TOWARD AN IMPROVED COMBINATORIC ALGORITHM

Ekaterina Nosova

DMI – Dept of Mathematics and Informatics,
University of Salerno, Italy

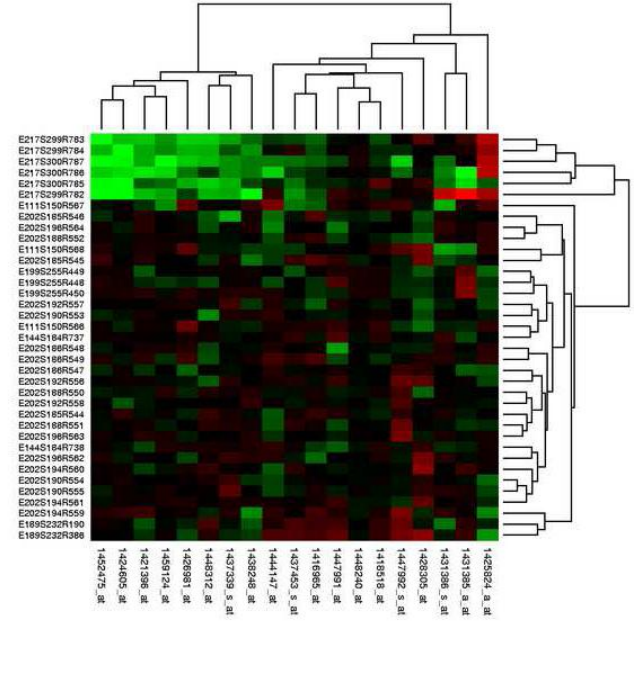
Outline

- Introduction to biclustering problem.
  - Data Sets
  - Biclustering
  - Task of biclustering
- Bicluster definition.
- Combinatoric algorithm.
  - CBA theory.
  - Error definition
  - Initial conditions
  - Obtaining of combinatorial matrix.
  - Bimax
- Results
- Conclusions

Introduction

Data sets

- Data sets are provided, for example, by the DNA Microarray technology. Where the results of the experiments carried out on genes under different conditions are the expression levels of their transcribed mRNA stored in DNA chips.



If two genes are related (they have similar functions or are co-regulated), their expression profiles should be similar.

Introduction

- **Clustering (Unsupervised):** Given a set of samples, partition them into groups containing similar samples according to some similarity criteria (CLASS DISCOVERING).
- **Classification (Supervised):** Find classes of the test data set using known classification of training data set (CLASS PREDICTION).
- **Feature Selection (Dimensionality reduction):** Select a subset of features responsible for creating the condition corresponding to the class (GENE SELECTION, BIOMARKER SELECTION).

Introduction

Biclustering

- If two genes are related, they can have similar expression patterns only **under some conditions** (e.g. they have similar response to a certain external stimulus, but each of them has some distinct functions at other time).
- Similarly, for two related conditions, some genes may exhibit different expression patterns (e.g. two tumor samples of different sub-types).
- As a result, each cluster may involve only a subset of genes and a subset of conditions.

Introduction

Biclustering

- Biclustering is a Simultaneous clustering of both rows and columns of a data Matrix.
- Concept can be traced back to the 70’(Hartigan, 1972), although it has been rarely used or studied.
- The term was introduced by (Cheng and Church, 2000) who were the first to use gene expression data analysis.
- The technique used in many fields, such as collaborative filtering, information retrieval and data mining.
- Other Names: simultaneous clustering, co-clustering, two-way clustering, subspace clustering, bi-dimensional clustering,.

Introduction

- Microarray data can be viewed as an $m \times n$ matrix X:

$$X = (x_{ij})_{m \times n}$$

  - Each of the m rows represents a gene (or a clone, ORF, etc.).
  - Each of the n columns represents a condition (a sample, a time point, etc.).
  - Each entry represents the expression level of a gene under a condition. It can either be an absolute value (e.g. Affymetrix GeneChip) or a relative expression ratio (e.g. cDNA microarrays).

Introduction

Biclustering

- An interesting criteria to evaluate a biclustering algorithm concerns the identification of the type of biclusters the algorithm is able to find.

- We identified three major classes of biclusters
  - Biclusters with constant values.
  - Biclusters with constant values on rows or columns.
  - Biclusters with coherent values.

$aij = \mu$

1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0

$aij = \mu + \beta j$

1.0	1.0	1.0	1.0
2.0	2.0	2.0	2.0
3.0	3.0	3.0	3.0
4.0	4.0	4.0	4.0

$aij = \mu + \alpha i + \beta j$

1.0	2.0	5.0	0.0
2.0	3.0	6.0	1.0
4.0	5.0	8.0	3.0
5.0	6.0	9.0	4.0

Additive Model

$aij = \mu \times \alpha i \times \beta j$

1.0	2.0	0.5	1.5
2.0	4.0	1.0	3.0
4.0	8.0	2.0	6.0
3.0	6.0	1.5	4.5

Multiplicative Model

Bicluster definition

$$n = n_g \times n_c;$$

$$x_{ij} = \mu + \alpha_i + \beta_j; \alpha_i = x_{iJ} - x_{IJ}; \beta_j = x_{Ij} - x_{IJ};$$

$$x_{Ij} = \frac{\sum_i x_{ij}}{n_g} = \frac{n_g \mu + \sum_i \alpha_i + n_g \beta_j}{n_g};$$

$$x_{iJ} = \frac{\sum_j x_{ij}}{n_c} = \frac{n_c \mu + \sum_j \beta_j + n_c \alpha_i}{n_c};$$

$$x_{IJ} = \frac{\sum_{ij} x_{ij}}{n} = \frac{n\mu + n_c \sum_i \alpha_i + n_g \sum_j \beta_j}{n};$$

$$d_{ij} = x_{ij} - (x_{IJ} + \alpha_i + \beta_j);$$

$$H = \sum_{ij} {d_{ij}}^2; G = \frac{\sum_{ij} {d_{ij}}^2}{n}$$

Let X be the bicluster of size n end the elements x_{ij}

bicluster row mean and

bicluster column mean

bicluster mean

residue [Cheng & Church, 2000]

Sum-squared residue and MSR

Overview of the Biclustering Methods

Method	Publish	Cluster Model	Goal
Cheng & Church	ISMB 2000	Background + row effect + column effect	Minimize mean squared residue of biclusters
Getz et al. (CTWC)	PNAS 2000	Depending on plugin clustering algorithm	Depending on plugin clustering algorithm
Lazzeroni & Owen (Plaid Models)	Bioinformatics 2000	Background + row effect + column effect	Minimize modeling error
Ben-Dor et al. (OPSM)	RECOMB 2002	All genes have the same order of expression values	Minimize the p-values of biclusters
Tanay et al. (SAMBA)	Bioinformatics 2002	Maximum bounded bipartite subgraph	Minimize the p-values of biclusters
Yang et al. (FLOC)	BIBE 2003	Background + row effect + column effect	Minimize mean squared residue of biclusters
Kluger et al. (Spectral)	Genome Res. 2003	Background × row effect × column effect	Finding checkerboard structures

Combinatorial Biclustering Algorithm

Problems of other techniques:

1. Precision
2. Noise Control
3. Initialization
4. Overlapping
5. Finding of all biclusters
6. Multi - biclustering solutions

CBA theory

$$x_{ij} = \mu + \alpha_i + \beta_j$$

1. Precision

$$\begin{vmatrix} \alpha_1+\beta_1 & \alpha_1+\beta_2 & \alpha_1+\beta_3 & \dots & \alpha_1+\beta_m \\ \alpha_2+\beta_1 & \alpha_2+\beta_2 & \alpha_2+\beta_3 & \dots & \alpha_2+\beta_m \\ \alpha_3+\beta_1 & \alpha_3+\beta_2 & \alpha_3+\beta_3 & \dots & \alpha_3+\beta_m \\ \dots & \dots & \dots & \dots & \dots \\ \alpha_n+\beta_1 & \alpha_n+\beta_2 & \alpha_n+\beta_3 & \dots & \alpha_n+\beta_m \end{vmatrix}$$

$$\begin{vmatrix} \alpha_1+\beta_1 & \alpha_1+\beta_2 & \dots & \alpha_1+\beta_m - \\ \alpha_2+\beta_1 & \alpha_2+\beta_2 & \dots & \alpha_2+\beta_m = \\ \hline \alpha_1-\alpha_2 & \alpha_1-\alpha_2 & \dots & \alpha_1-\alpha_2 \end{vmatrix}$$

If we calculate the difference between every two rows of the bicluster we obtain equal constant values. So we construct the matrix

$$T = \begin{bmatrix} G_1 \\ \dots \\ G_{N-1} \end{bmatrix}$$

Error definition

2. Noise Control

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \end{bmatrix}$$

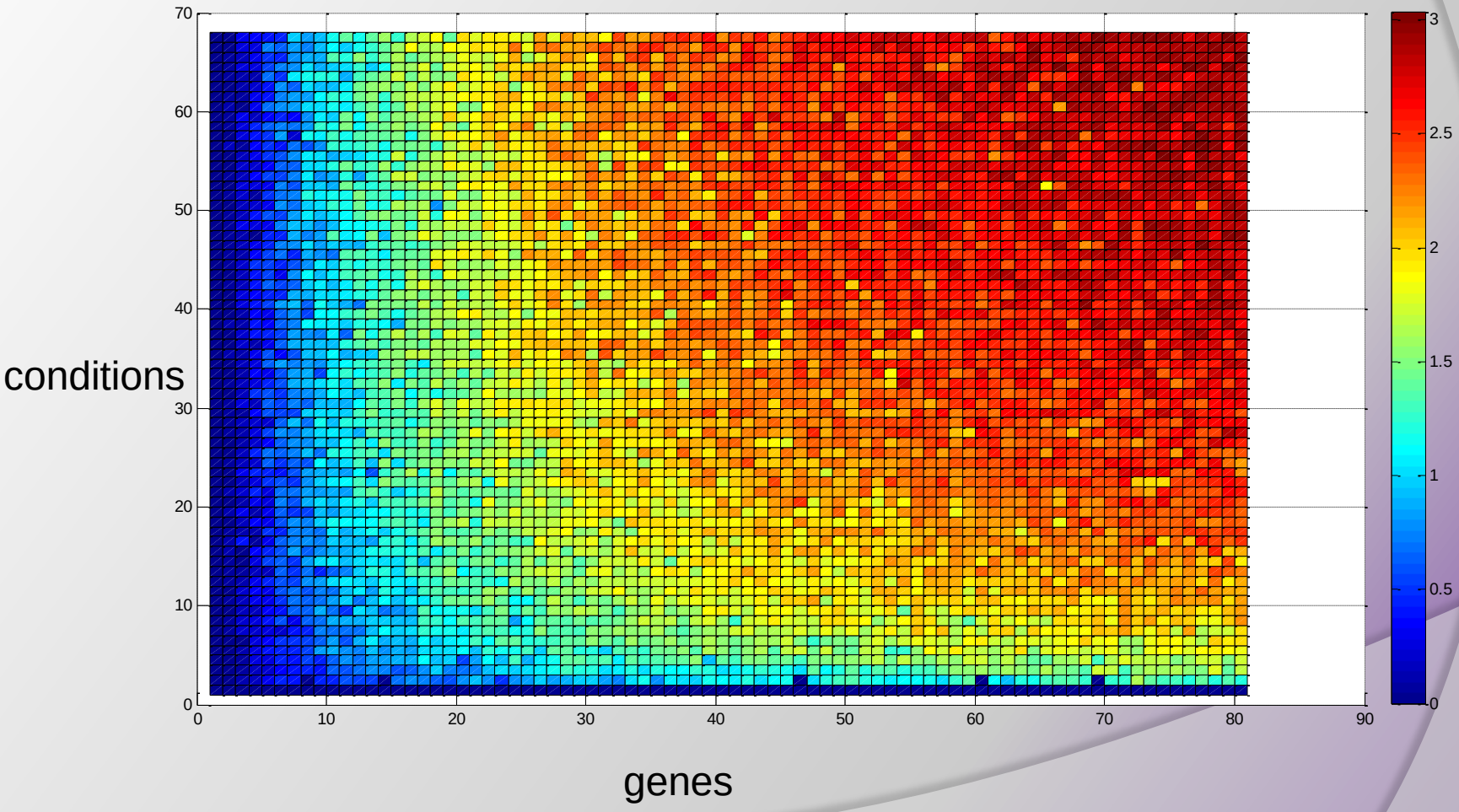
With the columns:

$$a_1 = [x_{11} \quad x_{21} \quad x_{31} \quad x_{41}]$$
$$a_2 = [x_{12} \quad x_{22} \quad x_{32} \quad x_{42}]$$
$$a_3 = [x_{13} \quad x_{23} \quad x_{33} \quad x_{43}]$$
$$a_4 = [x_{14} \quad x_{24} \quad x_{34} \quad x_{44}]$$

$$error = \max \begin{Bmatrix} \max(a_1) - \min(a_1) \\ \max(a_2) - \min(a_2) \\ \max(a_3) - \min(a_3) \\ \max(a_4) - \min(a_4) \end{Bmatrix}$$

Initial conditions

3. Initialization



Obtaining of combinatorial matrix

$$X = \begin{bmatrix} 1 & 1 & 1 & 2 & x & x & x \\ 1 & 1 & 1 & 2 & 3 & 4 & x \\ x & x & 0 & 1 & 2 & 3 & x \end{bmatrix}$$

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 & x & x & x \\ x & x & 1 & 1 & x & x & x \\ x & x & 1 & 1 & 1 & 1 & x \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

4. Overlapping

$$T = \begin{bmatrix} G_1 \\ \dots \\ G_{N-1} \end{bmatrix}$$

Obtaining of combinatorial matrix

4. Overlapping

Let us take the first row of *T* that contains 3 groups of the constants: c_1, c_2, c_3

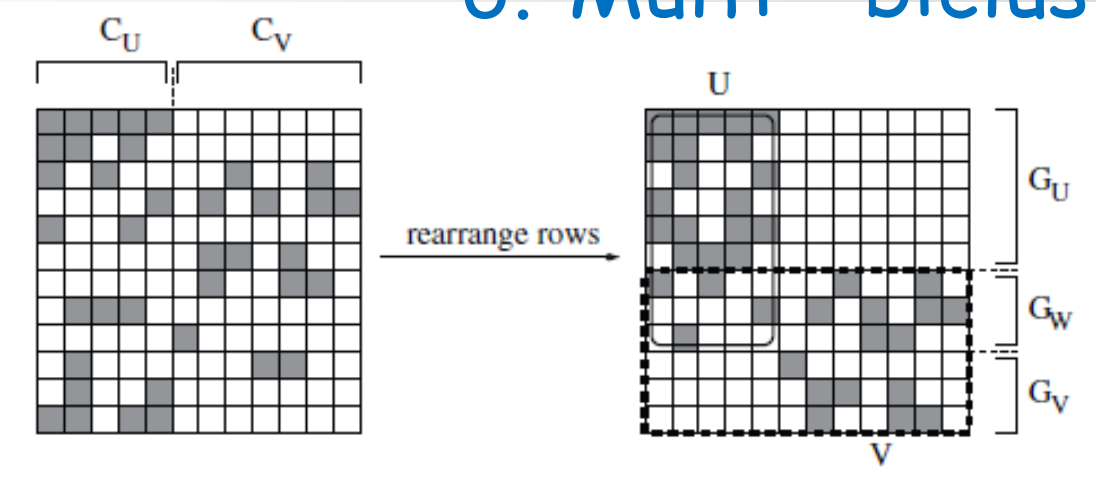
$$t_1 = \begin{bmatrix} c_1 & c_1 & c_1 & c_2 & c_2 & c_3 & c_3 \end{bmatrix}$$

We construct the matrix C_1 in the way:

$$C_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Bimax

5. Finding of all biclusters
6. Multi - biclustering solutions

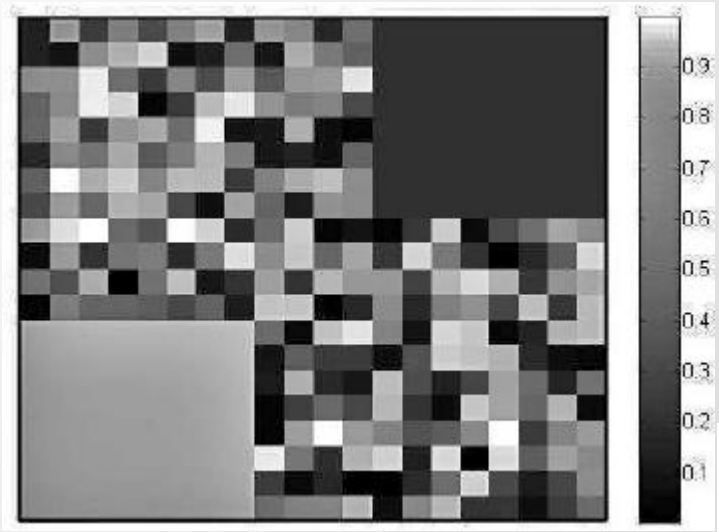


- We divide the input matrix E into two smaller sub-matrices U and V
- The set of columns is divided into two subsets C_U and C_V, here by taking the first row as a template.
- The rows of E are resorted:
  - 1. the genes that respond only to conditions given by C_U,
  - 2. those genes that respond to conditions in C_U and in C_V
  - 3. the genes that respond to conditions in C_V only.

  The corresponding sets of genes are GU, GW and GV

Results

The matrix 20×20

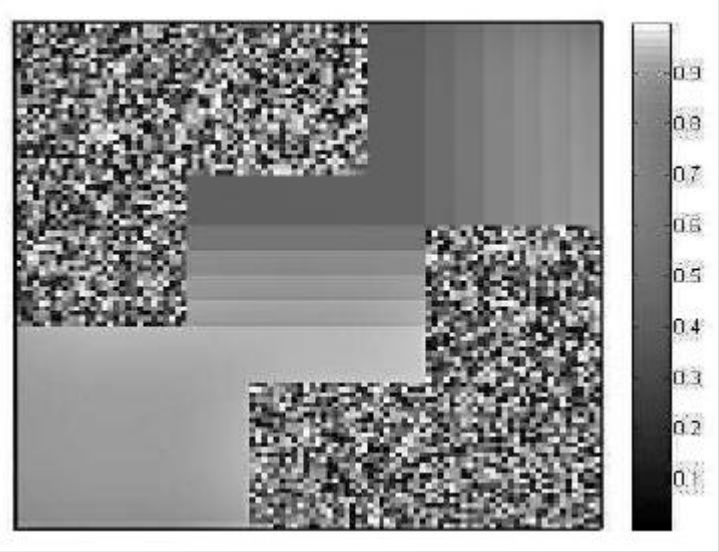


- 1. The simple matrix 20×20 with two biclusters

	Plaid	C&C	Spect	SAMBA	CBA	Theor
N of bic	1	2	22	-	2	2
no overl	1	2	2	-	2	2
the best	1	2	1	-	2	2
MSR1	0.02	0	-	-	0	0
MSR2	-	0	0.04	-	0	0
E1	1.67	2.5	-	-	2.5	2.5
E2	-	2.5	1.39	-	2.5	2.5
dim1	6×6	8×7	9×9	-	8×8	8×8
dim2	-	7×8	-	-	8×8	8×8

Results

The matrix 100×100



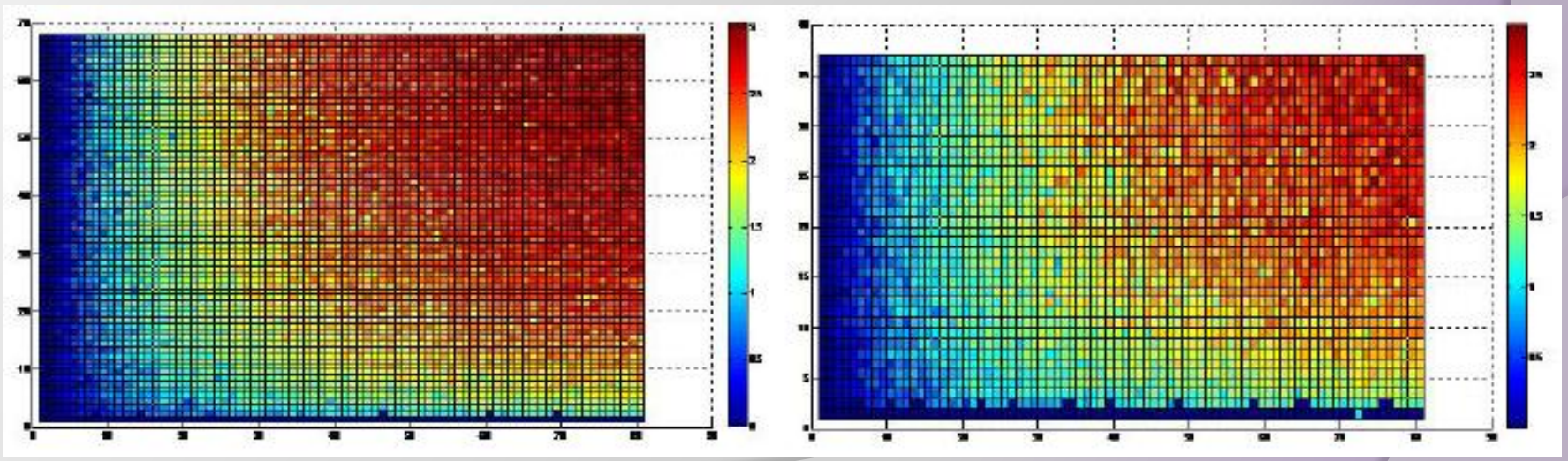
- The matrix 100×100 that contains 3 biclusters:

	Plaid	C&C	Spect	SAMBA	CBA	Theor
N of bic	6	3	2	5	3	3
no overl	3	3	2	3	3	3
the best	3	3	2	2	3	3
MSR1	0	0	0	-	0	0
MSR2	0	0	-	0.005	0	0
MSR3	0	0	0.013	0.011	0	0
E1	2.5	2.5	2.5	-	2.5	2.5
E2	2.44	2.44	-	2.5	2.5	2.5
E3	2.5	2.5	2.15	2.5	2.5	2.5
dim1	40×40	37×39	40×40	-	40×40	40×40
dim2	28×30	30×30	-	16×11	41×41	41×41
dim3	24×21	40×40	42×22	6×3	40×40	40×40

Results

The Gastric Cancer data

- 31 normal tissues
- 38 tumoral tissues:
  - 19 MSS
  - 19 MSI
- 82 genes



Normal/tumoral

N_g	5	10	11	12	15	20	25	30	35	40	45	55	60	65
Err	0.6	0.71	0.75	0.8	0.88	0.98	1.1	1.2	1.28	1.4	1.53	1.8	2.0	2.2
Er1	0.14	0.36	0.39	0.49	0.48	0.78	0.77	0.91	0.85	1.02	1.17	1.45	1.68	1.84
Er2	0.43	0.57	0.58	0.69	0.72	0.83	1.01	1.07	1.05	1.17	1.34	1.64	1.87	1.93
Er3	0.59	0.7	0.75	0.79	0.87	0.97	1.09	1.19	1.15	1.38	1.42	1.78	1.93	2.12
Msr1	0.01	0.02	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.06	0.07	0.08	0.1
Msr2	0.02	0.03	0.03	0.03	0.03	0.04	0.05	0.05	0.05	0.06	0.07	0.08	0.09	0.12
Msr3	0.03	0.03	0.03	0.03	0.04	0.04	0.05	0.06	0.06	0.07	0.08	0.09	0.12	0.12
S_n1	0.83	1.15	0.95	1.07	1.07	1.03	0.99	0.95	0.99	0.99	0.82	0.91	0.91	1.07
S_n2	1.25	1.41	1.38	1.43	1.4	1.22	1.28	1.06	1.13	1.13	0.99	1.17	1.37	1.24
S_n3	1.6	1.73	1.73	1.73	1.65	1.57	1.57	1.32	1.32	1.24	1.24	1.27	1.57	1.46
P_n1	0.79	0.12	0.43	0.25	0.25	0.25	0.43	0.43	0.43	0.43	0.79	0.62	0.62	0.25
P_n2	0.05	0.00	0.00	0.00	0.00	0.05	0.02	0.25	0.12	0.12	0.43	0.12	0.00	0.05
P_n3	1^{-4}	1^{-6}	1^{-6}	1^{-6}	1^{-6}	1^{-6}	1^{-6}	0.02	0.02	0.05	0.05	0.05	1^{-4}	7^{-4}
S_t1	0.4	0.4	0.4	0.4	0.47	0.54	0.54	0.74	0.74	0.81	0.81	0.78	0.54	0.63
S_t2	0.76	0.67	0.69	0.65	0.67	0.82	0.77	0.95	0.89	0.89	1.01	0.87	0.7	0.81
S_t3	0.92	0.87	1.03	0.94	0.94	0.97	1.01	1.03	1.01	1.01	1.14	1.07	1.07	0.94
P_t1	1	1	1	1	1	0.99	0.99	0.95	0.95	0.88	0.88	0.88	0.99	0.99
P_t2	0.95	0.99	0.98	0.98	0.98	0.88	0.95	0.57	0.75	0.75	0.38	0.75	0.98	0.88
P_t3	0.57	0.75	0.38	0.57	0.57	0.57	0.38	0.38	0.38	0.38	0.09	0.21	0.21	0.38
Rel	$\frac{20}{9}$	$\frac{22}{5}$	$\frac{17}{10}$	$\frac{21}{6}$	$\frac{15}{12}$	$\frac{16}{11}$	$\frac{14}{18}$	$\frac{14}{13}$	$\frac{14}{13}$	$\frac{14}{13}$	$\frac{13}{14}$	$\frac{15}{12}$	$\frac{18}{9}$	$\frac{15}{12}$
N	187	89	163	24	168	35	59	25	98	54	36	7	97	44

Mss/Msi

N_g	10	15	20	25	30	35	40	45	50	55	60	65	70	19t70g
Err	0.85	0.95	1.05	1.15	1.21	1.35	1.45	1.55	1.65	1.95	2.1	2.35	2.45	2.6
Er1	0.37	0.57	0.79	0.75	0.73	0.98	0.82	1.05	1	1.09	1.18	1.27	1.61	1.48
Er2	0.58	0.83	0.93	0.93	0.86	1.04	1.14	1.16	1.2	1.72	1.62	1.75	1.65	2.05
Er3	0.84	0.91	1.04	1.12	0.96	1.14	1.44	1.55	1.52	1.92	2.05	2.33	1.67	2.53
Msr1	0.02	0.04	0.05	0.05	0.06	0.06	0.06	0.07	0.07	0.08	0.09	0.11	0.12	0.13
Msr2	0.04	0.05	0.05	0.06	0.06	0.06	0.07	0.08	0.08	0.09	0.11	0.12	0.13	0.14
Msr3	0.05	0.05	0.05	0.06	0.06	0.07	0.08	0.08	0.09	0.12	0.13	0.14	0.13	0.14
S_s1	0.47	0.59	0.47	0.47	0.35	0.82	0.59	0.82	0.82	0.71	0.94	0.82	1.38	1.06
S_s2	0.92	0.75	0.75	0.64	0.52	0.96	0.86	1.01	1.08	1.02	1.27	1.17	1.46	1.49
S_s3	1.29	0.94	0.94	1.06	0.71	1.06	1.18	1.18	1.29	1.53	1.53	1.53	1.5	1.58
P_s1	0.99	0.99	0.99	0.99	1	0.74	0.99	0.74	0.74	0.9	0.74	0.74	0.02	0.5
P_s2	0.74	0.9	0.9	0.98	0.99	0.5	0.74	0.5	0.26	0.5	0.09	0.26	0.00	0.00
P_s3	0.09	0.74	0.74	0.26	0.9	0.26	0.09	0.09	0.09	0.00	0.00	0.00	0.00	0.00
S_i1	0.71	1.06	1.06	0.94	1.29	0.94	0.82	0.82	0.71	0.47	0.47	0.47	0.5	0.42
S_i2	1.08	1.25	1.25	1.36	1.48	1.04	1.14	0.99	0.92	0.97	0.73	0.83	0.54	0.52
S_i3	1.53	1.41	1.53	1.53	1.65	1.18	1.41	1.18	1.17	1.29	1.05	1.17	0.62	0.94
P_i1	0.9	0.26	0.26	0.74	0.09	0.74	0.74	0.74	0.9	0.99	0.99	0.99	0.02	0.5
P_i2	0.26	0.09	0.09	0.02	0.00	0.5	0.26	0.5	0.74	0.5	0.9	0.74	0.00	0.00
P_i3	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.09	0.25	0.09	0.26	0.26	0.00	0.00
Rel	$\frac{7}{11}$	$\frac{8}{9}$	$\frac{6}{12}$	$\frac{4}{13}$	$\frac{4}{14}$	$\frac{8}{9}$	$\frac{7}{10}$	$\frac{10}{9}$	$\frac{9}{10}$	$\frac{11}{8}$	$\frac{12}{5}$	$\frac{10}{7}$	$\frac{13}{5}$	$\frac{15}{5}$
N	34	10	38	77	29	11	74	83	79	146	89	124	3	21

Conclusions

- As shown by the experiments, Combinatorial algorithm gives always better and more accurate results than the other algorithms, because it reaches the maximal precision in the data sets analysis.
- In every experiment we a priori decided the maximal error and the minimal dimension of the desired biclusters.

Acknowledgments

- I thank my co-workers and co-authors:

Prof. Roberto Tagliaferri,
PhD Francesco Napolitano
Prof.Giancarlo Raiconi
(Dept. of Mathematics and Informatics, University of Salerno)

PhD.Roberto Amato
Prof. Gennaro Miele
Prof. Sergio Cocozza
(Dipartimento di Scienze Fisiche, Università degli Studi di Napoli "Federico I", Napoli, Italy)

- This work is partially supported by Istituto Nazionale di Alta Matematica Francesco Severi (INdAM) with the scholarship N U 2007/000458 07/09/2007

Thank you!!!