Relational Descriptive Analysis of Gene Expression Data

Igor Trajkovski 1, Filip Zelezny 2, Nada Lavrac 1, Jakub Tolar 3

Jozef Stefan Institute, Ljubljana, Slovenia 1
Czech Technical University, Prague, Czech Republic 2
University of Minnesota, Minneapolis, USA 3

29 August 2006
Riva del Garda, Italy

Overview

- DNA Microarray Technology
- Classification of Microarray Data
- Relational Descriptive Analysis of Gene sets
- Results
- Conclusion and future work
Molecular Biology: The central dogma

DNA

Gene = DNA subsequence

Genes code for proteins

Gene expression
DNA piece transcribes to RNA
RNA translates into a protein

Proteins `do the job`
- enzymes
- building blocks
- ...

DNA Microarray Technology

It allows scientists:
- To measure the expression level, in parallel, of thousands of genes in a tissue sample
- It works on the principle of hybridization
  (the process of joining two complementary strands of DNA)
DNA Microarray Technology (2)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>D26528_at</td>
<td>193</td>
</tr>
<tr>
<td>D26561_cds1_at</td>
<td>70</td>
</tr>
<tr>
<td>D26561_cds2_at</td>
<td>144</td>
</tr>
<tr>
<td>D26561_cds3_at</td>
<td>33</td>
</tr>
<tr>
<td>D26579_at</td>
<td>318</td>
</tr>
<tr>
<td>D26598_at</td>
<td>1764</td>
</tr>
<tr>
<td>D26599_at</td>
<td>1537</td>
</tr>
<tr>
<td>D26600_at</td>
<td>1204</td>
</tr>
<tr>
<td>D28114_at</td>
<td>707</td>
</tr>
</tbody>
</table>

Classification of patients with expression profiles

**Data:**
- Patient_X = (gene_1, gene_2, gene_3, ..., gene_N)
- Many genes (~ 10^4 - 10^5)
- Not so many patients (~ 10^1 - 10^2)

**Typical Questions:**
- Which disease has patient? (class A or class B)
- Which treatment should patient get? (Treatment A or B)
- Will patient develop side effects from drug X? (Yes or No)
**Standard Approach for classification**

Most learning algorithms look for non-linear combinations of features -- can easily find many spurious combinations given small # of records and large # of genes

1. Find genes that are differentially expressed between different conditions/phenotypes, e.g. two different tumor types (10 ~ 100)
2. Build a classifier taking into consideration genes from 1. (bayesian learners, SVM, decision trees, neural networks, ...)
3. Estimate the predictor accuracy (crossvalidation). (high dimensionality of the expression data causes overfitting problems)
4. Semantically interpret the results and found model. (our work comes here)

**Finding differentially expressed genes**

- You want to find genes that display a large difference in gene expression *between* groups and are homogeneous *within* groups
- Typically, you would use statistical tests (e.g. t-test) and calculate p-values (e.g. permutation test)
- p-values from these tests have to be corrected for multiple testing (e.g. Bonferroni correction)

The two sample t-statistic is used to test equality of the group means $\mu_1, \mu_2$.

$$T(g, c) = \frac{\mu_1(g) - \mu_2(g)}{\sqrt{\frac{\sigma_1(g)}{n_1} + \frac{\sigma_2(g)}{n_2}}}$$
Standard classification task

Decision trees | Artificial Neural Networks | Support Vector Machines

10-Fold Cross Validation

<table>
<thead>
<tr>
<th>Train</th>
<th>Train</th>
<th>Select</th>
<th>Train</th>
<th>Train</th>
</tr>
</thead>
</table>

Leave one out Cross-Validation

<table>
<thead>
<tr>
<th>Train</th>
<th>Train</th>
<th>Select</th>
<th>Train</th>
<th>Train</th>
</tr>
</thead>
</table>

Interpretation

<table>
<thead>
<tr>
<th>Gene</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene_1</td>
<td>score 1</td>
</tr>
<tr>
<td>gene_2</td>
<td>score 2</td>
</tr>
<tr>
<td>gene_3</td>
<td>score 3</td>
</tr>
<tr>
<td>gene_4</td>
<td>score 4</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
<tr>
<td>gene_100</td>
<td>score 100</td>
</tr>
<tr>
<td>gene_101</td>
<td>score 101</td>
</tr>
<tr>
<td>......</td>
<td>......</td>
</tr>
<tr>
<td>gene_99</td>
<td>score 9905</td>
</tr>
</tbody>
</table>

We what to describe the genes that are used by the classifier!

Selected genes have different influence on the classifier.

The weight of that influence can be extracted from the learned model (e.g. voting algorithm or SVM) or from the gene selection algorithm, in a form of a score, or weight.

Description of the selected genes, should prefer genes that have bigger weight.

The terms of the Gene Ontology were used as a vocabulary for the description of the genes.
Gene Ontology (GO)

- **GO** is a database of terms for genes
  - Function - What does the gene product do?
  - Process - Why does it perform these activities?
  - Component - Where does it act?
- Known genes are annotated to the terms (www.ncbi.nlm.nih.gov)
- Terms are connected as a directed acyclic graph (is_a, part_of)
- Levels represent specificity of the terms

Gene Ontologies (2)

- 10261 biological process
- 1677 cellular components
- 7375 molecular functions
Multi-Relational situation

Background Knowledge

Prolog facts:
- predicate(geneID, CONSTANT).
- interaction(geneID, geneID).
- component(2532,'GO:0016020').
- component(2532,'GO:0005886').
- component(2534,'GO:0008372').
- function(2534,'GO:0030554').
- function(2534,'GO:0005524').
- process(2534,'GO:0007243').
- interaction(2534,5155).
- interaction(2534,4803).

Basic, plus generalized background knowledge using GO

- zinc ion binding ->
- metal ion binding, ion binding, binding

fact(class, geneID, weight).
- fact('diffexp',64499, 5.434).
- fact('diffexp',2534, 4.423).
- fact('diffexp',5199, 4.234).
- fact('diffexp',1052, 2.990).
- fact('diffexp',6036, 2.500).

...
First order features

First order features with support > min_support

f(7,A):-function(A,'GO:0046872').
f(8,A):-function(A,'GO:0004871').
f(10,A):-process(A,'GO:0006952').
f(11,A):-process(A,'GO:0007165').
f(14,A):-process(A,'GO:0044267').
f(15,A):-process(A,'GO:0050874').
f(16,A):-component(A,'GO:0016021').
f(19,A):-component(A,'GO:0016020').
f(122,A):-interaction(A,B),function(B,'GO:0004872').
f(223,A):-interaction(A,B),function(B,'GO:0004871'),process(B,'GO:0009613').
f(224,A):-interaction(A,B),function(B,'GO:0016787'),component(B,'GO:0043231').

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
<th>f5</th>
<th>f6</th>
<th>...</th>
<th>...</th>
<th>fn</th>
</tr>
</thead>
<tbody>
<tr>
<td>g1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g8</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Relational Subgroup Discovery

- We don't look for single complex rule to describe A, but several rules that describe parts (subgroups) of A.

- We prefer rules that are accurate (cover only diff. exp. genes) and have big recall value (weight of covered genes is big).

- In our experiment we used the implementation of Propositionalization-Based Relational Subgroup Discovery algorithm, RSD (Lavrač, Žlezny 2002)

- RSD naturally uses gene weights in its procedure for repetitive subgroup generation, via its heuristic rule evaluation: weighted relative accuracy

RSD – grammar of the first order features

:-modeh(1, expr(+keygene)).
:-modeh(1, function(+keygene, -fnct)).
:-modeh(1, function(+gene, -fnct)).
:-modeh(1, instantiate(+fnct)).
:-modeh(1, process(+keygene, -prcs)).
:-modeh(1, process(+gene, -prcs)).
:-modeh(1, instantiate(+prcs)).

:-mode(1, component(+keygene, -cmpnt)).
:-mode(1, component(+gene, -cmpnt)).
:-mode(1, instantiate(+cmpnt)).
:-mode(1, pathway(+keygene, -pthw)).
:-mode(1, pathway(+gene, -pthw)).
:-mode(1, instantiate(+pthw)).
:-mode(1, interaction(+keygene, -gene)).
:-mode(1, instantiate(+gene)).

f(1,A): function(A,B), instantiate(B).  
f(7,A): pathway(A,B), instantiate(B).  
f(8,A): interaction(A,B), instantiate(B).  
f(9,A): interaction(A,B), function(B,C), instantiate(C).  
f(28,A): interaction(A,B), function(B,C), component(B,D), instantiate(C), instantiate(D).  
f(29,A): interaction(A,B), function(B,C), component(B,D), pathway(B,E), instantiate(C), instantiate(D), instantiate(E).  
f(30,A): interaction(A,B), function(B,C), pathway(B,D), instantiate(C), instantiate(D).  
f(47,A): interaction(A,B), process(B,C), instantiate(C).
## Data flow diagram

**Patients Microarray profiles**

- Background knowledge
  - Interesting genes
  - Random genes

- RSD
  - Medical knowledge

- Gene features

(ILP terms are, positive and negative examples)

---

## Experiments

- We apply the proposed methodology on three popular classification problems from gene expression data:
  - ALL vs. AML, 6817 genes, 73 labeled samples, 2 classes
  - Subtypes ALL, 22283 genes, 132 labeled samples, 6 classes
  - 14 cancers, 16063, 198 labeled samples, 14 classes

- For all three problems and all classes we selected the 50 most differentially expressed (highest t-score ranking) genes and the same number of randomly chosen non-differentially expressed genes.
Results – Discovered Subgroups

- Descriptions of differentially expressed genes for: Acute lymphoblastic leukemia (ALL) vs. Acute myeloid leukemia (AML), (Golub ‘99)
  - [12, 0] interaction(A,B),process(B,'humoral immune response').
  - [11, 0] interaction(A,B),function(B,'peptidase activity’).
  - [8, 0] interaction(A,B),process(B,'proteolysis').
    interaction(A,B),function(B,'peptidase activity’).
  - [10, 0] interaction(A,B),process(B,'immune response'),
    component(B,'extracellular space').
  - [8, 1] function(A,'signal transducer activity').

Results – Discovered Subgroups (2)

- Subtypes ALL, 6 classes + all other (Ross ‘03)
  - BCR subtype
    - [7, 0] interaction(A,B),function(B,'metal ion binding').
      component(A,'membrane').
      process(A,'cell adhesion').
    - [6, 0] interaction(A,B),function(B,'transmembrane receptor activity').
      function(A,'receptor activity').
    - [6, 0] interaction(A,B),function(B,'protein homodimerizat. activity').
      interaction(A,B),process(B,'response to stimulus').
### Results – 10-fold Crossvalidation

<table>
<thead>
<tr>
<th>Task</th>
<th>Data</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML-ALL diff. exp. g.</td>
<td>Train</td>
<td>100%</td>
<td>16%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>85%</td>
<td>13%</td>
</tr>
<tr>
<td>Subtypes of ALL diff. exp. g.</td>
<td>Train</td>
<td>95%</td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>78%</td>
<td>12%</td>
</tr>
<tr>
<td>14 types Cancers diff. exp. g.</td>
<td>Train</td>
<td>94%</td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>75%</td>
<td>12%</td>
</tr>
</tbody>
</table>

### Results – Influence of the BG Knowledge

<table>
<thead>
<tr>
<th>Task</th>
<th>None</th>
<th>- Interact.</th>
<th>- GO</th>
<th>- Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL-AML diff. exp. g.</td>
<td>85(±6)%</td>
<td>44(±12)%</td>
<td>72(±13)%</td>
<td>75(±8)%</td>
</tr>
<tr>
<td>Subtypes of ALL diff. exp. g.</td>
<td>78(±10)%</td>
<td>52(±13)%</td>
<td>74(±16)%</td>
<td>71(±12)%</td>
</tr>
<tr>
<td>14 types Cancers diff. exp. g.</td>
<td>75(±10)%</td>
<td>45(±16)%</td>
<td>56(±14)%</td>
<td>73(±14)%</td>
</tr>
</tbody>
</table>
Related Work

- There are several approaches for descriptive analysis of gene expression data: Onto-Expres, GOstat, GoMiner, FunSpec, FatiGO, GO::TermFinder.
- All of them calculate the significance of single GO term.
- They that don’t consider the weight of importance and don’t use the interaction information.
- In our relational approach we naturally use gene interaction information, and we use significantly richer language for description of the genes.

Conclusion

- Preliminary results show that our methodology is capable of automatic extraction of meaningful medical knowledge.
- Extracted knowledge can be used for guiding the medical research, generating different interpretations of the learned model or for constructing complex gene features for building interpretable classifiers.
- We have high hopes on discovering novel, yet reliable medical knowledge from the relational combination of gene expression data with public gene annotation databases.