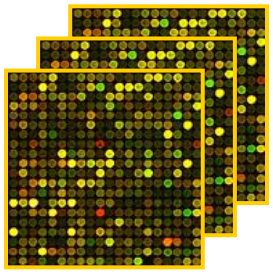
Gene-Markers Representation for Microarray Data Integration

Data Base and Data Mining Group of Politecnico di Torino

Elena Baralis, Elisa Ficarra, **Alessandro Fiori**, Enrico Macii
Department of Control and Computer Engineering
Politecnico di Torino (Italy)

Boston, 14-17 October 2007

Introduction



- Goals
  - Integrate heterogeneous datasets
  - Build a system independent to a-priori knowledge
  - New representation of data and synergies among genes
- Open problems of integration
  - Scaling issues
  - Error bias
  - Experimental condition
  - Different technology or protocol

Framework purpose

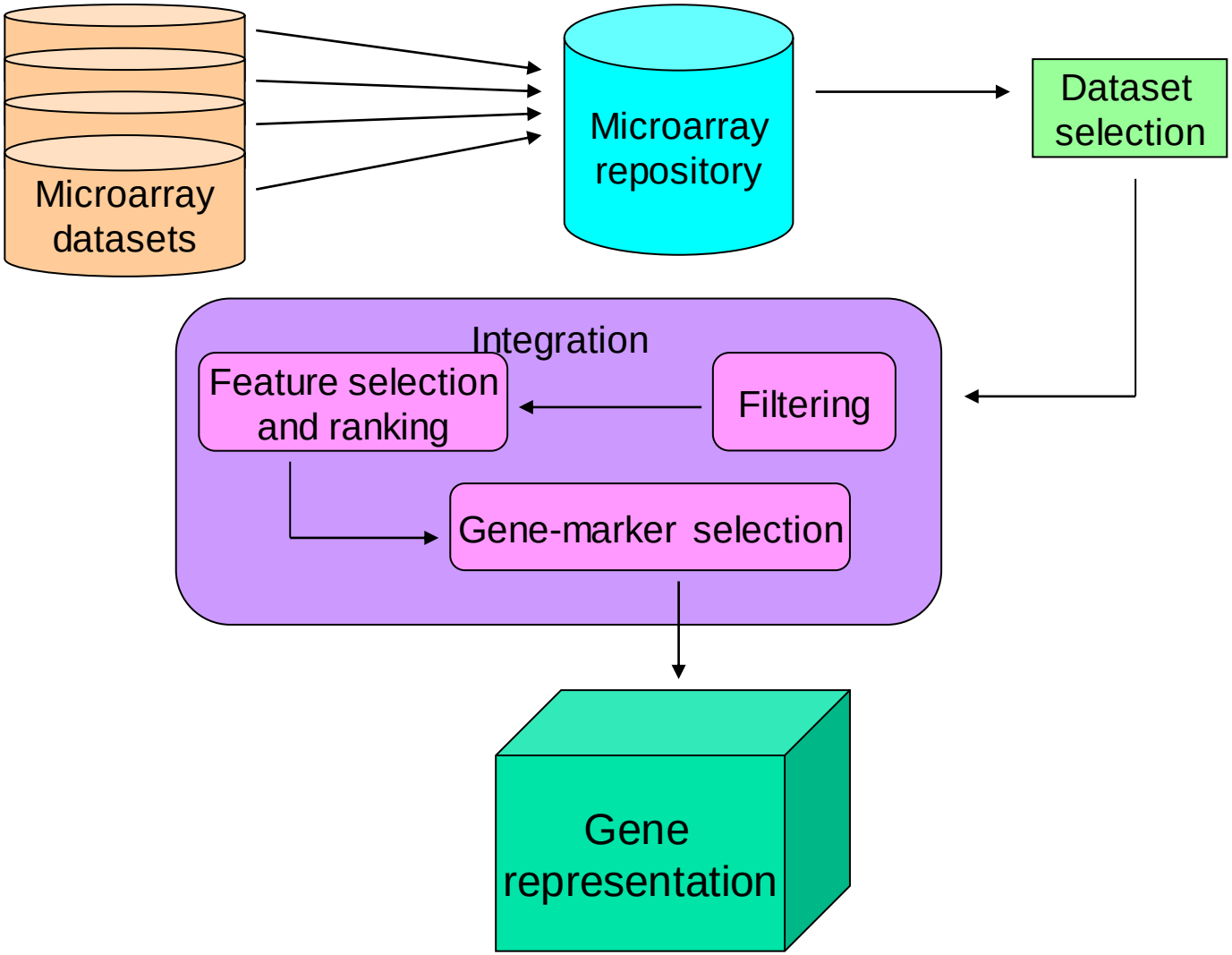
- Representation of synergies between genes
- Gene-markers selection
  - Common to all the datasets
  - Base of the new space representation
- Gene-markers characteristics
  - Common to all the datasets
  - "Highly" representative for each dataset
  - No outliers
  - Independency

Innovation

- Independence of a-priori knowledge
  - Biological information
  - Data distribution
- Fully automated
- Applicable to problems
  - With no knowledge
  - Few weak hypotheses

Kangl and al., "Integrating heterogeneous microarray data sources using correlation signatures," *Data Integration in the Life Sciences*, vol. 3615/2005, pp. 105–120, 2006

Framework


Microarray datasets
Microarray repository
Dataset selection
Integration
Feature selection and ranking
Filtering
Gene-marker selection
Gene representation

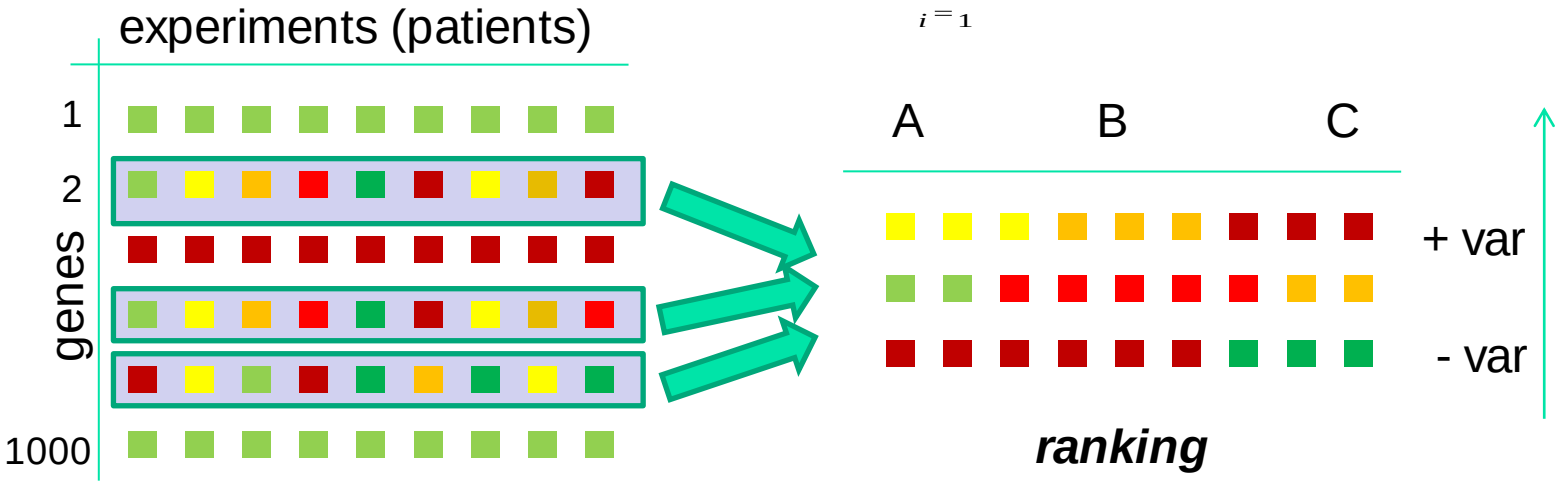
Filtering

- Remove flat genes
- Variance of a gene

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2}{N-1}$$

- Filter

$$\max \frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2} \cong \alpha \qquad \alpha = 0.9 \text{ (by default)}$$



Feature selection

- Eliminate less relevant features in K gene set
- Different techniques
  - Supervised
  - Unsupervised
- ANOVA in version 1.0 (Jeffery 2006)
  - Rank based on F-value
  - Binary and multi-class scenarios

Jeffery and al., "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data", *BMC Bioinformatics*, vol. 7, no. 1, p. 359, July 2006

Gene-marker selection

- Merge ranks

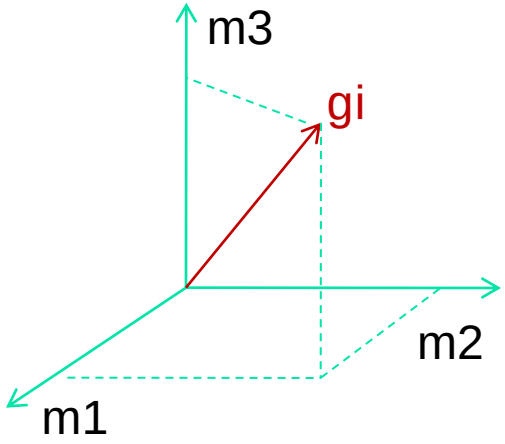
$$rank_i = \sum_{j=1}^{M} rank_{ij}$$

- Extraction of gene-markers
  - Gene with highest score removed from global rank and inserted in the gene-markers set
  - Pruning of the genes with average quadratic correlation with the selected gene-markers higher than a threshold (i.e. 20%)
  - Repeating procedure until L gene-markers are selected

Space transformation

- New representation
  - Matrix G, N_{tot}xL dimensions

$$g_{ij} = dist\left(\vec{g}_i, \vec{m}_j\right)$$



- g_{ij} elements measure distance
  - Cosine correlation
  - Pearson correlation
  - Euclidean
  - Manhattan

Experimental design

- Entropy evaluation
  - Evaluation of noise reduction
- Stability of the model
  - Conservative propriety with respect to biological information

Datasets	Patients	Genes	Classes
DLBCL	77	5469	2
Leukemia1	72	5327	3
Brain1	90	5921	5
Tumors9	60	5727	9

Entropy evaluation

- Description of data distribution
  - High value implies uniform distribution
- Entropy distance based (Manoranjan 2002)

$$E = -\frac{1}{N}\sum_{X_i}\sum_{X_j}\left[D_{ij}\log_2 D_{ij} + (1-D_{ij})\log_2(1-D_{ij})\right]$$

- Tests
  - Raw vs. transformed data
  - Impact of filtering phase

Manoranjan and al., "Feature selection for clustering - a filter solution", *IEEE International Conference on Data Mining (ICDM)*, pp. 115-122, 2002

Entropy on transformation

Datasets	Cosine correlation		Pearson correlation	
	Raw	*Transformed*	*Raw*	*Transformed*
DLBCL	0.750	0.127	0.947	0.639
Leukemia1	0.722	0.245	0.940	0.707
Brain1	0.813	0.305	0.943	0.664
Tumors9	0.813	0.292	0.976	0.762

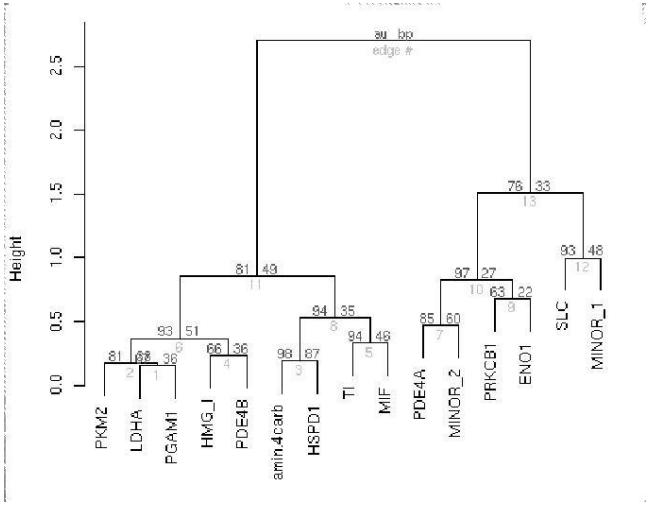
Impact of filtering phase

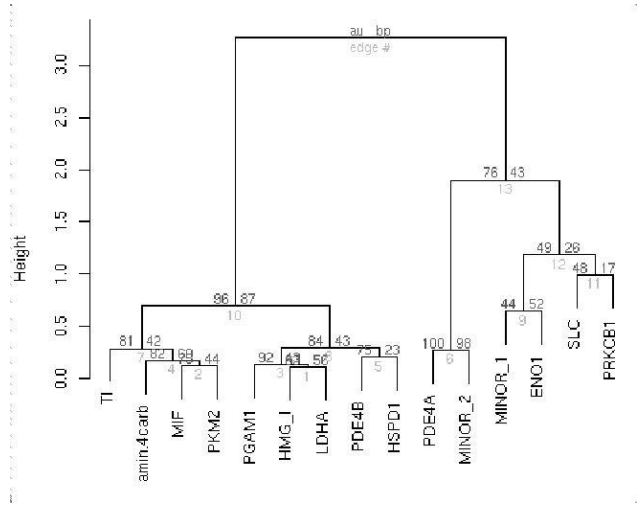
Datasets	Raw data	Data transformed without filter	Data transformed with filter
DLBCL	0.750	0.270	0.127
Leukemia1	0.722	0.296	0.245
Brain1	0.813	0.299	0.305
Tumors9	0.813	0.371	0.292

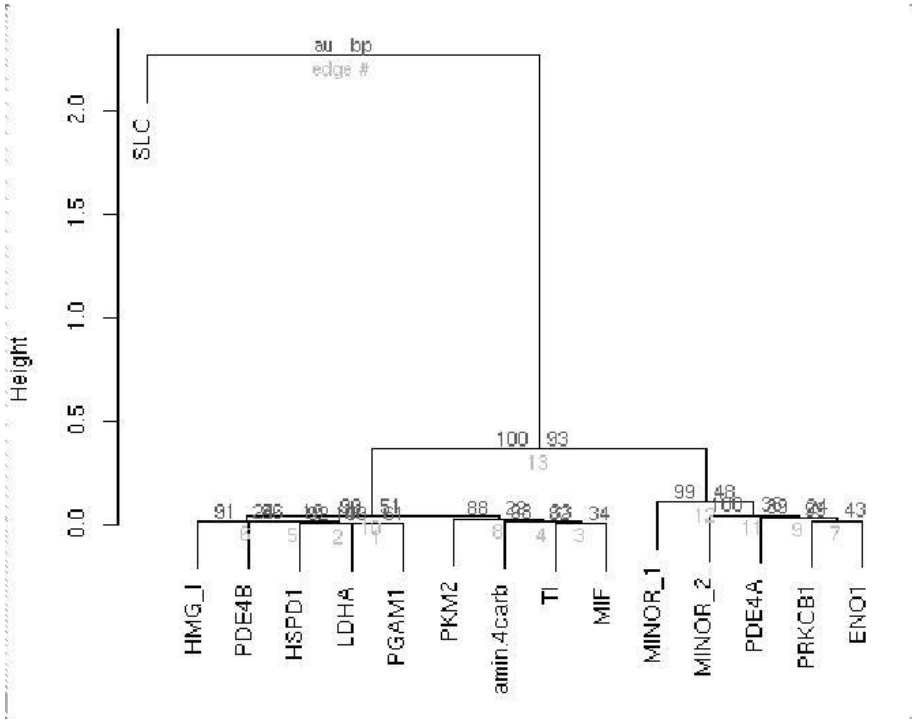
Subset genes

Reference	Description
TI	Triosephosphate Isomerase
HMG I	High mobility group protein gene exons 1-8
MIF	Macrophage migration inhibitory factor gene
PDE4B	Phosphodiesterase 4B, cAMP - specific (dunce (Drosophila) - homolog phosphodiesterase E4)
LDHA	Lactate dehydrogenase A
PRKCB1	clones lambda - hPKC - beta [15, 802]) protein kinase C - beta - 1
MINOR_1	Mitogen induced nuclear orphan receptor (MINOR_1) mRNA
PDE4A	Phosphodiesterase 4A, cAMP - specific (dunce (Drosophila) - homolog phosphodiesterase E2)
ENO1	ENO1 Enolase 1 (alpha)
MINOR_2	Mitogen induced nuclear orphan receptor (MINOR_2) mRNA
PKM2	Pyruvate kinase, muscle
amin4carb	5-aminoimidazole-4-carboxamide-1-beta-Dribonucleotide transformylase/inosinicase
SLC	SLC
HSPD1	Heat shock 60 kD protein 1
PGAM1	Phosphoglycerate mutase 1 (brain)

Stability of the model







Conclusion

- New method:
  - Based on dataset characteristics
  - Automatic selection of gene-markers based on microarray data
  - Independent on a-priori or pregressive knowledge
  - Definition of a new space representation
- Results
  - Reduction of entropy
  - Biological information content conservation
  - Improvement of knowledge about biological links between genes
- Future work:
  - Implementation of unsupervised and supervised feature selection methods
  - Integration of different kinds of information (ontologies)

Thanks for the attention!