Mining of Microarray, Proteomics, and Clinical Data for Improved Identification of Chronic Fatigue Syndrome

Hongbo Xie, Zoran Obradovic, Slobodan Vucetic

Information Science and Technology Center, Temple University
1805 N. Broad St., Philadelphia, PA 19122

Abstract. Chronic Fatigue Syndrome (CFS) is a recently recognized disease whose pathophysiology is insufficiently understood. The objective of this study was to explore if identification accuracy of CFS could be improved using microarray and proteomics data alone or when integrated with clinical data. First, a two-step approach for selection of genetic CFS biomarkers from microarray data is proposed. The underlying assumption is that CFS is characterized by deviations in expression of genes from a limited set of functions. The approach starts by selection of significantly differentially expressed genes by using standard statistical testing procedure. Using Gene Ontology (GO) resource, biological functions of the selected genes are studied to discover the ones that are highly overrepresented by the selection. Only the selected genes annotated with the most significant function are selected as biomarkers for identification of CFS. This approach results in a small set of biomarkers whose function is the most relevant to CFS. In our experiments Support Vector Machine (SVM) that uses as attributes genes obtained by the two-step approach achieved higher accuracy than when using genes obtained by the traditional one-step approach. (e.g. 72% vs 53% accuracy when selection based on p-value 0.05). Moreover, the finding that mRNA processing is the most representative function is consistent with the previously published results. In the second part of the study, benefits of combining microarray and proteomics data in CFS identification were explored. Using the standard procedure for preprocessing of ProteinChip data, we developed a proteomics-based predictor of CFS. Our results on the 38 samples with both microarray and ProteinChip data indicates that predictor combination can provide improved CFS identification (79% accuracy by a combination when two approaches agree vs. 72% obtained by microarray alone). However, an important observation is that the achieved accuracy of CFS identification of less than 80% is relatively low as compared to some other diseases, such as cancer. This suggests that identification of CFS biomarkers is a challenging task that requires significantly larger amounts of experimental data. Finally, we studied the clinical CFS data to discover factors that explain sources of CFS identification mistakes. We discovered significant difference in mental health, physical fatigue, and general fatigue indicators among cases differently classified by microarray and proteomics methods. This suggests that CSF identification could be improved by revising definitions of certain clinical conditions.

1. Introduction

Chronic Fatigue Syndrome (CFS) is a disease characterized by severe chronic fatigue lasting at least 6 months which is accompanied by symptoms such as impairment in short-term memory or concentration, sore throat, tender lymph nodes, and muscle pain [1]. Since there are no specific diagnostic tests for CFS, the diagnosis is made by ruling out other causes of fatigue [2]. CFS case definition, prevalence, clinical presentation, evaluation, and prognosis are still under investigation and encounter many challenges [3].

Microarray and proteomics technologies can be extremely helpful in deciphering the underlying mechanisms of CFS and lead to discovering CFS diagnostic biomarkers. The task of biomarker discovery is related to the task of attribute selection in machine learning. Attributes, or biomarkers, could be genes in microarray data or mass-charge peaks in proteomics data. As a standard approach, following the data normalization and preprocessing, attributes are selected by using various computationally fast heuristics or statistically-rooted measures [4]. However, the standard approach has many limitations. Gene expression measurements by the microarray technology are very noisy and this could lead to selection of many irrelevant biomarkers. Similarly, with proteomics data, experimental conditions, such as chip surface factors, washing stringency factors, spot factors, or laser energy factors could significantly influence the shape of the resulting spectra. Clearly, robust approaches for biomarker selection and integration of data from various technologies are necessary for successful knowledge extraction.

This study has two major objectives: one is to test the novel approach for biomarker selection from microarray data; another is to explore if it can be effectively integrated with proteomics and clinical information to gain better understanding of the CFS. In the following sections, we outline the methodology used, report the results of performed experiments, and discuss the relevance of the obtained results.

2. Methods and Materials

2.1. Biomarker Selection from Microarray Data

Let us assume we are given a gene expression data set D, $D = \{x_{ij}\}$, $i = 1\ldots N$, $j = 1\ldots K$, where N is the number of examples (arrays), K is the number of genes, and x_{ij} is expression of gene g_j on array a_i. In classification scenario, arrays are assigned class labels y_i, $i = 1\ldots N$, where $y_i \in \{1...C\}$, and C is the number of classes. The goal is to build a classification model that accurately predicts class of a new array represented by the vector of K gene expressions $[x_1...x_K]$. Given such a problem setup, N corresponds to number of labeled examples and K to number of attributes. To reduce the number of attributes and to focus on biologicaly relevant genes we used filter based attribute selection by the Kruskal-Wallis test together with domain knowledge of biological functions of genes as explained in this section.

2.1.1. Attribute Selection by the Kruskal-Wallis Test

Significance of each gene is measured by the Kruskal-Wallis test. Expressions of gene g_i over the N arrays are sorted and the arrays are assigned ranks based on their position in the sorted list. By denoting r_k as the average rank of arrays with class label k, the Kruskal-Wallis test statistics for gene g_i is calculated as

$$KW_i = \frac{12}{N(N+1)} \sum_{k=1}^{C} n_k (r_k - \frac{N+1}{2})^2 ,$$

where n_k is the number of labeled examples with class label k. Under the null hypothesis that expression of gene g_i is independent of the class label, KW_i follows the chi-square distribution with C–1 degrees of freedom. If gene expression is dependent on the class label, the value of KW_i will be significantly higher than expected by the chi-squared distribution. The significance of gene g_i can be measured by how likely it is that a value equal or larger than KW_i is generated by the chi-squared distribution. This quantity, called the *p-value* of gene g_i, is calculated as $G_i = \mathrm{P}(X \geq KW_i)$, where X is a random variable with the chi-squared distribution with C−1 degrees of freedom. Significant genes will have low p-values, and attribute selection can be performed by selecting all genes with p-value below a specified threshold θ.

2.1.2. Domain Dependent Attribute Selection

Given the meaning of the p-value, it can be concluded that the number of false positive, or irrelevant, genes selected by the Kruskal-Wallis test is equal to the product $K\theta$. In this study, we proposed to use domain knowledge to further improve quality of attribute selection. The approach is based on the assumption that the most discriminative genes are likely to correspond to a limited set of biological functions or pathways. Given the assumption, genes selected by the Kruskal-Wallis test whose biological properties deviate from other selected genes can be considered as noise and excluded from further consideration. In the following, we describe our algorithm for domain dependent attribute selection.

We used the Gene Ontology (GO) resource [5] as source of the high-quality domain knowledge about gene functions. The GO provides information about biological functions of genes using a controlled vocabulary. Given a set of M biological functions, called the GO terms, $go_1...go_M$, every gene is assigned an M-dimensional vector $[f_1 ... f_M]$, where $f_i = 1$, if the gene has been annotated with the GO term go_i, and $f_i = 1$, otherwise. Out of the nearly 20,000 GO terms, each gene is typically assigned to only a few of them.

Given a set of genes selected by with the Kruskal-Wallis test, $\{g_i, G_i < \theta\}$, we are interested in determining whether a given GO term is overrepresented by the selection. Let us denote k as the number of genes selected by the Kruskal-Wallis test, k_i as the number of genes annotated with the GO term go_i, and x_i as the number of selected genes annotated with the GO term go_i. If the gene selection were random, the quantity x_i would follow the hypergeometric distribution $H(K, k, k_i)$. If the selection were favorable to genes annotated with GO term go_i, tha value of x_i would be significantly higher then expected by the hypergeometric distribution. Therefore, the p-value $GO_i = \mathrm{P}(\mathrm{X} \geq x_i)$, where $X \sim H(K, k, k_i)$, can be used to measure the significance of go_i.

In our approach, the p-values GO_i , $i = 1...M$, for each GO term are evalueated and the one with the smallest p-value is seleced. Let us denote this GO term as go^*. Then, only genes selected by the Kruskal-Wallis test and annotated with go^*, $\{g_i, \; G_i < \theta \wedge g_i \in go^*\}$, are used as the attributes in classification.

2.2 Data Sets and Data Preprocessing

The datasets using in this study include microarray and proteomics data. Their preprocessing procedures are described in the following two sections.

2.2.1 Microarray data and its preprocessing

Microarray data provided by CAMDA organizers included 177 arrays. In this study we used 79 arrays representing 39 clinical identified CFS samples and 40 non-CFS samples. The excluded arrays include 3 arrays without associated clinical data which, 8 replicate arrays, and 54 arrays related to medical or psychiatric exclusionary conditions [3]. There are 20,160 genes spotted at every chip. Artifact-removed density values minus the background density value attributes provided by the organizers were used as the starting point of this study. As a result, 79×20,160 dataset was obtained with each row representing a sample (CFS/NF) and each column representing expression patterns of every gene. Using the SOURCE (source.stanford.edu) [6] resource, we obtained Gene Ontology (GO) information for 13,213 genes described by 4,110 unique GO terms.

2.2.2 Proteomics data and preprocessing

Proteomics data included 65 samples representing 33 CFS and 32 non-CFS samples. Each sample was profiled under 48 different conditions, with factors such as fractionation, ProteinChip surfaces, and binding and elution conditions. In addition, a control sample from the human serum from Centers for Disease Control (CDC) donor pool was provided. Data preprocessing was performed in accord with the standard procedure [7,8] which starts with baseline correction and follows by peak alignment to standardize the M/Z values, spectrogram smoothing to denoise the data, and spectrogram normalization by dividing each spectrogram by the corresponding control spectrogram. To reduce dimensionality of proteomics data, Kruskal-Wallis test was applied on all mass/charge values to select most discriminative peaks. The selected peaks were used as attributes in identification of CFS samples.

3. Results

3.1 CFS Classification from Microarray Data

In our experiments a 79-leave-one-out cross validation procedure was used to select significantly differentially expressed genes, develop a classifier based on these and to evaluate the prediction accuracy. More precisely, one out of 79 arrays was selected for testing and the remaining 78 were used as a training set for attribute selection and development of a classifier. The classifier used in all experiments was support vector machine (SVM) with quadratic kernel $k(x, y) = (C + x^T y)^2$. This procedure was repeated 79 times, each time leaving out a different array for testing.

Table 1. Performance of attribute selection using Kruskal-Wallis (KW) test (Section 2.1.1) vs. the domain dependent selection (Section 2.1.2). Support vector machine was used for classification and leave-one-out cross-validation scores are reported rounded to integers.

p-value threshold	Attribute selection with KW test		Two-step attribute selection	
	Accuracy (average)	Selected genes (average)	Accuracy (average)	Selected genes (average)
0.005	54%	126	54%	3
0.01	48%	257	52%	3
0.05	53%	1296	72%	17
0.1	57%	2560	58%	16
0.2	49%	3761	54%	19

Five sets of 79-leave-one-out cross validation experiments were repeated using p-value thresholds of 0.005, 0.001, 0.05, 0.1 and 0.2 using Kruskal-Wallis test described in Section 2. More than 500 genes satisfied the criteria with p-value threshold <0.01 and more than 2,000 genes were selected for larger p-value thresholds (the exact numbers are reported at Table 1 under KW test for attribute selection).

Following the approach described in Section 2.1.2, in each leave-one-out experiment we used hypergeometric distribution test to discover the most overrepresented GO term by the Kruskal-Wallis selection. Only GO terms with less than 30 associated genes were considered. Significantly expressed genes annotated with the most overrepresented GO term were used to build an SVM classifier. The percent correct accuracy by the leave-one-out procedure is reported in Table 1. The average size of the selected set of genes used as features in SVM classifier is also reported.

In our experiments the CFS and non-CFS arrays were almost indistinguishable using the traditional attribute selection (KW test) and SVM classification. The proposed domain dependent attribute selection allowed more accurate classification. For p-value of 0.05 the domain dependent selection was almost 20% more accurate than the traditional selection (72% vs 53%). Another important advantage of domain dependent feature selection was that it resulted in about two orders of magnitude smaller number of selected genes. As a result, the identified genes can be much easier evaluated as possible biomarker targets.

We also compared the result of the domain dependent selection to using the same number of most significantly expressed genes by the traditional selection. The resulting prediction accuracy based on the leave-one-out validation was in the range between 43-48%, which is far lower than the accuracy obtained when domain dependent attributes are used.

3.2 Biomarker Selection from Microarray Data

Towards biomarker identification goal we examined the most overrepresented GO terms among the genes selected by Kruskal-Wallis test with the p-value threshold of 0.05. Analyzing the results of 79 leave-one out cross-validation, we found that 13 different GO terms were identified as the most significant in at least one of the 79 experiments. However, the three GO terms that occurred most often (Cholesterol Metabolism, mRNA Processing, and Actin Binding) occurred in 56 out of the 79 experiments. In Table 2 we list the most overrepresented GO terms among genes selected by Kruskal-Wallis test with p-value threshold 0.05 from the complete set of 79 microarrays. The top two functions, with hypergeometric test p-values of 0.0016 and 0.0021, are consistent with the previously reported result indicating that expression of metabolic and RNA processing genes is related to CFS [9].

Table 2. P-values of the most overrepresented GO terms among the significantly differentially expressed genes. Numbers of differentially expressed genes annotated with the given GO term are also listed.

Gene Ontology ID	Function/Process Name	p-value	Number of selected genes
GO:0006397	mRNA processing	0.0016	10
GO:0008203	cholesterol metabolism	0.0021	7
GO:0003779	actin binding	0.0027	31
GO:00015629	actin cytoskeleton	0.0078	14
GO:00016564	transcriptional repressor activity	0.0105	9
GO:0005515	protein binding	0.0136	124
GO:0007187	G-protein signaling	0.0153	5
GO:0008009	chemokine activity	0.0153	5
GO:0007229	integrin-mediated signaling pathway	0.0155	9
GO:0007517	muscle development	0.016	14

In Table 3 we list the 10 significantly differentially expressed genes with mRNA Processing GO term. These 10 genes comprise a **set of potential biomarkers for CFS diagnosis**.

Table 3. mRNA binding genes identified as biomarkers.

Gene Name	Gene ID	Symbol	Unigene	P-value
Debranching enzyme homolog 1	AK000116	DBR1	Hs.477700	0.0086
Cleavage and polyadenylation specific factor 6	NM_007007	CPSF6	Hs.369606	0.0096
Small nuclear ribonucleoprotein polypeptide N	AF101044	SNRPN	Hs.525700	0.0131
Hypothetical protein	BC006407	MGC14151	Hs.333414	0.0186
Heterogeneous nuclear ribonucleoprotein L-like	BC008217	HNRPLL	Hs.445497	0.0212
TRNA splicing endonuclease 2 homolog	AK074794	SEN2L	Hs.335550	0.0223
Poly(A) polymerase beta	AF218840	PAPOLB	Hs.487409	0.0333
ELAV-like 4 (Hu antigen D)	BC036071	ELAVL4	Hs.213050	0.0376
ER to nucleus signalling 1	AF059198	ERN1	Hs.133982	0.0395
Nuclear ribonucleoprotein polypeptide	J04564	SNRPB	Hs.83753	0.0444
Nuclear RNA export factor 1	AF112880	NXF1	Hs.523739	0.0499

3.3 Microarray and Proteomics CFS Classification Analysis for Mining Clinical Data

Experiments on proteomics data revealed that accuracy of CFS identification using our approach does not exceed that of trivial predictor. based on the analysis of the 48 SVM classifiers for each of the 48 types of proteomics data, we concluded that IMAC chips provide the best overall results. The accuracy of an ensemble of 13 IMAC classifiers by the leave-one-out cross-validation was 51%.

For 38 subjects (20 CFS and 18 non-CFS samples), both proteomics and microarray data was provided. We compared the proteomics-based to the microarray-based CFS classification. Their predictions agreed on exactly 50% of the samples (19 samples). Among these 19 samples, the accuracy of the combined approach was 79%, suggesting that significant improvement are possible when relying on a combination of proteomics and microarray CFS identification approaches.

Further analysis using the provided clinical data was pursued to detect potential factors in the clinical records that reveal the reasons of disagreement between microarray and proteomics CSF classifiers. Using ANOVA of two groups of means, we found significant difference in three of the clinical classifying attributes. The 3 discovered attributes are mental heath, physical fatigue, and general fatigue. This result indicates that the reason for relatively low accuracy of CFS diagnosis could be partially blamed on the clinical definition of the disease. This strongly indicates that by **combining clinical data with microarray and proteomics data could lead to improved understanding and diagnosis of CFS**.

4. Conclusion

To improve identification accuracy of CFS using microarray data it is critical to develop better feature selection methods. A two step approach proposed in this study involves statistical gene selection process complemented by focus to the most significantly overrepresented Gene Ontology terms. The benefits of such a gene selection process were clearly demonstrated by improved prediction accuracy as compared to using standard statistical selection methods. In addition, a much smaller number (by two orders of magnitude) of genes were selected for evaluation as possible biomarker targets. Integrating information from multiple sources provides another opportunity for improving identification accuracy of CFS. A simple ensemble predictor based on microarray and proteomics data is demonstrated to improve CSF identification accuracy. In addition, joint analysis of microarray and proteomics identification mistakes together with clinical information resulted in a discovery that revising definitions of certain clinical conditions might lead to improved CSF identification.

References:

1. Fukuda, K., et al. The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group. in Ann Intern Med. 1994.
2. Smets, E.M., et al., The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res, 1995. 39(3): p. 315-25.
3. Reeves, W.C., et al., Chronic fatigue syndrome--a clinically empirical approach to its definition and study. BMC Med, 2005. 3: p. 19.
4. Almuallim, H. and T.G. Dietterich. Learning with many irrelevant features. National Conference on Artificial Intelligence. 1992: p. 547-552
5. Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. 25(1): p. 25-9.
6. Diehn, M., et al., SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res, 2003. 31(1): p. 219-23.
7. Carlson, S.M., Najmi, A., Whitin, J.C., Cohen, H.J., Improving Feature Detection and Analysis in SELDI-TOF Mass Spectra. PROTEOMICS, 2005. 5(11).
8. Pallavi N. Pratapa, E.F.P., Jr., Alexander J. Hartemink, Finding Diagnostic Biomarkers in Proteomic Spectra. Pacific Symposium on Biocomputing, 2006. 11: p. 279-290.
9. Toni Whistler,1 Elizabeth R Unger,1 Rosane Nisenbaum,1 and Suzanne D Integration of gene expression, clinical, and epidemiologic data to characterize Chronic Fatigue Syndrome, Transl Med. 2003; 1: 10