



# DNA array data analysis

Andres Figueroa, James Borneman, Tao Jiang  
 Computer Science Department, University of California, Riverside, CA 92521

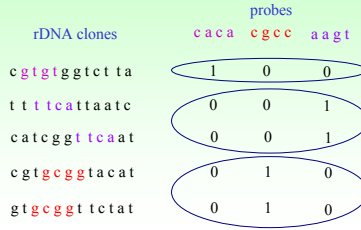
## ABSTRACT

Oligonucleotide fingerprinting is a powerful DNA array based method to characterize cDNA and ribosomal RNA gene (rDNA) libraries. A key step in the method is the cluster analysis of fingerprint data obtained from DNA array hybridization experiments. Most of the existing approaches to clustering use (normalized) real intensity values and thus do not treat positive and negative hybridization signals equally. In this work, we consider a discrete approach. Fingerprint data are first normalized and binarized using control DNA clones. Because there may exist unresolved (or missing) values in this binarization process, we formulate the clustering of (binary) oligonucleotide fingerprints as a combinatorial optimization problem that attempts to identify clusters and resolve the missing values in the fingerprints simultaneously.

We study the computational complexity of this clustering problem and a natural parameterized version, and present an efficient greedy algorithm based on Minimum Clique Partition on graphs.

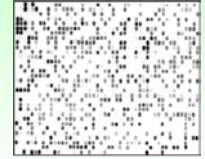
The algorithm takes advantage of some unique properties of the graphs considered here, which allow us to efficiently find the maximum cliques as well as some special maximal cliques efficiently. Our experimental results on simulated and real data demonstrate that the algorithm runs faster and performs better than popular clustering methods.

## 1 Characterize rDNA clones by their binary fingerprints

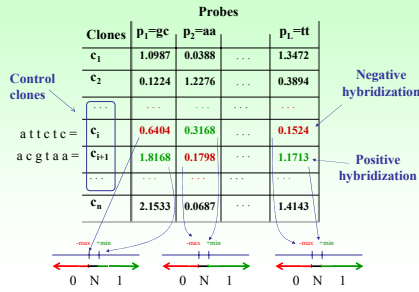


## 2 Microarray image

- Microarray image of a probe hybridized to 1536 clones.
- Control clones are included.



## 3 Hybridization matrix



## 4 Binary fingerprint vectors with missing values

Clones	$p_1=gc$	$p_2=aa$	$p_3=tt$
$c_1$	N	0	1
$c_2$	0	1	N
...	...	...	...
$c_i$	0	1	0
$c_{i+1}$	1	0	1
...	...	...	...
$c_n$	1	0	1

Annotations:  $c_1$  and  $p_2$  did not hybridize; missing value;  $c_{i+1}$  and  $p_1$  hybridized; fingerprint vector of  $c_n$ .

## 5 Binary clustering with missing values (BCMV(p))

- Instance:** a set  $F$  of 0-1-N fingerprint vectors with at most  $p$  missing values per vector.
- Feasible solution:** a partition of  $F$  into disjoint subsets  $F_1, F_2, \dots, F_k$  such that, for  $1 \leq i \leq k$ , any two fingerprint vectors in  $F_i$  are compatible.
- Measure:** Cardinality of the partition, to be minimized.

## 6 Computational complexity

- Theorem:** The problem BCMV( $p$ ) is NP hard, for any  $p \geq 3$ .
- Theorem:** The problem BCMV(1) can be solved in polynomial time.
- Theorem:** For any  $p$ , BCMV( $p$ ) can be approximated in polynomial time with ratio  $2^p$ .

## 7 Greedy clique partition algorithm (GCP)

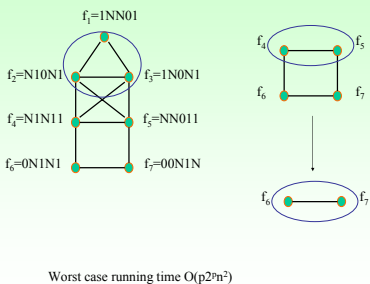
Algorithm: GCP( $F, \mathcal{C}$ )

```

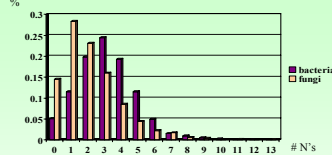
V ← F;
i ← 0;
repeat
  while VertexInUniqueMaximalClique(V, V) do
    V ← V \ V_i;
    i ← i + 1;
  endwhile;
  V_i ← FindMaximumClique(V);
  V ← V \ V_i;
  i ← i + 1;
until V = ∅;
C ← {V_i | i = 0, 1, ..., i};
end;
    
```

A Java implementation of GCP is available at [www.cs.ucr.edu/~andres](http://www.cs.ucr.edu/~andres)

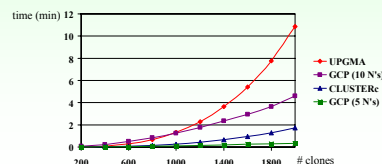
## 8 Example



## 9 Frequency of missing values in real data



### Average execution time



## 10 Experimental results

Performance of GCP vs. popular binary clustering methods on simulated data

n	l	d	p	GCP		UPGMA		CLUSTER	
				Jacard's coefficient	Number of clusters	Jacard's coefficient	Number of clusters	Jacard's coefficient	Number of clusters
2000	25	551	5	1.0	551	0.99950	551.8	0.85467	882.2
2000	25	551	10	0.99388	551	0.94449	626.4	0.35756	1548.2
2000	25	807	5	0.99944	807	0.99847	808.4	0.79545	1164.8
2000	25	807	10	0.95944	807	0.87338	872.8	0.33675	1643.8

Performance of GCP vs. popular binary clustering methods on real data

Data set	n	l	p	Number of Clusters		
				GCP	UPGMA	CLUSTER
Bacteria	1491	27	3.84	769	773	991
Fungi	1507	26	4.54	556	566	870

Performance of popular real intensity clustering methods in terms of incompatibility

	DRMANAC	R-UPGMA	R-CLUSTER	CLICK
Bacteria	0.7542	2.59	0.8064	2.75
Fungi	0.8315	2.21	0.6666	1.31