GENE SELECTION BASED ON CATEGORY DETECTION OF GENEONTOLOGY

Analytical objective:

The performance of classification methods is determined by the number of features, and can affect time complexity and convergence as well as detecting spurious patterns. The main idea of this work is to propose a gene selection methodology that retains most of the variability that is individually correlated with CFS diagnosis but at the same time, diminish the redundancy consequence of applying gene ranking. With this objective, a two stage procedure is defined which performs statistical gene ranking using the j5 test and selected genes are refined using clustering followed by class analysis of Gene Ontology to obtain meta-genes that can be associated to GO categories.

The main objective of this work is to present a gene selection methodology based on Gene Ontology information to discover latent components associated to biological process that, not only performs feature extraction and variable reduction but also generates a direct biological interpretation of such components allowing the validation against metabolic processes reported in the literature.

Additionally, the proposed technique is verified using the CAMDA microarray dataset, validating its predictive capability through cross validation. Based on the meta-genes obtained and its biological interpretation a biomarker for CFS is proposed.

Introduction

Chronic Fatigue Syndrome is a complex disease and have a multifactor etiology, the diagnostic is only made by the presence of a group of symptoms, among them, chronic fatigue and the presence of 4 or more symptoms that involve despair systems(1). Despite multiple studies, a biological marker for affected subjects has not been found yet. In fact, some researchers have claimed that it is not a single medical entity but a heterogeneous group of pathologies that converge into a physic pathologic common state characterized by chronic and severe fatigue(2).

With the information collected by Vernon's group at the CDC, differentially expressed genes between healthy and affected subjects were identified. It was determined that some of those genes were related to different signs and symptoms of the disease. Also, it was found that there are differences in expression levels in certain genes among groups induced by clinical characteristics (3,4,5,6).

However, the determination if the expression level of a gene or a group of genes allows to effectively diagnose the state of the disease is still a work in progress. Hongbo et al (7) worked in that direction. They use genomic and proteomic data, first selecting differentially expressed genes using a two stage approach, first with a ranking method based on the Kruskall Wallis statistical test and then taking genes with over represented categories in Gene Ontology. This methodology obtained a 72% diagnosis accuracy estimated using leave-one-out cross validation, compared with a 53% with just the hierarchical method, it proteomic data is included precision goes to 79%.

Methodology:

Data preprocessing
Assuming the empirical diagnosis as labels for the samples, only CFS Y NF diagnosed patients were selected, that is, individual with additional psychological and medical conditions were omitted in the training of the classification algorithm, therefore, only 79 patients, 40 presenting CFS and 39 control patients. For the microarray data normalization, the following steps were carried out: simple background subtraction, log-transformation, median shift, quantile normalization.

Information mining:
Through web service connection with NCBI xml format gene bank registers were obtained, from this files Gene Ontology categories were retrieved corresponding to gene annotation for about 14139 genes.

Gene Ranking
The gene ranking procedure was proposed as follows: the j5 statistic was used as an evaluator of the individual discriminative capability of each gene, applying this statistic to the expression profiles of the samples in the training set a raking was obtained, then a threshold was taken to obtain a preliminary gene set but with redundant information.

Clustering and meta-gene selection
To filter this preliminary gene set a hierarchical clustering was performed in the sample space using correlation distance and average link. Choosing different cut points in the tree, i.e. different values for k or the number of clusters, prototypes for each cluster were proposed. For each cluster, the most well represented categories were found and a new feature associated for each category was found taking the average of the gene expression value of the genes in such category. To restrict the number of categories and avoid ambiguity, only categories at a fixed level were taken. Experiments were performed taking two components of the ontology: “molecular function” and “biological process”.

Experimental design
To validate the discriminative capability of the algorithm, a standard experimental design was developed in which the data set was split in three groups using cross validation to
estimate the prediction of the method and bootstrap in the model selection phase. This work selected SVM as classification technique and was tested using different kernels as follows: Linear, polynomial and rbf kernel using different values for the box constraint parameter in order to adjust the complexity of the solution model

Using the results from cross validation the performance of the proposed method, its prediction capability (generalization) and meta-genes that improve the overall performance. Finally a model was trained using the whole data set and the best performance inducing meta-genes and a biological interpretation and contextualization was performed.

Results

After a large number of experiments the values of parameters that lead to better average performance were the following (accuracy 72%):

- Gene Ontology level= 3
- j5 threshold = 4.5
- Gene Ontology Categories p-value = 0.005
- SVM box Constraint = 0.35
- Ontology = biological process.

For the sake of validate the advantage of the two-stage gene selection method, a t-test was performed based on the average accuracy obtained in the cross-validation process for the proposed method against a simple gene ranking feature selection method, taking several runs. Finally we obtained p-value=0.008.

The Overrepresented Gene Ontology categories in the meta-genes selected from the best diagnosis prediction involve several molecular functions. Since the genes were previously grouped by level of expression to see if the groups were compatible with some grouping in gene ontology. As the best classification model 161 categories were found as overrepresented, most of them appearing just once (120), some twice (27) and 3, 4 and 5 times (9,3,1 and 1 respectively) . Categories that appeared more than three times are shown in table 1.

Category	Frequency
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	6
organelle organization and biogenesis	5
cell differentiation	4
positive regulation of cellular process	4
system development	4
cell death	3
cellular catabolic process	3
cellular macromolecule metabolic process	3
intracellular transport	3
organ development	3
positive regulation of biological process	3
regulation of cellular metabolic process	3
response to wounding	3
tissue development	3

Table 1. Categories of gene ontology that appeared more than three times.

The category that appears most times (nucleobase, nucleoside, nucleotide and nucleic acid metabolic process) (6 times) represents a category of molecular function that has been described in other works with different approaches and even with the same dataset, which lead us to relieve that this category is important and is consistently involved in the development of the disease.

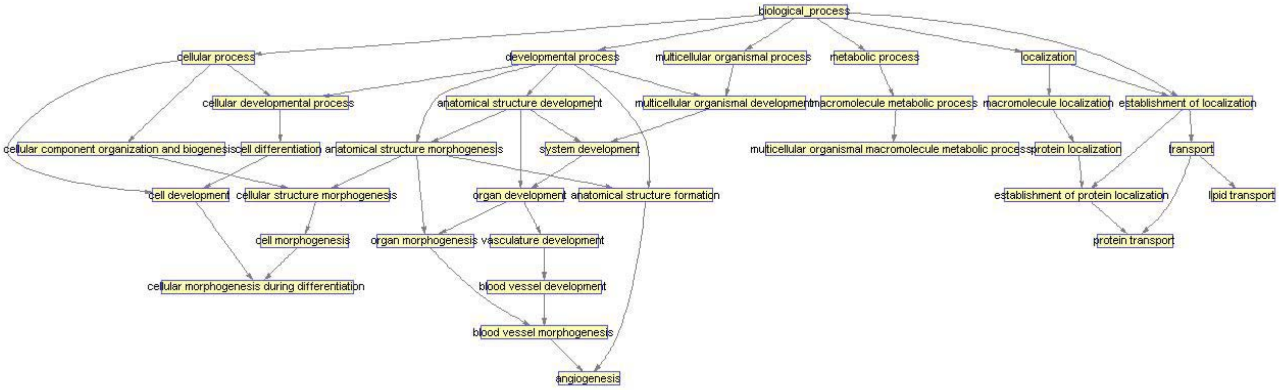
biological_process
cellular process
developmental process
multicellular organismal process
metabolic process
localization
cellular developmental process
anatomical structure development
multicellular organismal development
macromolecule metabolic process
macromolecule localization
establishment of localization
cellular component organization and biogenesis
cell differentiation
anatomical structure morphogenesis
system development
multicellular organismal macromolecule metabolic process
protein localization
transport
cell development
cellular structure morphogenesis
organ development
anatomical structure formation
establishment of protein localization
lipid transport
cell morphogenesis
organ morphogenesis
vasculature development
protein transport
cellular morphogenesis during differentiation
blood vessel development
blood vessel morphogenesis
angiogenesis

Figure 1. Sub-ontology extended over the overrepresented GO categories

Among other overrepresented categories (three times), some of them if as possible psychopathological networks of the disease such us cell differentiation, cell death, cellular catabolic process, cellular macromolecule metabolic process, regulation of cellular metabolic process, response to wounding, and other not as evident such us system development, intracellular transport, etc (figure 1). In the development of this work a careful analysis of the genes in each category will be made, the expression level of cases and controls and the possible implication of each of them

Algorithmic complexity:

Gene ranking using j5: O(n p)
Hierarchical clustering: O(n'^2 (lg n' + p))
Definition of a Meta-gene set: O(c + m n')
(where: n:=number of genes, c:=number of GO categories at the fixed level, m:=number of meta-genes, n':=number selected in the gene ranking stage, p:=sample size)

Discussion

Testing the classifier with subjects with chronic fatigue but with diagnosis exclusion most of them are classified in the CFS group (17 out of 19, table 2), this may indicate that people excluded from the diagnosis can present a similar physiopathologic process to those with CFS .

	Classification	
	CFS	NF
CFSMed	8	0
CFS/MDDm	8	2
CFSPsy	1	1

ISF	23	19
NF/MDDm	3	1
NFMed	4	3

Table 2. Clasificación de las categorías excluidas

However, the results are not consistent with the other excluded groups (ISF and healthy patients, table 2), this shows that our classifier has low specificity to separate people CFS from healthy people or with other diseases, probably because the group with CFS has a heterogeneous mix of pathologies with a similar clinical description, to clarify this in this work we will exclude subjects that do not allow to generalize the classifier in order to compare the result obtained in this with that reported in the present work.

References

1. Fukuda K, Straus SE, Hickie I, Sharpe MC, Dobbins JG, Komaroff A: The chronic fatigue syndrome: a comprehensive approach to its definition and study. Ann Intern Med 1994; 121:953–959
2. Afari N, Buchwald D, Chronic Fatigue Syndrome: A review; Am J Psyach, 2003; 160:221-236.
3. Aslakson E, Volmer-Conna U, White P. The validity of an empirical delineation of heterogeneity in chronic unexplained fatigue, Pharmacogenomics, 2006, 7, 365-373
4. Carmel L, Efroni S, White P, Aslakson E, Voller-Conna U, Rajeevan M, Gene expression profile of empirically delineated classes of unexplained chronic fatigue, Pharmacogenomics, 2006, 7, 374-880.
5. Fostel J, Boneva R, Lloyd A, Exploration of gene expression correlates of chronic unexplaines fatigue using factor analysis, Pharmacogenomics, 2006, 7, 881-890.
6. Wistler T, Taylor R, Craddock R, Broderick G, Klimas N, Unger E, Gene expression correlates of unexplained fatigue, Pharmacogenomics, 2006, 7, 395-405.
7. Hongbo Xie, Zoran Obradovic, Slobodan Vucetic Mining of Microarray, Proteomics, and Clinical Data for Improved Identification of Chronic Fatigue Syndrome. CAMDA 2006