Gene-Markers Representation for Microarray Data Integration

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

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
- Filter

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

\[
\max \frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2} \equiv \alpha
\]

\[\alpha = 0.9 \text{ (by default)}\]
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

Gene-marker selection

- **Merge ranks**

\[
\text{rank}_i = \sum_{j=1}^{M} \text{rank}_{ij}
\]

- **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 $G$, $N_{tot} \times L$ dimensions
    
    $$g_{ij} = \text{dist} \quad g_i, m_j$$

- $g_{ij}$ 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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Patients</th>
<th>Genes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>77</td>
<td>5469</td>
<td>2</td>
</tr>
<tr>
<td>Leukemia1</td>
<td>72</td>
<td>5327</td>
<td>3</td>
</tr>
<tr>
<td>Brain1</td>
<td>90</td>
<td>5921</td>
<td>5</td>
</tr>
<tr>
<td>Tumors9</td>
<td>60</td>
<td>5727</td>
<td>9</td>
</tr>
</tbody>
</table>
Entropy evaluation

- Description of data distribution
  - High value implies uniform distribution
- Entropy distance based (Manoranjan 2002)
  \[
  E = - \frac{1}{N} \sum_{x_i} \sum_{x_j} D_{ij} \log_2 D_{ij} + \sum_{x_i} \sum_{x_j} D_{ij} \log_2 D_{ij}
  \]
- 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

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Cosine correlation</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw</td>
<td>Transformed</td>
</tr>
<tr>
<td>DLBCL</td>
<td>0.750</td>
<td>0.127</td>
</tr>
<tr>
<td>Leukemia1</td>
<td>0.722</td>
<td>0.245</td>
</tr>
<tr>
<td>Brain1</td>
<td>0.813</td>
<td>0.305</td>
</tr>
<tr>
<td>Tumors9</td>
<td>0.813</td>
<td>0.292</td>
</tr>
</tbody>
</table>
## Impact of filtering phase

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Raw data</th>
<th>Data transformed without filter</th>
<th>Data transformed with filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLBCL</td>
<td>0.750</td>
<td>0.270</td>
<td>0.127</td>
</tr>
<tr>
<td>Leukemia1</td>
<td>0.722</td>
<td>0.296</td>
<td>0.245</td>
</tr>
<tr>
<td>Brain1</td>
<td>0.813</td>
<td>0.299</td>
<td>0.305</td>
</tr>
<tr>
<td>Tumors9</td>
<td>0.813</td>
<td>0.371</td>
<td>0.292</td>
</tr>
</tbody>
</table>
## Subset genes

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