Non-Unique Probe Selection by Matrix Condition Optimization

Sven Rahmann,[1] **Tobias Müller,**[2] **Martin Vingron**[3]

Keywords: microarray, DNA chip, design, probe selection, matrix, condition, optimization

1 Introduction: Non-Unique Probe Selection

We are interested in selecting oligonucleotide probes for DNA arrays [1]. In large transcript families, such as alternative splice variants of a gene, or in a large family of closely homologous genes (e.g., human heat shock proteins), it is often impossible to find enough unique 25-mer probes that can be taken as a signature for a specific variant. Therefore we consider *non-unique probes* [2].

For this study, we assume that we have many potential probe candidates and the task is to select an appropriate subset of them for use on the chip. We assume that we know (an approximation to) the probe-transcript *affinity matrix* A that relates transcript expression level to observed signal. We have $y = A \cdot x + c$, where $y \in \mathbb{R}^m$ are the observed probe signals, $x \in \mathbb{R}^n$ contains the transcript expression values, $A \in \mathbb{R}^{m \times n}$ contains the affinity coefficients between probes and transcripts, and c models additional noise or unspecific hybridization. A probe i that matches a transcript j leads to a high affinity value $A_{ij} \approx 0.1$ to 1, say. The target set of probe i is denoted by $T(i)$. A probe i that is unrelated to transcript j leads to a low affinity value of less than 10^{-4}, say.

From the $m \gg n$ probe candidates whose affinity values form the m rows of the affinity matrix A, we would like to select at most $\mu \leq m$ rows. We write H for the *hybridization matrix* defined by $H_{ij} := 1$ if $j \in T(i)$, and $H_{ij} := 0$ otherwise. We denote the index set of the chosen rows by D for *design*. We have $D \subset \{1, 2, \ldots, m\}$ and desire $|D| \leq \mu$. Let A^D and H^D denote the matrices obtained from A resp. H by removing all rows whose index is not in D. The requirements on D are that the equation $y = A^D \cdot x$ must be stably and robustly solvable for the n expression levels x, given the $|D|$ probe signals y.

2 Condition Optimization

Let A be an $m \times n$ matrix of full rank $n \leq m$, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ be the singular values of A. Then the condition of A is defined as $\mathrm{cond}(A) := \sigma_1/\sigma_n$. If A does not have full rank n, then $\sigma_n = 0$, and we set $\mathrm{cond}(A) := \infty$. The condition measures how changes in the measurement y influence the solution x of the minimization problem $\|y - A \cdot x\| \to \min$: We have $\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A) \cdot \frac{\|\Delta y\|}{\|y_A\|}$, where y_A is the projection of y on the range of A.

We assume that $\mathrm{cond}(A) < \infty$, i.e., that the affinity matrix that consists of all candidates has full rank n, and that the associated hybridization matrix H satisfies the minimum and average *coverage constraints* $\min_j \sum_i H_{ij} \geq \mathcal{M}$ and $\sum_{i,j} H_{ij} \geq n\mathcal{A}$. We let

[1] Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Ihnestr. 73, D-14195 Berlin, Germany; and Dept. of Mathematics and Computer Science, Free University of Berlin, Germany. E-mail: `Sven.Rahmann@molgen.mpg.de`

[2] Department of Bioinformatics, Biozentrum, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany. E-mail: `Tobias.Mueller@biozentrum.uni-wuerzburg.de`

[3] Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Ihnestr. 73, D-14195 Berlin, Germany. E-mail: `Martin.Vingron@molgen.mpg.de`

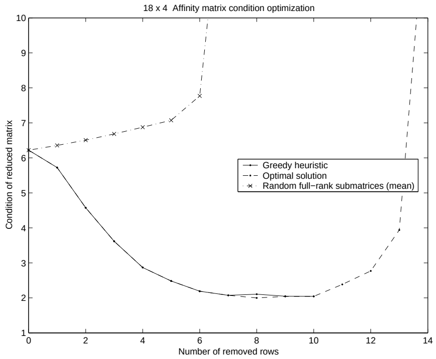
$\mathcal{D} := \{D \subset \{1, 2, \dots, m\} : |D| \le \mu,\ \mathrm{cond}(A^D) < \infty,\ \min_j \sum_i H^D_{ij} \ge \mathcal{M},\ \sum_{i,j} H^D_{ij} \ge n\mathcal{A}\}$ denote the set of admissible designs under the side conditions. It is assumed that $\mathcal{D}$ is not empty; otherwise we have to increase μ or decrease $\mathcal{M}$ or $\mathcal{A}$. The combinatorial problem is to minimize $\mathrm{cond}(A^D)$ among all $D \in \mathcal{D}$.

As far as we are aware, the *condition optimization problem* has not been posed before in the mathematical literature. It appears to be a difficult problem because the singular values of two matrices A^D and A^{D-i}, where $D - i := D \setminus \{i\}$, are not related in an obvious way, and also because the landscape of admissible designs potentially has many local minima. In spite of these difficulties, we propose a greedy heuristic to obtain a good admissible design.

```
Greedy Condition-based Design
Input: An $m \times n$ affinity matrix $A$ and hybridization matrix $H$
 1.  $D \leftarrow \{1, 2, \dots, m\}$
 2.  $B \leftarrow \emptyset$,  $C \leftarrow +\infty$
 3.  while ($|D| > n$)
 4.      $c \leftarrow +\infty$,  $i^* \leftarrow 0$
 5.      for each $i \in D$
 6.          if ($\min_j \sum_{i'} H^{D-i}_{i'j} \ge \mathcal{M}$) and ($\sum_{i',j} H^{D-i}_{i'j} \ge n\mathcal{A}$)
                 and ($\mathrm{cond}(A^{D-i}) < c$) then $c \leftarrow \mathrm{cond}(A^{D-i})$, $i^* \leftarrow i$
 7.      if $i^* = 0$ then break
 8.      $D \leftarrow D - i^*$
 9.      if ($|D| \le \mu$) and ($c < C$) then $B \leftarrow D$, $C \leftarrow c$
10.  if ($B = \emptyset$) then $B \leftarrow D$, $C \leftarrow c$
Output: Design $B$ with condition $C = \mathrm{cond}(A^B)$
```

The procedure starts with a full design and iteratively removes a single row to locally minimize the condition while still satisfying the coverage constraints(lines 5–6). If the resulting design is admissible it is compared against the current best admissible design B (line 9). This is repeated until the design size equals the number of targets (line 3) or no smaller design satisfying the coverage constraints can be found (line 7).

We evaluated the greedy heuristic against the optimal selection for small artificial matrices with 18 probe candidates and 4 targets. It was attempted to reduce the number of probes as far as possible with a minimum and average coverage requirement of 3 probes per target. Although the greedy heuristic does not always find the optimal solution, its performance is reasonably close to the optimal design (found by exhaustive search) and much better than choosing random subsets. The typical behavior is shown to the right: Removing a few probes improves the condition; removing too many will eventually worsen the condition again.



References

[1] S. Rahmann. Fast large scale oligonucleotide selection using the longest common factor approach. *Journal of Bioinformatics and Computational Biology*, 1(2):343–361, 2003.

[2] A. Schliep, D. C. Torney, and S. Rahmann. Group testing with DNA chips: Generating designs and decoding experiments. In *Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)*, pages 84–93. IEEE, 2003.