

Probe Selection Algorithms With Applications in the Analysis of Microbial Communities

James Borneman¹ Marek Chrobak² Gianluca Della Vedova³ Andres Figueroa²
Tao Jiang²

¹ Dept. Plant Pathology, Univ. of California at Riverside
² Dept. Computer Science, Univ. of California at Riverside
³ Dept. Statistics, Univ. Milano-Bicocca

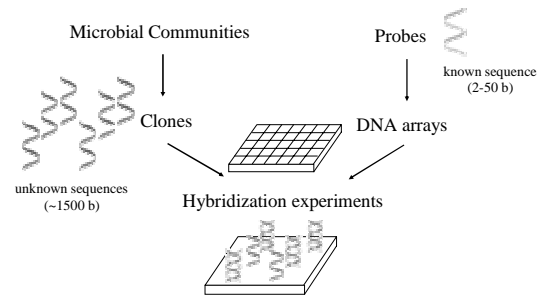
Talk Outline

- ⇒ • Motivations
- Introduction to the problem
- The Lagrangean relaxation-based algorithm
- The simulated annealing algorithm
- Experimental results

Motivations

- Microorganisms are fundamental for agriculture, biotechnology and medicine.
- Problem: Difficult to classify
- Solution: Analysis of microbial communities using ribosomal RNA genes (rDNA clones).
- Tools: Oligonucleotide fingerprinting.

Introduction



Microarray image



Arrayed bacterial rDNA clones hybridized with DNA probes.

Results of the Experiments

Clones	Probes				
	p ₁	p ₂	...	p _m	
c ₁	0	1	...	1	→ c ₁ and p ₂ hybridized
c ₂	1	0	...	0	→ c ₂ and p _m did not hybridized
...	
c _n	1	0	...	1	→ Fingerprint of c _n

Results of the Experiments

Clones	Probes				
	p_1	p_2	\dots	p_m	
c_1	1	1	\dots	1	p_2 distinguishes c_1 from c_2
c_2	1	0	\dots	0	
\dots	\dots	\dots	\dots	\dots	
c_n	1	0	\dots	1	

Results of the Experiments

Clones	Probes				
	p_1	p_2	\dots	p_n	
c_1	1	1	\dots	1	p_2 does not distinguishes c_2 from c_n
c_2	1	0	\dots	0	
\dots	\dots	\dots	\dots	\dots	
c_n	1	0	\dots	1	

Talk Outline

- Motivations
- ⇒ • Introduction to the problem
- The Lagrangean relaxation-based algorithm
- The simulated annealing algorithm
- Experimental results

The Problem

- Given a population C of m unknown rDNA clones.
- Choose a set S of probes of a given length l , such that any two clones c and d from C are distinguished by at least one probe in S .

What is a good probe set?

- The optimal probe set will contain as few oligonucleotides as possible.
- The number of probes used is exactly the number of hybridization experiments, so this will minimize experimental cost and effort.

Minimum Cost Probe Set (MCPS)

- Instance: a set C of clones and a set P of probes
- Feasible solution: a subset $S \subseteq P$ such that $A_S = C \times C$.
- MCPS: minimize the number of probes for distinguishing all pairs of clones
- Measure: the number of probes in S .

Bad News

- MCPS is NP-hard when the length of probes is unbounded.
- Approximation algorithm for Set Cover do not have good guaranteed ratios.

Mathematical formulation of MCSP

$$\begin{aligned} &\text{Minimize } |S| = \sum_{p \in P} x_p \\ &\text{Subject to } \sum_{p \in P} \delta_{p c d} \cdot x_p \geq 1, \forall (c, d) \in C^2 \\ &\quad x_p \in \{0, 1\} \quad \forall p \in P \end{aligned}$$

where

$$x_p = 1 \text{ iff } p \in S$$

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

Talk Outline

- Motivations
- Introduction to the problem
- ⇒ The Lagrangean relaxation-based algorithm
- The simulated annealing algorithm
- Experimental results

Lagrangean Lower Bound Program (LLBP)

$$\begin{aligned} &\text{Minimize } L(x, \lambda) = \sum_{p \in P} x_p + \\ &\quad \sum_{(c, d) \in C^2} \lambda_{c, d} (1 - \sum_{p \in P} \delta_{p c d} \cdot x_p) \end{aligned}$$

Lagrangean
Multipliers

$$\begin{aligned} &\text{Subject to } x_p \in \{0, 1\} \quad \forall p \in P \\ &\quad \lambda_{c, d} \geq 0 \quad \forall (c, d) \in C^2 \end{aligned}$$

LLBP (cont.)

$$\text{Minimize } L(x, \lambda) = \sum_{p \in P} C_p(\lambda) x_p + \sum_{(c, d) \in C^2} \lambda_{c, d}$$

$$C_p(\lambda) = 1 - \sum_{(c, d) \in C^2} \lambda_{c, d} \delta_{p c d}$$

Solution

$$x_p = \begin{cases} 1; & \text{if } C_p(\lambda) < 0 \\ 0; & \text{if } C_p(\lambda) > 0 \\ 0 \text{ or } 1; & \text{otherwise} \end{cases}$$

LLBP (cont.)

- Given a vector λ , let x be an optimal solution of LLBP.
- x is not necessarily feasible for MCSP
- A naïve feasible extension of x is, at each step, choose a probe that covers most yet uncovered pairs.

Feasible Extension

- Caprara *et al.*, 1999.

$$score_p(S, \lambda) = \begin{cases} \gamma_p(S) \mu_p(S); & \text{if } \gamma_p(S) < 0 \\ \gamma_p(S) / \mu_p(S); & \text{if } \gamma_p(S) > 0 \\ \infty; & \text{if } \mu_p(S) = 0 \end{cases}$$

where, $\mu_p(S) = |\Delta_p - \Delta_S|$, and
 $\gamma_p(S) = 1 - \sum_{(c,d) \in \Delta_p - \Delta_S} \lambda_{c,d}$

- at each step we add q that is not in S with the minimum score.

Subgradient Optimization

- Held & Karp, 1971.

$$\lambda^{i+1} = \max \left\{ \lambda^i + \alpha \frac{\langle \mathbf{x}^* | -L(\lambda^i) \rangle \nabla \lambda, \mathbf{0}}{\|\nabla \lambda\|} \right\}$$

where,

\mathbf{x}^* is the best feasible solution found so far,

$\alpha > 0$,

$\nabla \lambda = \langle \nabla_{c,d} \lambda | (c,d) \in C^2 \rangle$, and

$$\nabla_{c,d} \lambda = 1 - \sum_{p \in P} \delta_{p,c,d} \cdot x_p$$

Subgradient Optimization Procedure

```

Procedure SUBGRADOPTIMIZE( $\lambda, S^*, LB$ )
 $\alpha \leftarrow \text{initialAlpha}$ 
while  $\alpha > \text{MinAlpha}$  do
   $PrevS \leftarrow LB$ 
  repeat  $LB$ IncreaseTries times
     $LE \leftarrow \text{LFSolve}(\lambda, S)$ 
     $\lambda \leftarrow L(\lambda)$ 
     $LB \leftarrow \max(LB, LE)$ 
     $FS \leftarrow \text{FEASIBLESOLUTION}(\lambda, S)$ 
    if  $|S| < |S^*|$  then  $S^* \leftarrow S$ 
     $\lambda \leftarrow \max(\lambda, \alpha \frac{\langle \mathbf{x}^* | -L(\lambda) \rangle \nabla \lambda, \mathbf{0}}{\|\nabla \lambda\|})$ 
  if  $LB = PrevS$  then  $\alpha \leftarrow \alpha/2$ 
end
  
```

Lagrangean Relaxation Algorithm

```

Algorithm LR( $P, C$ )
 $S^* \leftarrow \emptyset$ 
 $\lambda_{c,d} \leftarrow \text{initial}(\lambda_{c,d})$  for all  $(c,d) \in C$ 
 $LB \leftarrow 0$ 
SUBGRADOPTIMIZE( $\lambda, S^*, LB$ )
Accuracy  $\leftarrow |S^*|/LB$ 
Print( $S^*, Accuracy$ )
end
  
```

Maximum Distinguishing Probe Set (MDPS)

- Instance: a set C of clones, a set P of probes and an integer k
- Feasible solution: a subset $S \subseteq P$ with $|S| = k$.
- MDPS: maximize the number of pairs of clones that can be distinguished by a certain number of probes
- Measure: $|\Delta_S|$, the number of pairs of clones that are distinguished by S .

Bad News

- MDPS is NP-hard when the length of probes is unbounded.
- Approximation algorithm for Maximum Coverage do not have good guaranteed ratios.

Talk Outline

- Motivations
- Introduction to the problem
- The Lagrangean relaxation-based algorithm
- ⇒ • The simulated annealing algorithm
- Experimental results

Simulated Annealing Algorithm

S := set of k random probes from P

t := initial temperature

repeat

S' := a random neighbor of S

$S := S'$ with probability

$$t = \frac{\beta}{\beta + t}$$

until $t \leq$ final temperature

return(S)

Neighbor Solution

- Each solution S is a set of k probes.
- A neighbor of S is a set S' of k probes iff $|S \cap S'| = k - 1$

Cost Functions

- number of pairs of clones that are distinguished by S , denoted by $|A_S|$
- entropy (Herwig *et al.*, 2000)
 - Let $\{C_1, \dots, C_Z\}$ be the clustering of C induced by A_S , then the entropy of this solution S is defined as
 - $-\sum_{i=1}^Z p_i \log p_i$, where $p_i = |C_i|/|C|$
- maximum size of a cluster

Initial Temperature

- The initial temperature t_0 is computed as follows:
 - An initial solution S is uniform randomly chosen.
 - 100 neighbors of S are extracted randomly, and let $h = \max \{|\text{cost}(S) - \text{cost}(S')|\}$ among all such solutions. h is an indicator of the maximum hill that the algorithm might climb in a single step.
 - t_0 is the temperature such that it is possible to climb a hill of height h with probability 0.9, that is,
 $0.9 = \exp(-h/t_0)$

Talk Outline

- Motivations
- Introduction to the problem
- The Lagrangean relaxation-based algorithm
- The simulated annealing algorithm
- ⇒ • Experimental results

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

Length of probes	Distinguishability	
	binary	non-binary
5	41(23)	21(11)
6	48(21)	29(15)
8	56(30)	46(28)

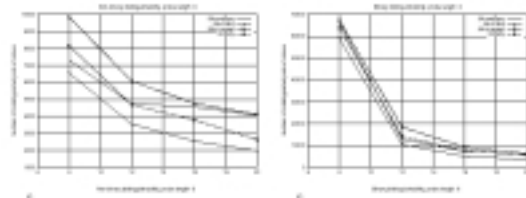
number of probes (lower bound)

Experimental Results, Dataset 2

Length of probes	Distinguishability	
	binary	non-binary
5	17(11)	11(8)
6	17(9)	14(10)
8	23(14)	21(13)

number of probes (lower bound)

Results



Some references

- Caprara, A., Fischetti, M & Toth, P. (1999). A heuristic method for the set covering problem. *Operations Research*, **47**, 730-743.
- Held, M. & Karp, R. M. (1971). The traveling salesman problem and minimum spanning trees: Part II. *Mathematical Programming*, **1**, 6-25.
- Herwig *et al.* (2000). Information theoretical probe selection for hybridization experiments. *Bioinformatics*, **10**, 890-898.

Thanks

?