Matrix tri-factorization for miRNA-gene association discovery in acute myeloid leukemia

Simone Marini, Andrea Demartini, Francesca Vitali, and Riccardo Bellazzi
AML and miRNA

**Acute myeloid leukemia (AML)**
- malignant disorder affecting myeloid hematopoietic cells

**Micro RNA (miRNA)**
- small non-coding RNAs (22–24 nucleotides)
- regulate mRNA
- in cancer, miRNA expression is altered
- miRNA → therapeutic applications for AML
Our goal

Find **new associations**

between

**miRNA and gene expressions**

in **AML** patients
Our goal

Find new associations between miRNA and gene expressions in AML patients through matrix tri-factorization
Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes

Moritz Gerstung, Andrea Pellegrati, Luca Malcovati, Aristoteles Giagounidis, Matteo G Della Porta, Martin Jädersten, Hamid Dolatshad, Amit Verma, Nicholas C. P. Cross, Paresh Vyas, Sally Killick, Eva Hellström-Lindberg, Mario Cazzola, Elli Papaemmanuil, Peter J. Campbell & Jacqueline Boulwood

Affiliations | Contributions | Corresponding authors

*Nature Communications* 6, Article number: 5901 | doi:10.1038/ncomms6901
Received 18 August 2014 | Accepted 18 November 2014 | Published 09 January 2015

PCA + logistic regression
Validated on TCGA-AML data
TCGA-AML

Publicly available

Data to be extracted and interpreted

Several data sources
TCGA-AML

Gene expressions
~20,000
Patients
198
miRNAs
705
Filtering data through knowledgebases

1177 Genes

→ 552 AML-related from COSMIC (Catalogue of Somatic Mutations in Cancer)

→ 991 AML-related from the Atlas of Genetics and Cytogenetics in Oncology and Haematology
Integrating data through knowledgebases

750 miRNAs
→ 705 from TCGA-AML
→ 45 (interacting with selected genes) from miRTarBase
Data Fusion by Matrix Factorization

Marinka Žitnik and Blaž Zupan

Abstract—For most problems in science and engineering we can obtain data sets that describe the observed system from various perspectives and record the behavior of its individual components. Heterogeneous data sets can be collectively mined by data fusion. Fusion can focus on a specific target relation and exploit directly associated data together with contextual data and data about system’s constraints. In the paper we describe a data fusion approach with penalized matrix tri-factorization (DFMF) that simultaneously factorizes data matrices to reveal hidden associations. The approach can directly consider any data that can be expressed in a matrix, including those from feature-based representations, ontologies, associations and networks. We demonstrate the utility of DFMF for gene function prediction task with eleven different data sources and for prediction of pharmacologic actions by fusing six data sources. Our data fusion algorithm compares favorably to alternative data integration approaches and achieves higher accuracy than can be obtained from any single data source alone.
Tri-factorization

Data Fusion by Matrix Factorization

Marinka Žitnik and Blaž Zupan

Abstract—For most problems in science and engineering we can obtain data sets that describe the observed system from various perspectives and record the behavior of its individual components. Heterogeneous data sets can be collectively mined by data fusion. Fusion can focus on a specific target relation and exploit directly associated data together with contextual data and data about system’s constraints. In the paper we describe a data fusion approach with penalized matrix tri-factorization (DFMF) that simultaneously factorizes data matrices to reveal hidden associations. The approach can directly consider any data that can be expressed in a matrix, including those from feature-based representations, ontologies, associations and networks. We demonstrate the utility of DFMF for gene function prediction task with eleven different data sources and for prediction of pharmacologic actions by fusing six data sources. Our data fusion algorithm compares favorably to alternative data integration approaches and achieves higher accuracy than can be obtained from any single data source alone.

Semi-Supervised Clustering via Matrix Factorization*

*The work of Fei Wang and Changshui Zhang is funded by China Natural Science Foundation No.60675009. The work of Tao Li is partially supported by NSF IIS-0546280.

Abstract

The recent years have witnessed a surge of interests of semi-supervised clustering methods, which aim to cluster the data set under the guidance of some supervisory information. Usually those supervisory information takes the form of pairwise constraints that indicate the similarity/dissimilarity between the two points. In this paper, we propose a novel matrix factorization based approach for semi-supervised clustering. In addition, we extend our algorithm to co-cluster the data sets of different types with constraints. Finally the experiments on UCI data sets and real world Bulletin Board Systems (BBS) data sets show the superiority of our proposed method.
The plan sounds simple

Replicate the algorithm

Extract TCGA-AML data

Apply it to TCGA-AML data

Find putative new associations
The algorithm

Basic idea:

Exploit fusion of different data sources

**objects**
patients, genes, miRNAs

objects show **relations** (e.g. patient-gene expression)
The algorithm

Represent relations with a block matrix

one block is the target (associations)

$\Theta \rightarrow$ object(i,i) $\rightarrow$ diagonal

relation -1 $\rightarrow$ 1 no relation

$R \rightarrow$ object(i,j) $\rightarrow$ upper/lower part

relation 1 $\rightarrow$ 0 no relation
# The algorithm

<table>
<thead>
<tr>
<th>Patients</th>
<th>Gene expression</th>
<th>miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>Gene Expression TCGA-AML</td>
<td>miRNA Expression TCGA-AML</td>
</tr>
<tr>
<td>Gene expression</td>
<td>gene-gene Interaction BioGrid</td>
<td>miRNA-gene interaction MiRTarBase</td>
</tr>
<tr>
<td>miRNA</td>
<td>miRNA Expression TCGA-AML</td>
<td>miRNA-gene interaction MiRTarBase</td>
</tr>
</tbody>
</table>
The algorithm

Exploit relations from knowledgebases
   BioGrid
   MiRTarBase

Exploit relations from TCGA-AML data
The algorithm

Patients 

Gene expression

miRNA

Patients

Gene expression

miRNA

Gene Expression TCGA-AML

Gene-gene Interaction BioGrid

miRNA-gene interaction MiRTarBase

Gene Expression TCGA-AML

Gene-gene Interaction BioGrid

miRNA-gene interaction MiRTarBase

Gene Expression TCGA-AML

Gene-gene Interaction BioGrid

miRNA-gene interaction MiRTarBase

Gene Expression TCGA-AML

Gene-gene Interaction BioGrid

miRNA-gene interaction MiRTarBase
The algorithm

Patients
Gene expression
miRNA

Patients
Gene expression
miRNA

Gene Expression TCGA-AML
Gene-gene Interaction BioGrid
miRNA-gene interaction MiRTarBase

Gene Expression TCGA-AML
Gene-gene Interaction BioGrid
miRNA-gene interaction MiRTarBase

Gene Expression TCGA-AML
Gene-gene Interaction BioGrid
miRNA-gene interaction MiRTarBase

Gene Expression TCGA-AML
Gene-gene Interaction BioGrid
miRNA-gene interaction MiRTarBase
The algorithm

Patients  | Gene expression  | miRNA

Patients  | gene expression  | miRNA expression  
TCGA-AML  | TCGA-AML  

Gene expression  | gene-gene interaction  | miRNA-gene interaction  
BioGrid  | MiRTarBase  

miRNA  | miRNA-gene interaction  | ?  
Expression  
TCGA-AML  | MiRTarBase  

Patients

Gene expression

miRNA

0

Gene Expression TCGA-AML

miRNA-gene interaction MiRTarBase

miRNA Expression TCGA-AML

0

miRNA-gene interaction MiRTarBase

miRNA Expression TCGA-AML

0

Gene Expression TCGA-AML

Patients

Gene expression

miRNA
<table>
<thead>
<tr>
<th></th>
<th>Patients</th>
<th>Gene expression</th>
<th>miRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>-1, 1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gene expression</td>
<td>0</td>
<td>gene-gene</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interaction</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BioGrid</td>
<td></td>
</tr>
<tr>
<td>miRNA</td>
<td>0</td>
<td>0</td>
<td>-1, 1</td>
</tr>
</tbody>
</table>
The algorithm

Decompose block matrix into in 3 smaller matrices
PCA to determine size

Convergence granted

\[ J = \sum_{i,j} \left\| R_{i,j} - \hat{R}_{i,j} \right\|^2 + tr(\hat{\Theta}) \]
The algorithm

Convergence → arbitrary threshold
The algorithm

Convergence → arbitrary threshold

Random initialization → consensus
The algorithm

Convergence $\rightarrow$ arbitrary threshold

Random initialization $\rightarrow$ consensus

And after convergence...?
Results

How we do interpret the results?

- Stop criteria $10^{-5}$
- Presence of association > 5/15 runs
- Association > 0.25
Results

How we do interpret the results?

- Stop criteria $10^{-5}$
- Presence of association $> 5/15$ runs
- Association $> 0.25$

Twelve new putative miRNA-gene associations found
## Validation

<table>
<thead>
<tr>
<th>Gene</th>
<th>miRNA</th>
<th># counts</th>
<th>Predicted by</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFLAR</td>
<td>hsa-miR-101-1</td>
<td>9</td>
<td>microRNA, ComiR, TargetScanHuman</td>
</tr>
<tr>
<td>ERG</td>
<td>hsa-miR-10b</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>FAIM</td>
<td>hsa-miR-10b</td>
<td>9</td>
<td>ComiR</td>
</tr>
<tr>
<td>FSCN1</td>
<td>hsa-miR-10b</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>BCOR</td>
<td>hsa-miR-101-1</td>
<td>8</td>
<td>TargetScanHuman</td>
</tr>
<tr>
<td>BTK</td>
<td>hsa-miR-101-1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>FGFR1</td>
<td>hsa-miR-10b</td>
<td>8</td>
<td>ComiR, TargetScanHuman</td>
</tr>
<tr>
<td>IGF1R</td>
<td>hsa-miR-10b</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>LIMK1</td>
<td>hsa-miR-101-1</td>
<td>8</td>
<td>ComiR</td>
</tr>
<tr>
<td>MLLT4</td>
<td>hsa-miR-101-1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>PTPRK</td>
<td>hsa-miR-10b</td>
<td>8</td>
<td>microRNA, TargetScanHuman</td>
</tr>
<tr>
<td>PRDM16</td>
<td>hsa-miR-1233</td>
<td>6</td>
<td>TargetScanHuman</td>
</tr>
</tbody>
</table>
That's all folks

Thank you

Questions...?