Probe Selection Algorithms with Applications in the Analysis of Microbial Communities

James Borneman\textsuperscript{1} Marek Chrobak\textsuperscript{2} Gianluca Della Vedova\textsuperscript{3} Andres Figueroa\textsuperscript{2} Tao Jiang\textsuperscript{2}

\textsuperscript{1} Dept. Plant Pathology, Univ. of California at Riverside
\textsuperscript{2} Dept. Computer Science, Univ. of California at Riverside
\textsuperscript{3} Dept. Statistics, Univ. Milano-Bicocca
Talk Outline

Introduction to the *in vitro* problem
Talk Outline

Introduction to the *in vitro* problem

Introduction to the *in silico* problem
Talk Outline

Introduction to the *in vitro* problem

Introduction to the *in silico* problem

The Lagrangian Relaxation-based algorithm

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Talk Outline

Introduction to the *in vitro* problem

Introduction to the *in silico* problem

The Lagrangian Relaxation-based algorithm

The Simulated Annealing algorithm
Talk Outline

Introduction to the *in vitro* problem

Introduction to the *in silico* problem

The Lagrangian Relaxation-based algorithm

The Simulated Annealing algorithm

Experimental Results

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Talk Outline

Introduction to the *in vitro* problem
Introduction to the *in silico* problem
The Lagrangian Relaxation-based algorithm
The Simulated Annealing algorithm
Experimental Results
Conclusions
Motivations

Microorganisms are fundamental for agriculture, biotechnology and medicine.
Motivations
Microorganisms are fundamental for agriculture, biotechnology and medicine

Problem: Difficult to classify
Motivations

Microorganisms are fundamental for agriculture, biotechnology and medicine

Problem: Difficult to classify
Solution: Analysis of microbial communities using rRNA sequences
Motivations

Microorganisms are fundamental for agriculture, biotechnology and medicine

Problem: Difficult to classify

Solution: Analysis of microbial communities using rRNA sequences

Tools: Hybridization fingerprints and DNA arrays
In Vitro Experiment

Microbial communities

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
In Vitro Experiment

Microbial communities

Clones

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
In Vitro Experiment

Microbial communities

Clones

Probes

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
In Vitro Experiment

Microbial communities

Clones

Probes

DNA array

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
In Vitro Experiment

Microbial communities

Clones

DNA array

Probes

Hybridization Experiment

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
### Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>c₁</th>
<th>c₂</th>
<th>...</th>
<th>cₘ</th>
</tr>
</thead>
<tbody>
<tr>
<td>p₁</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>p₂</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>pₙ</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>
## Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>Clones</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>...</td>
<td>$c_m$</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>$p_2$</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$p_n$</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

$p_1$ and $c_2$ hybridized
## Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>...</th>
<th>$c_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>$p_2$</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>$p_n$</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

$p_1$ and $c_2$ hybridized

$p_n$ and $c_m$ did not hybridize

Probes Selection Algorithms with Applications in the Analysis of Microbial Communities
## Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>c&lt;sub&gt;1&lt;/sub&gt;</th>
<th>c&lt;sub&gt;2&lt;/sub&gt;</th>
<th>...</th>
<th>c&lt;sub&gt;m&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>p&lt;sub&gt;1&lt;/sub&gt;</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>p&lt;sub&gt;2&lt;/sub&gt;</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>p&lt;sub&gt;n&lt;/sub&gt;</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Probe Selection Algorithms with Applications in the Analysis of Microbial Communities**

**Fingerprint**: \{p<sub>1</sub>, p<sub>2</sub>\}(\{c<sub>1</sub>\})
### Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>Clones</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$\ldots$</td>
<td>$c_m$</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
</tr>
<tr>
<td>$p_2$</td>
<td>1</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$p_n$</td>
<td>1</td>
<td>0</td>
<td>$\ldots$</td>
<td>0</td>
</tr>
</tbody>
</table>

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
### Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>Clones</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$\ldots$</td>
<td>$c_m$</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
</tr>
<tr>
<td>$p_2$</td>
<td>1</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
</tr>
<tr>
<td>$p_n$</td>
<td>1</td>
<td>0</td>
<td>$\ldots$</td>
<td>0</td>
</tr>
</tbody>
</table>

$\{p_1, p_2\}$ distinguishes $c_1$ from $c_2$.
### Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>( c_1 )</th>
<th>( c_2 )</th>
<th>( \ldots )</th>
<th>( c_m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 )</td>
<td>0</td>
<td>1</td>
<td>( \ldots )</td>
<td>1</td>
</tr>
<tr>
<td>( p_2 )</td>
<td>1</td>
<td>1</td>
<td>( \ldots )</td>
<td>1</td>
</tr>
<tr>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
<td>( \ldots )</td>
</tr>
<tr>
<td>( p_n )</td>
<td>1</td>
<td>0</td>
<td>( \ldots )</td>
<td>0</td>
</tr>
</tbody>
</table>

\( \{ p_1, p_2 \} \) distinguishes \( c_1 \) from \( c_2 \) but not \( c_2 \) from \( c_m \).
## Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>Clones</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$\ldots$</td>
<td>$c_m$</td>
</tr>
<tr>
<td>$p_1$</td>
<td>0</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
</tr>
<tr>
<td>$p_2$</td>
<td>1</td>
<td>1</td>
<td>$\ldots$</td>
<td>1</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>2?</td>
<td>$\ldots$</td>
</tr>
<tr>
<td>$p_n$</td>
<td>1</td>
<td>0</td>
<td>$\ldots$</td>
<td>0</td>
</tr>
</tbody>
</table>

$\{p_1, p_2\}$ distinguishes $c_1$ from $c_2$ but not $c_2$ from $c_m$. 

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
### Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>Clones</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$c_1$</td>
<td>$c_2$</td>
<td>$\cdots$</td>
<td>$c_m$</td>
</tr>
<tr>
<td>$p_1$</td>
<td></td>
<td>0</td>
<td>1</td>
<td>$\cdots$</td>
<td>1</td>
</tr>
<tr>
<td>$p_2$</td>
<td></td>
<td>1</td>
<td>1</td>
<td>$\cdots$</td>
<td>1</td>
</tr>
<tr>
<td>$\cdots$</td>
<td></td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
<td>$\cdots$</td>
</tr>
<tr>
<td>$p_n$</td>
<td></td>
<td>1</td>
<td>0</td>
<td>$\cdots$</td>
<td>0</td>
</tr>
</tbody>
</table>

The set $\{p_1, p_2\}$ distinguishes $c_1$ from $c_2$ but not $c_2$ from $c_m$. 

**Non-binary distinguishability**
### Result of the Experiment

<table>
<thead>
<tr>
<th>Probes</th>
<th>Clones</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(c_1)</td>
<td>(c_2)</td>
<td>(\ldots)</td>
<td>(c_m)</td>
</tr>
<tr>
<td>(p_1)</td>
<td>0</td>
<td>1</td>
<td>(\ldots)</td>
<td>1</td>
</tr>
<tr>
<td>(p_2)</td>
<td>1</td>
<td>1</td>
<td>(\ldots)</td>
<td>1</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(2?)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(p_n)</td>
<td>1</td>
<td>0</td>
<td>(\ldots)</td>
<td>0</td>
</tr>
</tbody>
</table>

\(\{p_1, p_2\}\) distinguishes \(c_1\) from \(c_2\) but not \(c_2\) from \(c_m\).

Non-binary distinguishability

\(\Delta_{[p_1,p_2]}\)

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Previous Work

G+C content [Cuticchia et.al. 1993][Fu et.al. 1992]

free energy and melting temperature [Li, Stormo 2000]

greedy algorithm [Herwig et.al. 2000]
The Problems

**Clones:** genomic sequences of species to be classified

**Probes:** synthetic oligonucleotides
The Problems

**Clones:** genomic sequences of species to be classified

**Probes:** synthetic oligonucleotides

**MCPS:** minimize the number of probes for distinguishing all pairs of clones

**MDPS:** maximize the number of pairs of clones that can be distinguished by a certain number of probes
The Problems

**Clones:** genomic sequences of species to be classified

**Probes:** synthetic oligonucleotides

**MCPS:** minimize the number of probes for distinguishing all pairs of clones

**MDPS:** maximize the number of pairs of clones that can be distinguished by a certain number of probes
Minimum Cost Probe Set

**Instance:** a set $C$ of clones and a set $P$ of $l$-mers

**Feasible solution:** a subset $S \subseteq P$ such that $\Delta_S = C \times C$

**Measure:** the number of probes in $S$
**MDPS**

**Maximum Distinguishing Probe Set**

**Instance**: a set $C$ of clones, a set $P$ of $l$-mers and an integer $k$

**Feasible solution**: a subset $S \subseteq P$ with $|S| = k$

**Measure**: $|\Delta_S|$, the number of pairs of clones that are distinguished by $S$. 

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Cost Functions

number of pairs of clones that are distinguished by $S$.

entropy

number of clusters

maximum size of a cluster
Bad News

Both MCPS and MDPS are \textbf{NP-hard} when the length of probes is unbounded.
Bad News
Both MCPS and MDPS are **NP-hard** when the length of probes is unbounded. Approximation algorithm for Set Cover and Maximum Coverage **do not have good guaranteed ratios**.
Lagrangian Relaxation

Minimize \(|S| = \sum_{p \in P} x_p\)

subject to \(\sum_{p \in P} \delta_{p,c,d} \cdot x_p \geq 1 \quad \forall (c, d) \in C^2\)

\(x_p \in \{0, 1\} \quad \forall p \in P\)

\(x_p = 1\) iff \(p \in S\)

\(\delta_{p,c,d} = 1\) iff \(p\) distinguishes \(c\) from \(d\)
Minimize $|S| = \sum_{p \in P} x_p$

subject to $\sum_{p \in P} \delta_{p,c,d} \cdot x_p \geq 1 \quad \forall (c, d) \in C^2$

$x_p \in \{0, 1\} \quad \forall p \in P$

$x_p = 1$ iff $p \in S$

$\delta_{p,c,d} = 1$ iff $p$ distinguishes $c$ from $d$

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Lagrangian Relaxation

\[ \text{Min} \quad L(x, \lambda) = \]

\[ = \sum_{p \in P} x_p + \sum_{(c,d) \in C^2} \lambda_{c,d} (1 - \sum_{p \in P} \delta_{p,c,d} \cdot x_p) \]
Lagrangian Relaxation

Min \( L(x, \lambda) = \)

\[
= \sum_{p \in P} x_p + \sum_{(c,d) \in C^2} \lambda_{c,d} (1 - \sum_{p \in P} \delta_{p,c,d} \cdot x_p)
\]

Given \( \lambda \), \( x_p = 0 \) if and only if

\[
\sum_{p \in P} (1 - \sum_{(c,d) \in C^2} \lambda_{c,d} \delta_{p,c,d}) x_p \geq 0
\]
Lagrangian Relaxation

\[ \text{Min } L(x, \lambda) = \]

\[ = \sum_{p \in P} x_p + \sum_{(c,d) \in C^2} \lambda_{c,d} \left(1 - \sum_{p \in P} \delta_{p,c,d} \cdot x_p\right) \]

Given \( \lambda, x_p = 0 \) if and only if

\[ \sum_{p \in P} \left(1 - \sum_{(c,d) \in C^2} \lambda_{c,d} \delta_{p,c,d}\right) x_p \geq 0 \]

Lagrangian Multipliers
Implementation

Good Multiplier

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Implementation

Good Multiplier

Actual Solution

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Implementation Issues

Problems

The constraint matrix has about 5000 columns and 12,000,000 rows.
The constraint matrix is dense (4GB).
Implementation Issues

Problems

The constraint matrix has about 5000 columns and 1200000 rows.
The constraint matrix is dense (4GB).

Solutions

Alternative representation of the constraint matrix.
Begin solving smaller subinstances.
The Experiments

**Dataset 1**: 1158 small-subunit ribosomal genes from GenBank (NCBI).

**Dataset 2**: 131 large-subunit ribosomal genes from the Ribosomal Database Project II.
### Experimental results - Dataset 1

<table>
<thead>
<tr>
<th>Length of probes</th>
<th>Distinguishability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>binary</td>
</tr>
<tr>
<td>5</td>
<td>42 (23)</td>
</tr>
<tr>
<td>6</td>
<td>48 (21)</td>
</tr>
<tr>
<td>8</td>
<td>56 (30)</td>
</tr>
</tbody>
</table>

number of probes (lower bound)
## Experimental results - Dataset 2

<table>
<thead>
<tr>
<th>Length of probes</th>
<th>Distinguishability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>binary</td>
</tr>
<tr>
<td>5</td>
<td>17 (11)</td>
</tr>
<tr>
<td>6</td>
<td>17 (9)</td>
</tr>
<tr>
<td>8</td>
<td>23 (14)</td>
</tr>
</tbody>
</table>

number of probes (lower bound)

---

Probes Selection Algorithms with Applications in the Analysis of Microbial Communities
Simulated Annealing

\( S := \text{set of } k \text{ random probes from } P \)

\( t := \text{initial temperature} \)

repeat

\[ S' := \text{a random neighbor of } S \]

\[ S := S' \text{ with probability } \exp\left(\frac{\text{cost}(S) - \text{cost}(S')}{t}\right) \}

\[ t := \frac{\beta t}{\beta + t} \]

until \( t \leq \text{final temperature} \)

return \( S \)
Simulated Annealing II

Two sets of probes $S_1$ and $S_2$ are neighbors if they differ in one probe.
Simulated Annealing II

Two sets of probes $S_1$ and $S_2$ are neighbors if they differ in one probe.

Cost functions considered: number of pairs of clones that are distinguished by $S$, entropy, maximum size of a cluster.
Implementation Issues

Problems

Memory for the fingerprint matrix.

Computing the cost of a solution. Naive algorithm $O(m^2k)$ time.
Implementation Issues

Problems

Memory for the fingerprint matrix.

Computing the cost of a solution. Naive algorithm $O(m^2k)$ time.

Solutions

Compact storing of the matrix.

Radix sorting using fingerprints as keys, $O(mk)$ time.
The Experiments

Dataset 1: 1158 small-subunit ribosomal genes from GenBank (NCBI).

Dataset 3: 5000 eubacteria samples.
Experimental Results - Dataset 1

Non-binary distinguishability, probe length 8

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Experimental Results - Dataset 1

Binary distinguishability, probe length 8

Entropy vs. Number of probes graph showing the performance of different probe selection algorithms: SA+entropy, SA+Pairs, SA+Largest, and Greedy.
Experimental Results - Dataset 3

Probe Selection Algorithms with Applications in the Analysis of Microbial Communities
Experimental Results - Dataset 3

Binary distinguishability, probe length 8

Probes Selection Algorithms with Applications in the Analysis of Microbial Communities
Further Research

Improve the running time of LR by choosing the initial solution in a different way.
Further Research

Improve the running time of LR by choosing the initial solution in a different way.

Improve the running time of the Simulated Annealing algorithm by adopting a different algorithm for computing the cost of a given solution.
Conclusions

Two algorithms for computing a set of probes for classifying microbial communities
Conclusions

Two algorithms for computing a set of probes for classifying microbial communities

They show good behavior in practice
Conclusions

Two algorithms for computing a set of probes for classifying microbial communities

They show good behavior in practice

- good approximate solutions
- feasible on real-world instances
Conclusions

Two algorithms for computing a set of probes for classifying microbial communities

They show good behavior in practice
- good approximate solutions
- feasible on real-world instances

We propose some directions for enhancements.