

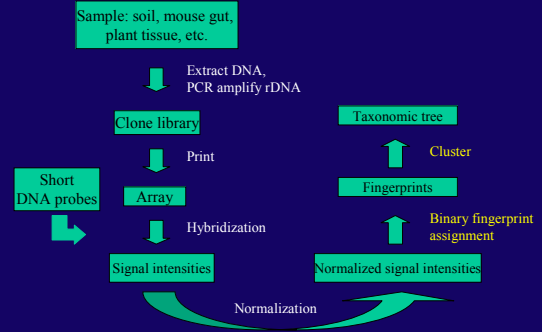
Clustering binary fingerprint vectors with missing values for DNA array data analysis

Andres Figueroa¹ James Borneman² Tao Jiang¹
¹ Dept. Computer Science, Univ. of California at Riverside
² Dept. Plant Pathology, Univ. of California at Riverside

8/12/03

University of California, Riverside, CA

Microorganism classification by Oligonucleotide Fingerprinting of Ribosomal Genes (OFRG)

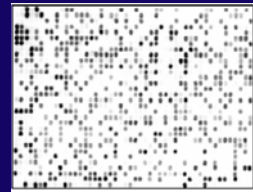


rDNA clones characterized by their binary fingerprints

rDNA clones	probes		
	caca	cgcc	aagt
cgtgtggtctta	1	0	0
tttctattaatc	0	0	1
catcggtcaat	0	0	1
cgtgcggtacat	0	1	0
gtgcggttctat	0	1	0

Microarray image

- Microarray image of 1536 rDNA clones hybridized with the probe AATTCGATGC.
- Control clones are included.



Hybridization Matrix

Clones	Probes			
	p ₁	p ₂	...	p _L
c ₁	1.0987	0.0388	...	1.3472
c ₂	0.1224	1.2276	...	0.3894
...
c _i	0.6404	0.3168	...	0.1524
c _{i+1}	1.8168	0.1798	...	1.1713
...
c _n	2.1533	0.0687	...	1.4143

Control clones

Clones	p ₁ =gc	p ₂ =aa	...	p _L =tt
c ₁	1.0987	0.0388	...	1.3472
c ₂	0.1224	1.2276	...	0.3894
...
attctc = c _i	0.6404	0.3168	...	0.1524
acgtaa = c _{i+1}	1.8168	0.1798	...	1.1713
...
c _n	2.1533	0.0687	...	1.4143

Positive hybridization

Negative hybridization

Binary conversion

Clones	$p_1 = gc$	$p_2 = aa$...	$p_L = tt$
c_1	1.0987	0.0388	...	1.3472
c_2	0.1224	1.2276	...	0.3894
...
c_i	0.6404	0.3168	...	0.1524
c_{i+1}	1.8168	0.1798	...	1.1713
...
c_n	2.1533	0.0687	...	1.4143

Control clones

attctc =
acgtaa =

0-1-N classification matrix

Clones	Probes				
	p_1	p_2	...	p_L	
f_1	N	0	...	1	f_1 and f_n did not hybridized
f_2	0	1	...	N	missing value
...	
f_i	0	1	...	0	f_{i+1} and f_i hybridized
f_{i+1}	1	0	...	1	
...	fingerprint vector of f_n
f_n	1	0	...	1	

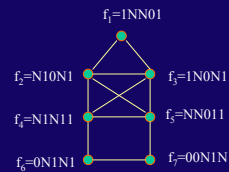
Question: How to cluster these imperfect binary fingerprints while resolving the N's?

- Definition: Let f and g be two 0-1-N fingerprint vectors of length L . We say the f and g are **compatible**, if $f[i] \neq g[i]$, then $f[i]=N$ or $g[i]=N$, for $0 \leq i \leq L$. A set of mutually compatible fingerprint vectors form a **cluster**.

f_1 and f_2 are compatible
 f_1 and f_n are not compatible

Clones	Probes			
	p_1	p_2	...	p_L
f_1	0	1	...	N
f_2	0	1	...	0
...
f_n	1	N	...	1

- Definition: Given a set of 0-1-N fingerprint vectors F , define a graph $G_F = (F, E_F)$ where two vertices (fingerprints) are adjacent if and only if they are compatible. The graph G_F will be called the **compatibility graph** of F .



Binary clustering with missing values

- BCMV
 - Instance: a set F of 0-1-N fingerprint vectors.
 - 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 0-1-N fingerprint vectors in F_i are compatible.
 - Measure: Cardinality of the partition, to be minimized.
- BCMV(p)
 - Instance: a set F of 0-1-N fingerprint vectors with at most p missing values per vector.

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 .

An efficient heuristic for BCMV(p):

Greedy Clique Partition Algorithm (GCP)

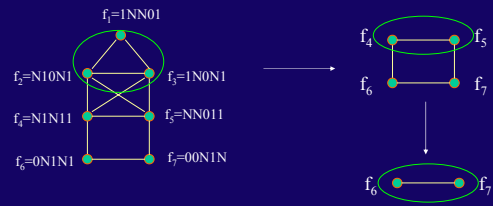
Algorithm: GCP (F, \mathcal{C})

```

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

A java implementation of GCP is available at www.es.ucr.edu/~andres

Example



Worst case running time $O(p^2n^2)$

Experimental results on simulated data

n	L	d	p	GCP		UPGMA ¹		CLUSTERc ²	
				Jaccard's coefficient	Number of clusters	Jaccard's coefficient	Number of clusters	Jaccard's coefficient	Number of clusters
2000	25	500	5	0.99965	500	0.99665	501.6	0.75497	872.2
2000	25	500	10	0.98512	500	0.85452	575.6	0.35392	1618.8
2000	25	551	5	1.0	551	0.99950	551.8	0.83467	892.2
2000	25	551	10	0.99388	551	0.94449	626.4	0.35756	1548.2
2000	25	800	5	0.99999	800	0.99990	802.4	0.93228	1176
2000	25	800	10	0.99902	800	0.99002	891	0.34841	1510
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

¹ D.L. Swofford. *PAUP: Phylogenetic Analysis Using Parsimony version 4.0 beta 10*. Sinauer Associates, Sunderland, Massachusetts, 2002.

² M.B. Eisen *et al.* *Cluster analysis and display of genome-wide expression patterns*. Proc. Nat'l Acad Sci USA, 95:14863-14868, 1998

Real data

- The first data set is a collection of 1491 bacterial small subunit rDNA genes. [Borneman *et al.* 2002].
- The second data is a set of 1507 fungal small subunit rDNA genes. [Valinsky *et al.* 2002]

Experimental results on real data

Data set	n	L	\bar{p}	Number of Clusters		
				GCP	UPGMA ¹	CLUSTERc ²
Bacteria	1491	27	3.84	769	773	991
Fungi	1507	26	4.54	556	566	870

¹ D.L. Swofford. *PAUP: Phylogenetic Analysis Using Parsimony version 4.0 beta 10*. Sinauer Associates, Sunderland, Massachusetts, 2002.

² M.B. Eisen *et al.* *Cluster analysis and display of genome-wide expression patterns*. Proc. Nat'l Acad Sci USA, 95:14863-14868, 1998

Performance of popular real intensity clustering methods in terms of incompatibility

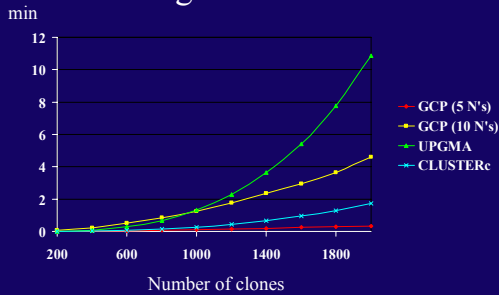
	DRMANAC ¹		R-UPGMA		R-CLUSTERc		CLICK ²	
Bacteria	0.7542	2.59	0.8064	2.75	0.7041	1.97	0.9310	4.13
Fungi	0.8315	2.21	0.6666	1.31	0.6809	1.53	0.9184	2.94

Average # of incompatible positions between fingerprints and average # of incompatible fingerprints in a cluster

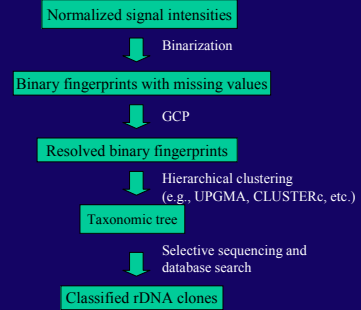
¹ S. Dramanac, *et al.* Gene representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics*, 37:29-40, 1996.

² R. Sharan and R. Shamir. CLICK: A clustering algorithm with applications to gene expression analysis. *Proc. ISMB* 2000, 307-316, 2000.

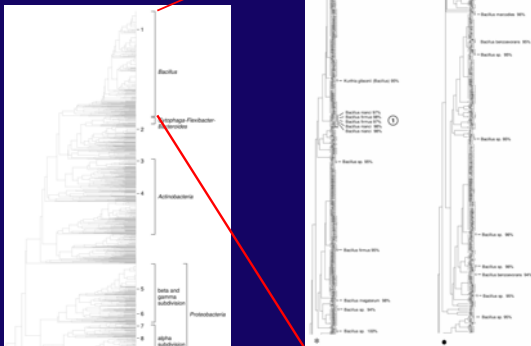
Average execution times



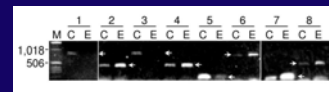
Applications in Microorganism Classification



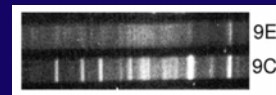
Taxonomic tree



Compositional differences between two soil samples, 9C and 9E, by OFRG and DGGE



PCR primer distribution of clones in the 8 clusters obtained by OFRG.



PCR primer distribution of clones by using DGGE.

Summary

- An efficient heuristic for $BCM(p)$.
- $BCM(p)$ is NP hard, for any $p \geq 3$.
- $BCM(1)$ can be solved in polynomial time.
- Experiments on simulated and real data.
- A java implementation of GCP is available at www.cs.ucr.edu/~andres

Thanks.
Questions? Comments?