Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Predictive modeling of therapy response in multiple sclerosis using gene expression data

[*]Sara Mostafavi, Sergio Baranzini, Jorge Oksernberg, and Parvin Mousavi

Abstract—**Transcription profiling studies reveal important insights in regards to molecular events that manifest in phenotypic outcomes such as response to drug therapy. Construction of computational models that accurately predict therapy response is only possible when precise data measurements, robust feature/gene selection, and advanced computational modeling methods are combined with stringent statistical validation and large scale verification of results. Due to the large number of gene expression measurements in transcriptional profiling studies, feature selection represents a bottleneck when constructing computational models. The degree of compromise between selection of the optimal feature set and computational efficiency results in many choices for candidate gene sets which leads to a wide range of classification accuracies. Furthermore, constructing a classification model using a larger-than-necessary gene set along with small number of samples may cause over-fitting the data, resulting in highly optimistic classification accuracies. In this study we present OSeMA, a fast, robust and accurate gene selection-classification framework which results in construction of classification models that are highly predictive of the rIFNB therapy response in multiple sclerosis patients. We assess the performance of OSeMA on held out test data. Additionally, we extensively evaluate OSeMA by comparing it to an exhaustive combinatorial gene selection-classification approach.**

I. INTRODUCTION

Multiple sclerosis (MS), the most common disease of central nervous system in young adults, affects over 2,500,000 people worldwide, causing devastating symptoms such as loss of speech, vision, and paralysis, yet remains without a cure [1]. Although IFNβ is widely used to reduce autoimmune attacks in relapsing-remitting multiple sclerosis (RRMS), patients exhibit considerable interindividual heterogeneity in their clinical response, adverse side effects, and almost half of them do not experience significant benefits [2]. Gene expression profiling prior to INFβ therapy in MS patients allows for identification of specific interindividual heterogeneity at the molecular level that causes differential response to therapy [3]. A classifier, constructed from the most informative gene signature, can be used to predict the therapy response of new patients. However, several challenges remain in accurately classifying gene expression data.

Although gene expression studies measure the expression values of hundreds to several thousands of sequences, the expression patterns of a few genes is often sufficient to account for the underlying grouping of the data [4]. In parallel, constructing a classification model using a larger-than-necessary gene set along with small number of samples causes over-fitting to the data, resulting in highly optimistic classification accuracies. Thus, constructing a generalizable and robust classification model requires identification of a minimal set of most informative genes. However, due to the small number of samples and a very large number of features (gene measurements) in expression datasets, gene selection poses a computational challenge.

In the machine learning literature two approaches to feature selection predominate: "filter" and "wrapper". In the filter approach, feature selection is usually performed by evaluating the 'goodness' of each gene independently of the classifier design. Most filter approaches are also univariate; the importance of each feature in distinguishing different classes is assessed individually. Signal-to-noise ratio, *t*-test, and the ratio of between-groups to within-groups sum of squares (BSS/WSS) are some examples of the filter approach for feature selection (see [5] for a review). In contrast, in the wrapper approach, the 'goodness' of a subset of features is directly evaluated by their corresponding classification performance. Wrapper approaches consider the combinatorial effects of a subset of features and are multivariate [6]. The major disadvantage of this latter approach is computational complexity, especially when the dataset is comprised of a large number of features and feature subsets are exhaustively compared. Essentially, the degree of compromise between selection of the optimal feature set and computational efficiency in different feature selection algorithms result in many choices for candidate gene sets which leads to a wide range of classification accuracies [5].

In this paper we present OSeMA (Orthogonal Search Model Analysis), a fast and robust gene selection-classification framework based on orthogonal search algorithm, quadratic discriminant analysis (QDA) and a committee classification framework. OSeMA identifies a

Manuscript received April 03, 2006.

[*]To whom correspondence should be addressed.

S. Mostafavi is at School of Computing, Queen's University, Kingston, ON, K7L 3N6, Canada (e-mail: sara@cs.queensu.ca)

S. Baranzini is at Department of Neurology, University of California, San Francisco

J. Oksenberg is at Department of Neurology, University of California, San Francisco.

P. Mousavi is at School of Computing, Queen's University, Kingston, ON, K7L 3N6, Canada (phone: (613)533-6070, e-mail: pmousavi@cs.queensu.ca)

minimal set of orthogonal genes that are most relevant for classification of gene expression data and constructs models of different classes in the dataset. We employ OSeMA to construct a committee of classifiers that are highly predictive of the response of RRMS patients to IFNβ drug therapy based on gene expression data acquired prior to therapy initiation. Furthermore, we extensively evaluate OSeMA by comparing it to an exhaustive wrapper approach feature selection/classification procedure based on the QDA framework.

This manuscript is organized as follows: in section II we describe OSeMA and the general classification and validation procedures. In section III we describe our dataset and present and discuss our results. Section IV consists of summary and conclusions.

II. Methods

OSeMA constructs a classification model of the data in two steps. In the first step, a search is conducted to identify a minimal set of genes that are most relevant to distinguishing the underlying classes in the dataset. In the second step, the identified gene set is utilized to construct a committee of QDA classifiers.

A. Gene Selection with OSeMA

The initial gene search of OSeMA is based on Fast Orthogonal Search (FOS) introduced in [9]. FOS is a supervised forward regression algorithm that has been successfully employed in a variety of applications. The goal in solving the regression problem given by:

$$\hat{y}(x) = \sum_{k=1}^{p} a_k f_k(x) \qquad (1)$$

is to determine the weights, *a,* and *p* functions, *f,* among *N* candidate functions, such that the mean squared error (MSE) between $\hat{y}$ and 'true' output *y* is minimal. Orthogonal Search aims to identify the *p* functions *f,* among a large pool of candidate functions, and their corresponding weights *a*. Functions are selected iteratively, one at a time by holding previously selected functions fixed, such that the new selected function results in the most decrease in MSE. This procedure is repeated until an error criterion is met or a desired number of functions are selected [8]. By projecting the functions in the orthogonal space, with the Gram-Schmid Orthogonalization algorithm, the contribution of each function in reducing the MSE can be assessed independently. In FOS, the orthogonalized function associated with each f_k is not explicitly calculated; this fact significantly reduces the computational burden of the procedure [8, 9].

For the purpose of gene selection we propose to use FOS for assessing the 'significance' of each gene in distinguishing between different classes in the dataset, where each gene *k* in the dataset is considered to be a function f_k. In this procedure we set a threshold *t* and use FOS to identify the top *t* most relevant genes.

In order to identify the minimal set of genes that are most relevant in distinguishing the classes in the dataset we propose the following procedure. First we randomly divide the dataset into four equal parts and use three parts to identify a predefined number of genes with the FOS algorithm. We then repeat this procedure 100 times and rank the identified genes based on their frequency of appearance in all the random divisions of the data. As it will be shown in the results, there is a clear separation between the frequency of appearance of the highly ranked genes and the remaining genes.

B. Creating a Classifier with OSeMA

Subsetquet to the search step, OSeMA constructs a committee of QDA classifiers. In previous studies in the literature, various classification methods have been successfully employed to gene expression (see [7] for a review). However, gene expression studies are characterized by high degree of measurement noise and variability [10], therefore probabilstic approaches offer a well suited framework for predictive analysis. Linear and Quadratic Discriminant Analysis (LDA and QDA) are classical probabilistic classification approaches that are popular because of their solid theoretical foundation as well as their simplicity to implement and interpret [11].

A gene expression dataset $D_{n,p}$ comprises measurements of expression levels of *p* genes in *n* samples. Each sample has a class label y_i and $y \in \{1,...K\}$, where *K* is the number of different classes in the data. Assuming a multivariate Gaussian distribution for each class with different mean and common (LDA) or different (QDA) covariance matrix and utilizing Bayes rule, classifying a new sample entails calculating the probability of generation of the sample from each of the underlying class distributions:

$$p(k \mid x) = \frac{N_p(x; \mu_k, \Sigma_k)\, p(k)}{\sum_{m=1}^{K} N_p(x; \mu_m, \Sigma_m)\, p(m)} \qquad (2)$$

Where μ_k ,Σ_k are the mean and the covariance of class *k* of the *p*-dimensional multivariate Gaussian distribution N_p, $x \in \Re^p$ is a *p*-dimensional new sample, and *p(k)* is the prior probability of observing class *k*. Note that μ_k ,Σ_k are not known and must be estimated from the data [12].

In the second step of OSeMA, the identified minimal gene set is utilized to construct committee of QDA classifiers using the leave one out cross-validation (LOOCV) scheme. In this procedure, *n* QDA classifiers are constructed, each on different *n-1* samples in the data. In order to classify a new sample all committee members make a prediction and the sample is assigned to the class with majority of votes. The committee scheme proves to improve the generalizability of the classification models.

C. Classification and Evaluation Procedures

Due to the small number of available samples, stringent evaluation criterion is essential to avoid over-fitting to the training data. To evaluate our models critically, prior to each search-classification procedure, we randomly split the input gene expression dataset to 75% training data and 25% test data, while keeping the ratios of samples in each class consistent between training and test data. The training data is then used to train a committee of classifiers, using the LOOCV scheme, and the performance of the committee of classifiers is assessed on the test data. Thus, "prediction accuracy" refers to the classification accuracy on the held out test data. To ensure that the prediction accuracies were not just fortuitous and particular to a given split of the data, we created 100 random splits of data into the training and test sets and recorded the mean prediction accuracy as well as the 10th percentile over the 100 different data divisions.

To further evaluate the OSeMA framework we compare our results with that of IBIS [3]. IBIS is a search-classification framework that integrates exhaustive gene search and QDA classification in a committee framework. By exhaustively searching through all gene sets of size *r*, where *r* is pre-assigned by the user, IBIS identifies sets of *r* genes with the highest LOOCV accuracy and lowest mean squared prediction error (MSE).

III. Results

A. Dataset

The dataset in this study contains the expression levels of seventy seven genes in fifty two RRMS patients, obtained by Quantatative (q) *RT-PCR* from peripheral blood mononuclear cells. These patients have been followed for two years subsequent to initiation of rIFNβ drug therapy (time zero), blood specimens were collected every three months (resulting in measurements at time three, time six,…, time twenty-four) and at the end of the two years they have been classified as either poor- responders (19 patients) or good-responders (32 patients) to the treatment following strict clinical criteria. In a previous study, we have reported 3 dimensional IBIS classification results on predicting response of RRMS patients to rINFβ therapy prior to initiation of the treatment [3]. Here we compare gene selection and classification results of IBIS with that of OSeMA.

B. OSeMA – Gene Search

OSeMA was applied to the gene expression data of RRMS patients at time zero. As described in the II.C, the gene search step was solely conducted on the training portion of the data and similarly, the corresponding classifiers were creating using only the training data. In order to identify the most robust gene set, in the search step of OSeMA, we further divided the training data 100 times. On each division of the training data we randomly selected 75% of the samples to identify four most relevant genes. Subsequently, we ranked all the genes identified on different divisions of the data by their frequency of appearance. Figure 1 illustrates the frequency of identifications of genes in the search phase of OSeMA. Caspase 10 and MAP3k1 were consistently identified as 'significant' genes on different divisions of the training data.

Caspase2 and Caspase10 belong to the a family of proteins that are act as anti- or pro- apoptotic regulators and are involved in a wide variety of cellular activities. Previous gene expression studies have involved IFNβ in the regulation of apoptosis in MS [13, 14]. MAP3k1, occupies a pivotal rule in a network of phosphorylating enzymes integrating cellular response to a number of mitogenic and metabolic stimuli [15].

C. OSeMA – Classification and Validation

In the classification step, utilizing high ranking genes identified in the search step, committees of classifiers were constructed to model different classes in the data. The constructed classifier committees were then evaluated based on their prediction accuracy on the held out test dataset. To ensure that the prediction accuracies were not particular to the first split of the data into training and test set we report the mean prediction accuracy as well as 10th percentile of prediction accuracies on 100 splits of data. The 10th percentile indicates the minimum prediction accuracy in 90% of the instances if multiple predictions were made. Figure 2 illustrates the histogram of prediction accuracies, along with the mean and 10th percentile, of the two top ranking genes. The gene pair Caspase10 and MAP3k1 was identified in the search step of OSeMA and is sufficient to account for most of the variations between good- and poor- responder patients, as it results in high prediction accuracies on test data.

Furthermore, we compared the identified gene sets and corresponding accuracies of OSeMA with that of IBIS. IBIS reaches a mean prediction accuracy of 90% with a 4 dimensional search. In three dimensional IBIS search, the gene triplet Caspase 10, Caspase 2, and MAP3k1, is one of the top triplets and results in low MSE on the training data. Similarly, in a 2 dimensional search Caspase 10 and MAP3k1 were among the top ranking gene pairs. However, the exhaustive search in IBIS is computationally expensive and may not be achievable on datasets with large number of genes. IBIS requires exponential time with respect to the dimensionality of the search step whereas OSeMA requires linear time. A four dimensional IBIS search on this dataset require approximately 8 hours of computational time on a desktop PC whereas OSeMA require a few seconds on each division of the data. In addition, the gene ranking method in the OSeMA framework has the advantage that it results in identifying a robust minimal gene set, irregardless of the initial data split.

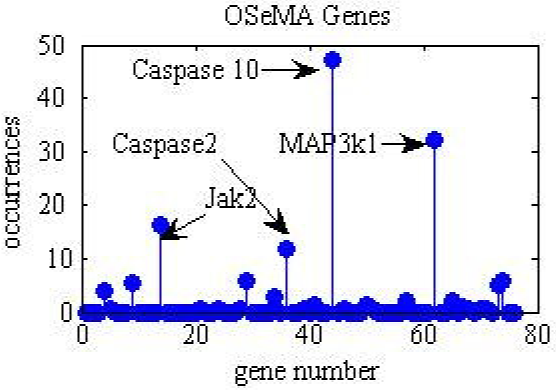

Figure 1. The frequency of occurrence of each of the 77 genes in the dataset in the search step of OSeMA

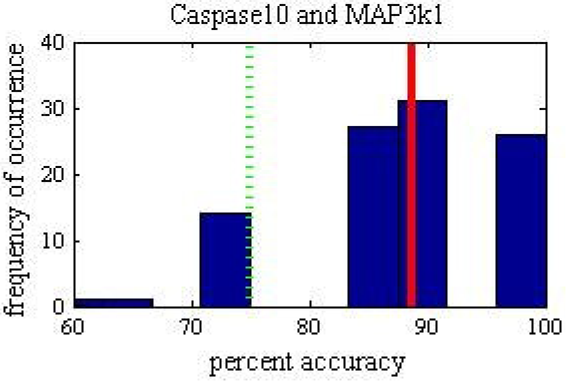

Figure 2. Histogram of prediction accuracies of OSeMA on 100 divisions of data into training and test sets using the top gene pair. The solid line represents the mean accuracy and the dashed line the 10th percentile of the histogram.

IV. Summary, conclusion and future work

In this study we present OSeMA, an integrated gene search/classification framework, which identifies a minimal gene set that is highly predictive of the therapy outcome in RRMS patients. We stringently evaluate the classification performance of OSeMA on held out test samples as well as through statistical approaches. Furthermore we compare our results to that of an exhaustive search with IBIS. The classification accuracies along with the identified gene sets are similar between IBIS and OSeMA. However, the search step in the OSeMA framework is very fast; requiring less than a minute on a personal PC, whereas the exhaustive IBIS search requires exponential time versus dimensionality of search. Additionally, the proposed ranking method in the search step of OSeMA allows for identification of a minimal and robust gene set, regardless of the particular division of the data into test and training sets.

We have successfully identified minimal gene groups, whose combinatorial transcriptional profiles distinguish poor- and good- responding RRMS patients, however, larger prospective studies are required to confirm our findings. We are currently investigating applying our poor- and good-responders predictive models to an independently produced RT-PCR gene expression data of a group of RRMS patients that have been classified to poor- or good- responder to drug therapy.

References

[1] National Multiple Sclerosis Society. http://www.nmss.org.

[2] Rio J, *et al.* (2002) Assessment of different treatment failure criteria in a cohort of relapsing-remitting multiple sclerosis patients treated with interferon beta: Implications for clinical trials. *Ann Neurol* 52: 400–406.

[3] Baranzini SE, *et al.*(2005) Transcription-based prediction of response to IFNbeta using supervised computational methods. *PLoS Biol.* Jan 3, 1:e2.

[4] Li, W. and Yang, Y. (2002) How many genes are needed for a discriminant microarray data analysis? *Methods of Microarray Data Analysis: Papers from CAMDA'00, eds.* SM Lin, KF Johnson (Kluwer Academic), pp. 137-150.

[5] Li, T., *et al*. (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20**,** 2429–2437.

[6] Blum, A., and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 245-271.

[7] Dudoit, S, Fridlyand, J, and Speed, T. (2002) Comparison of Discriminant Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*: Mar(v.97, No. 457):77-87.

[8] Chen, S. Cowan, CFN. And Grant, PM. (1991). Orthogonal least squares learning algorithm for radial basis functions networks. *IEEE Transactions on Neural Networks,* 2(2), 302-309.

[9] Korenberg, MJ. (1988). Identifying nonlinear difference equation and functional expansion representations: the fast orthogonal algorithm. *Ann Biomed, Eng.* (16) 123-142.

[10] Kendziorski, M, *et al*. (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* 22: 3899-3914.

[11] Huberty CJ. (1994) Applied Discriminant Analysis. New York: John Wiley & Sons.

[12] Hastie, T, Tibshirani, R, and Friedman, J. (2001).The Elements of Statistical Learning: Data Mining, Inference, and Prediction *Springer*: New York.

[13] Chawla-Sarkar M, *et al.* (2003) Apoptosis and interferons: Role of interferon-stimulated genes as mediators of apoptosis. *Apoptosis* 8: 237–249.

[14] Gniadek P, *et al.* (2003) Systemic IFN-beta treatment induces apoptosis of peripheral immune cells in MS patients. *J Neuroimmunol* 137: 187–196.

[15] Xia, JX. *et al.* (2005). B7 molecules mRNA expression in colorectal carcinoma. *World J Gastroenterol.*Sep 28;11(36):5655-8.