this case, we can still say a target $t_i$ is conditionally distinguishable (without the removal assumption) if it is distinguishable, provided we know a certain set of targets $T'$ is not in the sample; the set $T'$ must be distinguishable. Therefore we get the following ILP/Pseudo-Boolean problem:

$$\begin{aligned} \max z = \; & F(x,w,y,u) & \\ & d \cdot x_i \leq \sum_{p_j \in P_U(t_i)} y_j & \forall t_i \in T \\ & d \cdot w_i \leq \sum_{p_j \in P_U(t_i)} y_j + \sum_{p_j \in P_N(t_i)} u_{ij} & \forall t_i \in T \\ & u_{ij} \leq y_j & \forall t_i \in T, \forall p_j \in P_N(t_i) \\ & d \cdot u_{ij} \leq \sum_{\substack{p_r binds\ to\ t_k \\ p_r binds\ only\ to\ R_i(p_j)}} y_r & \forall t_i \in T, \forall p_j \in P_N(t_i), \forall t_k \in R_i(p_j) \\ & t_i + w_i \leq 1 & \forall t_i \in T \end{aligned}$$

where $w_i = 1$ if $t_i$ is conditionally distinguished and $u_{ij} = 1$ if $p_j$ is unique for $t_i$, provided the set of targets $R_i(p_j)$ is not in the sample.

## IV. Experimental results

We solve both problems with the state of the art ILP solver IBM-ILOG Cplex 11 and the Pseudo-Boolean solver SAT4J on a Core Duo machine (1.8GHz). In particular, the SAT solver is used to provide Cplex with a high quality warm start. We tested the approach on the benchmark set given in [5], consisting of a real world (M) and 10 synthetic datasets (A1-5,B1-5); the input data included binding relations and probe qualities. Table 1 and 2 report the results when the removal assumption is respectively considered/non considered. We solve several problem variants with different objective function (**F**) and constraints on some quality metrics (**C**). The tables report the percentage of identified targets (**tgts**) or selected probes (**prbs**), the effectiveness and the reliability improvement (**impr**) over a basic case where only unique probes are used. The number of targets and probes is reported under the instance names.

| | | | A1 | A2 | A3 | A4 | A5 | B1 | B2 | B3 | B4 | B5 | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | targets | 256 | 256 | 256 | 256 | 256 | 400 | 400 | 400 | 400 | 400 | 679 |
| **F** | **C** | probes | 2786 | 2821 | 2871 | 2954 | 2968 | 6292 | 6283 | 6311 | 6223 | 6285 | 15,139 |
| eft | | tgts. | 58% | 56% | 63% | 63% | 59% | 72% | 73% | 71% | 71% | 72% | 55% |
| | | impr. | 17% | 13% | 17% | 10% | 12% | 15% | 20% | 16% | 14% | 17% | 9% |
| | | cnd. tgt. | 26% | 20% | 20% | 19% | 25% | 29% | 29% | 26% | 28% | 26% | 14% |
| rel | eft | impr. | 10% | 15% | 6% | 2% | 7% | 41% | 44% | 33% | 40% | 45% | 1928% |
| | | cnd. tgt. | 55% | 59% | 47% | 36% | 47% | 59% | 58% | 54% | 55% | 49% | 41% |
| cst | eft | *prbs.* | 23% | 21% | 23% | 23% | 22% | 18% | 19% | 18% | 18% | 19% | 11% |
| | | cnd. tgt. | 59% | 59% | 51% | 47% | 52% | 2% | 5% | 6% | 4% | 59% | 0% |
| cst | eft,rel | *prbs.* | 22% | 20% | 23% | 22% | 21% | 21% | 21% | –% | 21% | 22% | –% |
| | | cnd. tgt. | 45% | 0% | 0% | 0% | 0% | 64% | 61% | –% | 58% | 57% | –% |

**Table 1: results with removal assumption**

| | | | A1 | A2 | A3 | A4 | A5 | B1 | B2 | B3 | B4 | B5 | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | targets | 256 | 256 | 256 | 256 | 256 | 400 | 400 | 400 | 400 | 400 | 679 |
| **F** | **C** | probes | 2786 | 2821 | 2871 | 2954 | 2968 | 6292 | 6283 | 6311 | 6223 | 6285 | 15,139 |
| eft | | tgts. | 77% | 76% | 81% | 83% | 79% | 95% | 95% | 94% | 96% | 96% | 63% |
| | | impr. | 55% | 54% | 50% | 45% | 49% | 52% | 57% | 53% | 53% | 55% | 24% |
| | | *cnd. tgt.* | 36% | 35% | 33% | 31% | 33% | 34% | 36% | 35% | 35% | 35% | 20% |
| rel | eft | impr. | 22% | 23% | 22% | 10% | 17% | 62% | 69% | 52% | 56% | 54% | 3458% |
| | | *cnd. tgt.* | 28% | 26% | 27% | 23% | 25% | 27% | 28% | 25% | 27% | 25% | 17% |
| cst | eft | *prbs.* | 28% | 28% | 29% | 29% | 27% | 30% | 29% | 28% | 28% | 27% | 12% |
| | | *cnd. tgt.* | 26% | 26% | 25% | 20% | 20% | 27% | 28% | 25% | 28% | 29% | 0% |
| cst | eft,rel | *prbs.* | 28% | 28% | 29% | 29% | 27% | 30% | 29% | 28% | 28% | 27% | 13% |
| | | *cnd. tgt.* | 22% | 25% | 26% | 23% | 27% | 30% | 33% | 25% | 29% | 33% | 17% |

**Table 2: results without removal assumption**

It is interesting to remark that conditional detection yields a good improvement in the number of distinguishable targets and of selected probes. In particular the reliability gain can be enormous (see the real world M instance) if probes differ significantly in terms of quality. Further details will be presented in a follow-up paper.

## V. Conclusions

In this paper, we have proposed methods to design high quality DNA microarrays. We introduced conditional deductions as a way to exploit the advantages of non-unique probes. While previous methods exploiting non-unique probes assume that there are no more than certain number of targets in the input sample, our models make no such assumption. An empirical evaluation shows how our approach leads to more effective and more reliable devices, compared to a basic method using unique probes only. Finally this work points out the importance of explicitly taking into account probe quality and device reliability in microarray design, in particular when one is interested in minimizing the number of probes on an array. Even though the methods proposed in this paper are applied on DNA microarrays, they are applicable to all types of probe-sensing based biosensor arrays.

[1] L. Benini et al., "Electronic Detection of DNA Hybridization: Toward CMOS Microarrays", IEEE Design & Test of Computers, vol. 24, no. 1, pp. 38–48, 2007.

[2] C. Meneses et al., "A new approach to the non-unique probe selection problem," Annals of Biomedical Engineering, vol. 35, no. 4, pp. 651–658, 2007.

[3] S. Tomiuk and K. Hofmann, "Microarray probe selection strategies", Briefings in bioinformatics, vol. 2, no. 4, p. 329, 2001.

[4] S. Peng et al., "Probe Selection with Fault Tolerance", in Proc. of ICIC 2008. Springer, p. 219.

[5] G. Klau et al., "Optimal robust non-unique probe selection using integer linear programming", BIOINFORMATICS, vol. 20, pp. 186–193, 2004.

[6] P. Deng et al., "Efficient non-unique probes selection algorithms for DNA microarray", BMC genomics, vol. 9, no. Suppl 1, p. S22, 2008