A GO based representation for prognosis and inference from microarray data

Shashank Srivastava, Snigdha Chaturvedi, Arnab Bhattacharya
Department of Computer Science and Engineering, IIT Kanpur

Introduction

Study of microarray data can assist in mining valuable biological information, building predictive models for pathological conditions, and unravelling latent correlations signifying biological pathways.

Current microarray analysis techniques focus on:
- Concise representations of microarrays through dimensionality reduction techniques (clustering , PCA etc.)
- Identifying differentially expressed genes
- And recently, evaluating expression patterns of gene groups instead of individual genes

However, none of these allow direct inference of relations between gene-expression values and higher level biological concepts.

We propose a concise representation of microarray data in terms of biological concepts, using a knowledge infusion from the structure of the Gene Ontology database. The method identifies important biological phenomena that differentiate two disease conditions, and the approach can be especially useful for biological conditions where underlying biological processes and critical pathways are still unknown.

Our approach

1. Create a functional profile for each gene :-

- Choose an abstraction level of the Gene Ontology
- Express each gene of the microarray in terms of the GO terms present at the chosen abstraction level
- Functional profiles of genes is vector containing the number of times the gene appears in the directed sub-graph of the gene ontology rooted at the corresponding GO term)

Genes	GO term 1	GO term 2	GO term 3	...	GO term N
Gene 1	2	1	0	...	0
Gene 2	0	0	1	...	1
. .	. .	. .	. .	...	. .
Gene M	0	1	0	...	0

2. Representation by gene functions :-

- Estimate the degree of up/down regulation of each gene as the difference between its expression level and the expression level for normal cells (average over all population classes)
- Using this value, each disease instance of the microarray is expressed in terms of the selected GO annotations

$$\overline{y_k} = \sum_{i=1}^{i=M} (x_{i_k} - \mu_i)\overline{f_i}$$

where, $\overline{y_k}$ is the functional representation for the k[th] disease sample

x_{i_k} is the gene expression value of the i[th] gene for the k[th] disease sample

μ_i is the average expression value of the gene i over all population samples

$\overline{f_i}$ is the functional profile for the i[th] gene.

3. Inferring Biological Concepts :-

- Identify the differentiating GO terms to characterise differences between the concerned classes
  - Rank by class separability criteria. eg. Student's t-test
  - Examination of class distribution of the GO terms
  - Class differences can yield insights into etiological differences between two physiologically indistinguishable pathological conditions.

Subsequent mining

- Often, a highly differentiating GO term might be too general to make any significant inference
- The steps described above can be iteratively called to mine deeper into the GO graph, starting at the relevant GO term

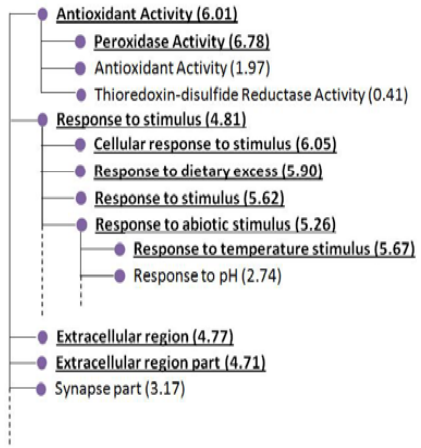
Datasets

- Follicular lymphoma (FL) vs Diffused Large B cell lymphoma (DLBCL) : grade 3 FL is physiologically indiscernible from FL.
- Two variants of leukemia : ALL vs AML

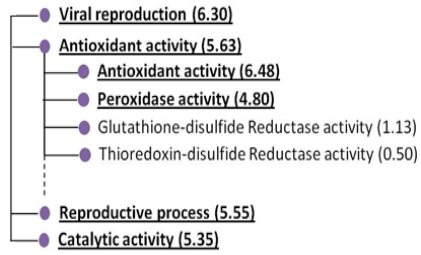
Dataset name	No. of genes	Disease class
Leukemia	5147	ALL (47) and AML (25)
DLBCL	7070	DLBCL (58) and FL (17)

Experiments

Most differentiating GO terms for Leukemia dataset :-

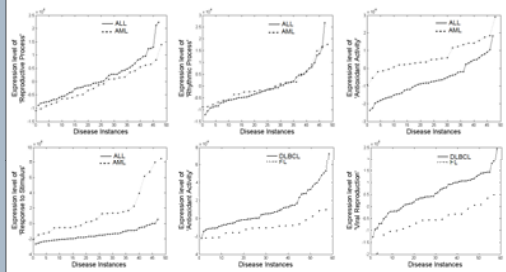


Most differentiating GO terms for DLBCL dataset :-



Class distributions

A small minority of GO terms are seen to be significantly different in the concerned classes .These GO terms could be expected to characterise differences between the concerned classes, and have major discriminatory potential.



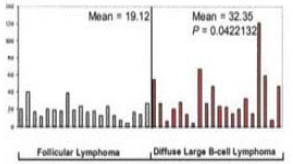
Results

On basis of differences in class distributions of GO terms, we predict DLBCL and FL to be distinct in terms of differences such as:
- Higher levels of viral reproduction in DLBCL
- Lower anti-oxidant activity in FL
- Lower peroxidase activity in FL

Similarly, we predict:
- Greater response to abiotic stimulus (specifically temperature stimulus) in AML
- Higher anti-oxidant activity in AML
- Correlation of AML with dietary excess

Whereas traditionally both ALL/AML and DLBCL/FL have been hard to differentiate, recent research by [1]-[5] support five (out of all six) claims above, whereas observation 3 about AML is suspected, especially for infants [6].

The current pointer indicates that this could be a promising viable direction for future biochemical research.



Expression levels of PLA_2 (antioxidant protein) in cases of FL and DLBCL [5]

Potential for predictive modeling (classification)

The proposed approach offers a quantitative representation of a biological condition in terms of gene function and concepts. This also suggests a roundabout to overcome the curse of dimensionality involved with microarray data, and could be very useful in building predictive statistical models. The proposed representation incorporates a knowledge infusion from the structure of the gene ontology, as distinct from models built purely on gene expression values.

Technique	Leukemia data	DLBCL data
Expression based	93.8%	88.3 %
Functional profile based	90.3 %	90.9 %
Our approach	94.4 %	84.4 %

References

1. Moreira et al. Dermatology Online Journal 14,7 (2008),17
2. Almasri et al. Saudi Med J 26,2 (2005), 251-5
3. Rodig et al. Clinical Cancer Research 12,23 (2006), 7174-79
4. Poongothai et al. Indian Journal of Human Genetics 10,1 (2004), 9-12
5. Johnson et al. PNAS USA 100,12 (2003) , 7259-64
6. Ross et al. PNAS USA 2000, 97(9), 4411-13