EURASIP Journal on Bioinformatics and Systems Biology
a SpringerOpen Journal

RESEARCH Open Access

Subtyping glioblastoma by combining miRNA and mRNA expression data using compressed sensing-based approach

Wenlong Tang[1], Junbo Duan[1], Ji-Gang Zhang[2] and Yu-Ping Wang[1,2,3*]

Abstract

In the clinical practice, many diseases such as glioblastoma, leukemia, diabetes, and prostates have multiple subtypes. Classifying subtypes accurately using genomic data will provide individualized treatments to target-specific disease subtypes. However, it is often difficult to obtain satisfactory classification accuracy using only one type of data, because the subtypes of a disease can exhibit similar patterns in one data type. Fortunately, multiple types of genomic data are often available due to the rapid development of genomic techniques. This raises the question on whether the classification performance can significantly be improved by combining multiple types of genomic data. In this article, we classified four subtypes of glioblastoma multiforme (GBM) with multiple types of genome-wide data (e.g., mRNA and miRNA expression) from The Cancer Genome Atlas (TCGA) project. We proposed a multi-class compressed sensing-based detector (MCSD) for this study. The MCSD was trained with data from TCGA and then applied to subtype GBM patients using an independent testing data. We performed the classification on the same patient subjects with three data types, i.e., miRNA expression data, mRNA (or gene expression) data, and their combinations. The classification accuracy is 69.1% with the miRNA expression data, 52.7% with mRNA expression data, and 90.9% with the combination of both mRNA and miRNA expression data. In addition, some biomarkers identified by the integrated approaches have been confirmed with results from the published literatures. These results indicate that the combined analysis can significantly improve the accuracy of classifying GBM subtypes and identify potential biomarkers for disease diagnosis.

Keywords: Glioblastoma, Data integration, Compressed sensing, Classification, mRNA, miRNA

Introduction

Many diseases including cancers have multiple subtypes. For example, leukemia has four main categories: acute lymphoblastic leukemia (ALL), acute myelogenous leukemia, chronic lymphocytic leukemia, and chronic myelogenous leukemia. Each of these categories can be further divided into different subtypes [1]; for example, ALL can be further subtyped into six types [2]. Glioma has four subtypes, including oligodendroglioma, anaplastic oligodendroglioma, anaplastic astrocytoma, and glioblastoma multiforme (GBM) [3]. Prostate cancer has three major subtypes [4]. An accurate and effective classification of those subtypes based on genomic data will result in personalized treatments of the cancer in terms of a particular subtype. In this article, we are interested in the subtyping of GBM, which is a kind of glioma and is the most common form of malignant brain cancer in adults [5]. There is an increasing interest in classifying multiple subtypes of GBM based on its genomic measurements. Most of the existing works are based on gene expression data only. Benjamin et al. [6] classified two types of GBM in adults and found that the genes EGFR and TP53 were important in discriminating the two subtypes. Nutt et al. [7] built a *k*-nearest neighbor model with 20 features to classify 28 glioblastomas and 22 anaplastic oligodendrogliomas and found that the class distinctions were significantly associated with survival outcome (p = 0.05). Noushmehr et al. [8] separated a subset of samples in GBM from The Cancer Genome Atlas (TCGA) project, which displayed concerted

* Correspondence: wyp@tulane.edu
[1]Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA
[2]Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, USA
Full list of author information is available at the end of the article

Springer

© 2013 Tang et al.; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

hypermethylation at a large number of loci. The datasets we used to subtype GBM are also from TCGA. The subtypes of GBM samples in TCGA includes: pro-neural, neural, classical, and mesenchymal [9]. The GBM data we have tested include both miRNA expression and mRNA expression data. The miRNAs, also called microRNAs, are short non-coding RNA molecules that were recently found in all eukaryotic cells except fungi, algae, and marine plants. The human genome may contain over 1,000 miRNAs [10]. Aberrant expressions of miRNAs have been found to be related to many diseases, including cancers [11,12]. They play an essential role in tissue differentiation during normal development and tumorigenesis [13].

In the last decade, the development of genomic techniques enables the availability of multiple data types on the same patient, such as mRNA or gene expression, SNP, miRNA expression, and copy number variation data. It is well recognized that a more comprehensive analysis result could be obtained based on integrating multiple types of genomic data than using an individual dataset. Soneson et al. [14] investigated the correlation between gene expression and copy number alterations using canonical correlation analysis for leukemia data. A web-based platform, called Magellan, was developed for the integrated analysis of DNA copy number and expression data in ovarian cancer [15], which found significant correlation between gene expression and patient survival. Troyanskaya et al. [16] developed a Bayesian framework to combine heterogeneous data sources to predict gene function with improved accuracy. A kernel-based statistical learning algorithm was also proposed in the combined analysis of multiple genome-wide datasets [17]. In this article, we propose a novel classifier based on the compressed sensing (CS) theory that we have been working with.

The CS technique enables compact storage and rapid transmission of large amounts of information. The technique can be used to extract significant statistical information from high-dimensional datasets [18]. The CS technology has been proven to be a powerful tool in the signal processing and statistics fields. It demonstrates that a compressible signal can be recovered from far fewer samples than that needed by the Nyquist sampling theorem [19]. Our recent work used a CS-based detector (CSD) for subtyping leukemia with gene expression data [20]. The CSD achieved high classification accuracies, with 97.4% evaluated with cross-validation and 94.3% evaluated with an independent dataset. The CSD showed better performance in subtyping two types of leukemia compared to some traditional classifiers such as the support vector machine (SVM), indicating the advantage of the CSD in analyzing high-dimensional genomic data. In this article, we extended the CSD to multiple data types and proposed a detector called MCSD. In particular, we applied the MCSD to the subtyping of four types of GBM by combining miRNA expression and mRNA expression data. We present a novel combined analysis method based on the CS and demonstrate that the classification performance can significantly be improved in subtyping four types of GBM, with both miRNA expression and mRNA expression data.

Methods and materials

Data collection

The GBM data used in this study are publicly available from the website of TCGA [21]. The patients in the dataset can be classified into four subtypes, i.e., pro-neural, neural, classical, and mesenchymal [9]. The genomic data include miRNA expression (1,510 probes) and mRNA expression data (22,277 probes). We randomly divided the data (including 115 patients with both miRNA and mRNA expression data) into two sets: training and testing datasets. The total number of patients in the training dataset was 60 with 15 patients in each group. The testing dataset had 55 patients, with 17 pro-neural, 3 neural, 17 classical, and 18 mesenchymal subtypes (as listed in Table 1). The same number of patients in each subtype for training data was used for reducing the bias in the model building. Meanwhile, the numbers of patients in training and testing were approximately the same.

For multiple types of genomic data (e.g., miRNA expression data, mRNA expression data, etc.), we used $\boldsymbol{x}_{1i}$ to denote the data vector for the ith sample in data 1 (e.g., miRNA expression), $\boldsymbol{x}_{2i}$ to denote the data vector for the ith sample in data 2 (e.g., mRNA expression), and $\boldsymbol{x}_{ni}$ to denote the data vector for the ith sample in data n. The combined data for the ith sample is $x_i = \left(x_1{}_i^T, x_2{}_i^T, \ldots, x_n{}_i^T\right)^T$, which is arranged in a cascaded manner.

MCSD

Bayesian classifier

To classify a given observation $\boldsymbol{y}$ to one of n classes, we define the actual class ("ground truth") to which it belongs as g; the class to which it is assigned ("decision") as d. The n classes are defined as: $\pi_1, \pi_2, \ldots, \pi_n$. Let $U_{\pi i}$ $(\boldsymbol{y}, g)$ be the

Table 1 GBM subtypes and their corresponding samples used for the training and the testing

Glioblastoma subtypes	Training (total 60)	Testing (total 55)
Pro-neural	15	17
Neural	15	3
Classical	15	17
Mesenchymal	15	18

These datasets are publically available from the TCGA project.

utility of assigning $\boldsymbol{y}$, actually from π_g, to π_i. The "utility" is negative relevant to the Bayes Risk (BR) [22], which is the minimum classification error. Thus, we make: U = 1-BR. The two-class one-dimensional BR (shaded area in Figure 1) can be calculated by

$$\mathrm{BR} = P_2 \int_{-\infty}^{y_0} p_2(y)dy + P_1 \int_{y_0}^{\infty} p_1(y)dy, \tag{1}$$

where y_0 is the decision boundary, P_1,P_2 are the prior probabilities and $P_1(y)$,$P_2(y)$ are the conditional probability density functions of the two classes, respectively (shown in Figure 1).

Let us extend the BR to n classes and N dimensions. Then Equation (1) can be rewritten as

$$\mathrm{BR}_N = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} P_j \int_{\Omega i} p_j(y)dy, \tag{2}$$

where P_j is the prior probability of a given subject belonging to the class π_j, j = 1,. . .,n; $P_j(y)$ is the conditional probability density function of the class π_j, and Ω_i is the Bayesian decision region for class π_i [23].

For multi-class classification, an ideal detector should yield

$$P\left(d = \pi_i | g = \pi_j\right) = \delta_{ij},$$

where

$$\delta_{ij} = \begin{cases} 0, & i\neq j \\ 1, & i = j \end{cases}, \tag{3}$$

where $P(d = \pi_i|g = \pi_j)$ denotes the probability of assigning a given observation $\boldsymbol{y}$, actually belonging to π_j, to π_i. δ_{ij} is the Kronecker's delta.

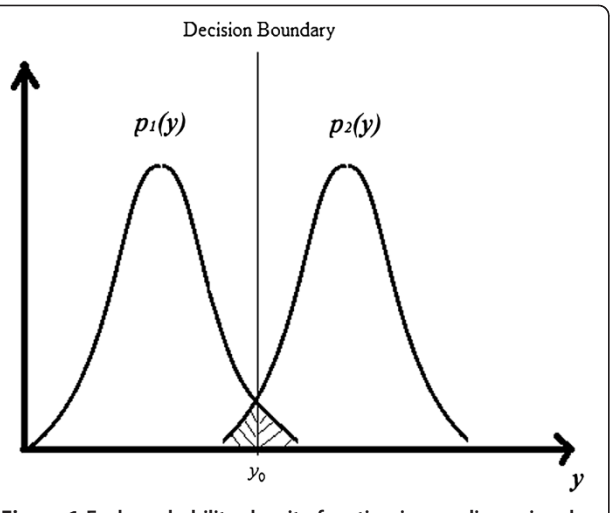

Figure 1 Each probability density function is one-dimensional normal distribution (area under each curve sums to 1).

According to the ideal observer decision theory [22], a decision is selected only if its expected utility is greater than the expected utility of any others. Thus, for any given observation $\boldsymbol{y}$, we decide $d = \pi_i$ iff

$$E\{U_{\pi_i}(\boldsymbol{y},g)|\boldsymbol{y}\} > E\{U_{\pi_j}(\boldsymbol{y},g)|\boldsymbol{y}\}, i\neq j \quad j = 1,2,\ldots,i-1,i+1,\ldots,n. \tag{4}$$

From Equation (2) and the relationship of utility and BR, we know "utility" is a number that can be calculated. We denote that number as $U_{i|j}$ to express the utility of assigning a given observation $\boldsymbol{y}$, actually belonging to π_j, to π_i. The inequality (4) can be written as

$$\sum_{k=1}^{n} U_{i|k} P(g = \pi_k|\boldsymbol{y}) > \sum_{k=1}^{n} U_{j|k} P(g = \pi_k|\boldsymbol{y}), \quad i\neq j, j = 1,2,\ldots,i-1,i+1,\ldots,n,. \tag{5}$$

We apply Bayes' rule

$$P(g = \pi_k|\,\boldsymbol{y}) = \frac{P_{\boldsymbol{y}}(\boldsymbol{y}|g = \pi_k)P(g = \pi_k)}{P_{\boldsymbol{y}}(\boldsymbol{y})}, \tag{6}$$

where $P_y(\boldsymbol{y}|g = \pi_k)$, k = 1,. . .,n, is the probability density function for the signal observations. According to Inequality (5), we decide $d = \pi_i$ iff

$$\sum_{k=1}^{n} U_{i|k} P(g = \pi_k) p_{\boldsymbol{y}}(\boldsymbol{y}|g = \pi_k) > \sum_{k=1}^{n} U_{j|k} P(g = \pi_k) \quad p_{\boldsymbol{y}}(\boldsymbol{y}|g = \pi_k)\ i\neq j\ j = 1,2,\ldots,i-1,i+1,\ldots,n. \tag{7}$$

That is known as maximum likelihood estimation. Specifically, the class label of the testing sample $\boldsymbol{y}$ is given by

$$\mathrm{ID} = \underset{l}{\mathrm{argmax}} \left[\sum_{k=1}^{n} U_{l|k} P(g = \pi_k) p_{\boldsymbol{y}}(\boldsymbol{y}|g = \pi_k)\right]. \tag{8}$$

If we assume π_1, π_2,. . .,π_n have the same prior probability, i.e., $P(g = \pi_k) = \frac{1}{n}$. The detector (8) can be rewritten as

$$ID = \underset{l}{\mathrm{argmax}} \left[\sum_{k=1}^{n} U_{l|k} p_{\boldsymbol{y}}(\boldsymbol{y}|g = \pi_k)\right]. \tag{9}$$

The calculation of the utility is shown in the Additional file 1.

Dimension reduction using CS

To reduce the dimension of original sample, we design a projection (sparse) matrix $\boldsymbol{\Phi}$, called compress matrix.

The generation of the compress matrix can be formulated as a sparse representation problem as in Equation (10)

$$\boldsymbol{Y} = \boldsymbol{\Phi S}, \tag{10}$$

where $\boldsymbol{Y} = \{\boldsymbol{y}_i\} \in \mathbb{R}^{M \times c}$ is the projected sample, M is the dimension of the sample after the projection, $\boldsymbol{y}_i$ is the ith column in the compressed signal, c is the total number of columns in the compressed signal, $\boldsymbol{S} = \{\boldsymbol{s}_i\} \in \mathbb{R}^{N \times c}$ is the original signal, and N is the dimension of the original signal and $N >> M$. The matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ is a sparse matrix, with most of the entries '0's. The compress matrix $\boldsymbol{\Phi}$ projects the original sample $\boldsymbol{S}$ to a much smaller dimensional signal $\boldsymbol{Y}$. The original sample may contain redundancy; through this projection, the original sample can significantly be compressed and compactly represented, which usually lead to better classification performance. Suppose we have n groups, with c_1 training samples in group 1, c_2 training samples in group 2, and so forth, c_n training samples in group n, and $c = c_1 + c_2 + \cdots + c_n$ for $\boldsymbol{S} = \left[\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_c\right] \in \mathbb{R}^{N \times c}$ and $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_c\right] \in \mathbb{R}^{M \times c}$. The transpose of Equation (10) is

$$\boldsymbol{S}^T \boldsymbol{\Phi}^T = \boldsymbol{Y}^T. \tag{11}$$

Let $\left(\boldsymbol{\Phi}^T\right)_j \in \mathbb{R}^{N \times 1}$ denote the jth column of $\boldsymbol{\Phi}^T$, and $\left(\boldsymbol{Y}^T\right)_j \in \mathbb{R}^{c \times 1}$ denote the jth column of $\boldsymbol{Y}^T$, where $j = 1, 2, \ldots, M$. Then Equation (11) can be rewritten as

$$\boldsymbol{S}^T \left(\boldsymbol{\Phi}^T\right)_j = \left(\boldsymbol{Y}^T\right)_j. \tag{12}$$

The linear system given by (12) is an underdetermined system, which can be solved by using l-1 norm minimization algorithm such as Homotopy method, or the least angle regression method [24]. The l-1 norm optimization problem reads

$$\left(\boldsymbol{\Phi}^T\right)_j = \underset{\left(\boldsymbol{\Phi}^T\right)_j}{\operatorname{argmin}} \left\| \left(\boldsymbol{\Phi}^T\right)_j \right\|_1, \tag{13}$$

subject to

$$\boldsymbol{S}^T \left(\boldsymbol{\Phi}^T\right)_j = \left(\boldsymbol{Y}^T\right)_j,$$

where $\|(\boldsymbol{\Phi}^T)_j\|_1$ is the l-1 norm of the vector $(\boldsymbol{\Phi}^T)_j$, i.e., the sum of the absolute values of entries in vector $(\boldsymbol{\Phi}^T)_j$. Obviously, the compress matrix $\boldsymbol{\Phi}$ projects the original signal $\boldsymbol{s}_i \in \mathbb{R}^{N \times 1}$ to a much smaller dimensional signal $\boldsymbol{\Phi s}_i \in \mathbb{R}^{M \times 1}$. Instead of dealing with the original signal, we only use $\boldsymbol{\Phi s}_i \in \mathbb{R}^{M \times 1}$ and $\boldsymbol{\Phi\Phi}^T \in \mathbb{R}^{M \times M}$ in the subtyping procedure, leading to a fast classification.

Determination of feature vector

We need to select significant features to represent the original data before we classify the data. For each sample, we extracted five feature characteristics [20]: the mean and the standard deviation of each group's standard deviation (*MeanStd, StdStd*), the standard deviation of the means of all the groups (*StdMean*), and the mean and standard deviation of Pearson's linear correlation coefficient (*MeanCorr, StdCorr*) between the samples and their class label vector. Therefore, for the ith sample, we have a five-dimensional feature vector as follows:

$$\boldsymbol{V}_i = \{MeanStd_i, StdStd_i, StdMean_i, MeanCorr_i StdCorr_i\}$$

where i = 1,2,...,N, and N is the number of samples. Each element in the vector $\boldsymbol{V}_i$ has been normalized by its overall maximum value so that its value is between 0 and 1, i.e., $\boldsymbol{V}_i \in [0, 1]$. A number of M informative features were selected by setting the threshold values of $\boldsymbol{V}_i$. If a feature is informative or significant, we expect that the values from different patients within the same subtype are similar while the differences among different subtypes are relatively large. In addition, it is easy to understand that, if the correlation between the feature vector and the class label is high, the feature vector can serve as a significant biomarker to distinguish the subtypes. According to the above analysis, matrix $\boldsymbol{Y}$ in Equation (10) is built by those features with low *MeanStd, StdStd, StdCorr* while high *StdMean, MeanCorr*, which are significant for the classification.

Classifier based on CS

In this particular study of subtyping four types of GBM with miRNA expression and mRNA expression data, we make a hypothesis that the data follow a normal distribution. In other words, the probability density function for the data is

$$p_y(\hat{\boldsymbol{y}} | g = \pi_k) = \left(2\pi\sigma^2\right)^{-\frac{N}{2}} \exp\left(-\frac{\|\hat{\boldsymbol{y}} - \boldsymbol{s}\|_2^2}{2\sigma^2}\right), \tag{14}$$

where $\hat{\boldsymbol{y}} \in \mathbb{R}^N$ is a given observation; $\boldsymbol{s} \in \mathbb{R}^N$ is the mean of a sample; and σ is the standard deviation of the data.

After compressing the original sample, the probability density function (Equation 14) is still Gaussian but with different mean and standard deviation given by [18]

$$p_y(\boldsymbol{y} | g = \pi_k) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\Phi s})^T \left(\sigma^2 \boldsymbol{\Phi\Phi}^T\right)^{-1} (\boldsymbol{y} - \boldsymbol{\Phi s})\right)}{\left|\sigma^2 \boldsymbol{\Phi\Phi}^T\right|^{\frac{1}{2}} (2\pi)^{\frac{N}{2}}} \tag{15}$$

where $\hat{\boldsymbol{y}} \in \mathbb{R}^M$ is a compressed observation; $\boldsymbol{s} \in \mathbb{R}^N$ is a known signal and $\boldsymbol{\Phi}$ is the compress matrix. The MCSD used in this study is constructed by substituting Equation (15) into Equation (9) for maximum likelihood estimation.

The classification algorithm is described as below.

1. Inputs: training dataset and testing dataset
2. Normalize the rows of the training and the testing datasets to the range of [0,1]
3. Select informative features according to the feature selection criteria
4. Calculate compress matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ by the training dataset by Equation (14)
5. Identify the class of the compressed testing data by Equation (9), where the probability density function is given by Equation (15).

There are many other classifiers such as SVM that can be used. But our purpose here is to show that dimension reduction with the CS can improve subsequent classification and the often used Bayesian classifier is chosen.

Results

We subtyped four types of GBM with multiple genomic data types (e.g., miRNA expression, mRNA expression, and their combinations) from TCGA. The MCSD was first trained by the training data with known class labels, and was then employed to detect subtypes in another independent testing dataset. The classification accuracy by the MCSD was compared with that without using MCSD. The classification performance between using the combined data types and using a single type of data was also compared.

Table 2 shows the comparison of the GBM classification accuracy for the testing dataset, with and without the compress matrix used in our algorithm (see Section "Methods and materials"). The results were obtained on three types of data, i.e., miRNA expression data, mRNA expression data, and their combinations. The classification accuracy is defined as the ratio between the number of correctly labeled samples and the number of total samples. The result calculated by the non-compressed detector had a classification accuracy of 41.8% with miRNA expression data. However, when we used the MCSD to classify the four subtypes, the accuracy of classifying the testing dataset was 69.1%, with 54 selected informative features out of 1,510 features. When we tested the classifiers on the mRNA expression data, the result calculated by the non-compressed detector was 32.7%. However, the classification result with the MCSD was 52.7%, which employed a subset of the features, 432 out of 22,277 features.

Table 2 Comparison of classification accuracy between MCSD and non-compressed detector using combined and single data type

	MCSD		Non-compressed accuracy (%)
	Accuracy (%)	Number of features	
Combined analysis	90.9	121	32.7
miRNA	69.1	54	41.8
Gene expression	52.7	432	32.7

We also tested if the classification performance of the MCSD was better than non-compressed detector in the combined analysis of both miRNA expression and mRNA expression data as shown in Table 2. The subtyping accuracy by the non-compressed detector was 32.7%. The classification accuracy by the MCSD showed a significant improvement over the non-compressed detector. The accuracy was 90.9% (121 informative features selected or 145 informative features selected). The 121 features selected are shown in Additional file 2 with the probes and the corresponding symbols.

Figure 2 demonstrates the classification accuracy when different numbers of informative features were employed. The combined analysis of the two types of genome-wide data was always able to achieve a significant higher subtyping accuracy than any single data type analysis when the same number of informative features were used (with a subset of features less than 450), indicating the advantages of the combined analysis. Figure 2 also shows that the classification accuracy was low when only a few features were used, indicating that the subset was too small to represent the characteristics of the entire dataset. When we increased the number of features used in the MCSD, the classification accuracy went up. The accuracy of classifying the testing dataset reached the highest value, 69.1, 52.7, and 90.9% on the miRNA expression, mRNA expression, and their combinations, respectively. However, more features may also add redundancy and thus cause the decrease of the classification accuracy. Therefore, we conclude that the use of

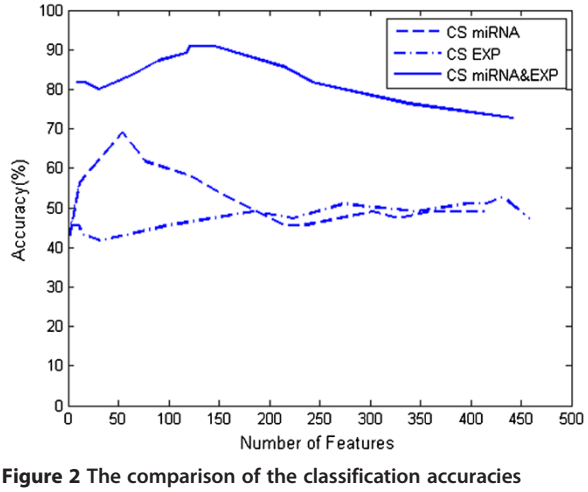

Figure 2 The comparison of the classification accuracies between the combined analysis and the single data type analysis. All of them employed MCSD method to subtype four types of GBM. Note that a significant improvement of the classification accuracy has been achieved by using the combined analysis.

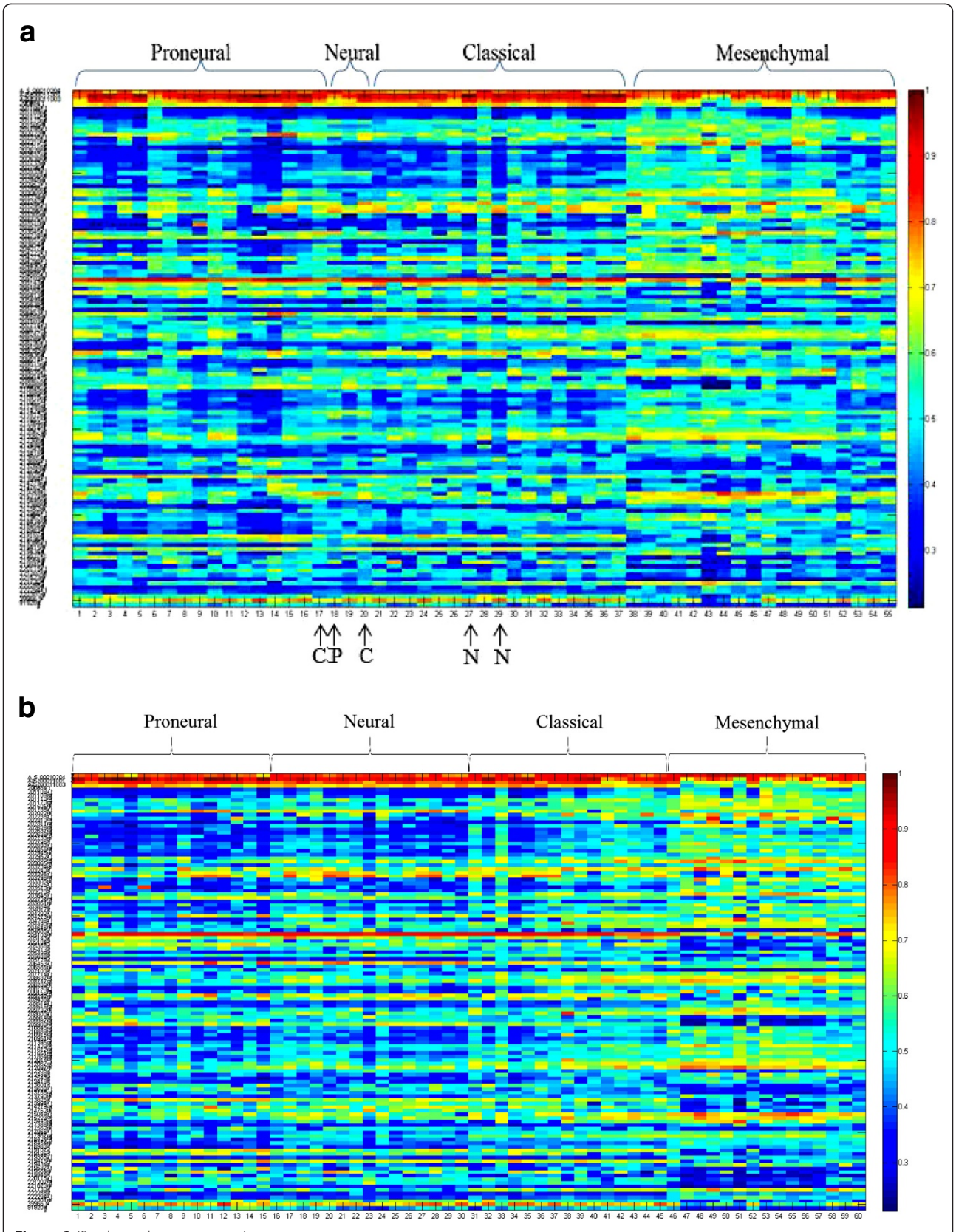

Figure 3 (See legend on next page.)

(See figure on previous page.)
Figure 3 Display of the selected features in distinguishing the four subtypes of GBM, i.e., pro-neural (P), neural (N), classical (C), and mesenchymal (M) for the testing dataset (a) and the training dataset (b). 121 features (3 miRNA expression probes on the top and followed by 118 mRNA expression probes) were chosen from both miRNA expression and mRNA expression data. Each row represents a feature and each column represents a sample/patient. Each feature is normalized by the largest value in each row. The samples with arrows were misclassified to the subtypes as denoted by the arrow.

fewer but significant features will achieve better classification accuracy.

Figure 3a displays the normalized levels of the 121 selected features (118 mRNAs and 3 miRNAs) from both miRNA expression and mRNA expression data for the combined analysis, with the highest classification accuracy of 90.9%. If using the mRNA and miRNA data separately, they only give the accuracy as 49.1 and 47.3%, respectively. The samples with arrows were misclassified to the subtypes pointed by the arrows (e.g., the 17th sample that belongs to pro-neural was misclassified to classical). Each column represents a patient/sample and each row represents a feature (a probe from miRNA expression or mRNA expression data). The four subtypes of GBM are pro-neural, neural, classical, and mesenchymal. Each feature was normalized by the largest value in each row. It can be found that the misclassification only happens among the subtypes of pro-neural, neural, and classical. The number of misclassified samples in each subtype is one sample in pro-neural, two samples in neural, two samples in classical, and zero samples in mesenchymal. The expression levels in the subtype mesenchymal exhibit a significant difference from other three subtypes as shown in Figure 3a. Figure 3b displays the same selected features in the training dataset.

Conclusion and discussion

In this study, we applied the proposed MCSD to subtype four types of GBM: pro-neural, neural, classical, and mesenchymal with multiple genetic data from TCGA. High classification accuracy was achieved by using CS-based technique (i.e., MCSD) along with the combination of multiple datasets. The results from combining two types of genomic data were compared with those from single type of data. Moreover, the performance of the classification with and without MCSD technique had also been compared. The comparisons showed that the CS-based combined analysis of multiple types of genetic data could significantly improve the accuracy of detecting GBM subtypes.

Combining different types of genomic data allows us to interpret the information in the datasets comprehensively. The information from miRNA and mRNA are complementary to each other; so a combined analysis can give a better result than single data type analysis. miRNAs are a recently discovered class of small non-coding RNAs that regulate gene expression [25], which can be combined with mRNA data for better disease subtyping. However, if no dimension reduction with CS was applied, we found from Table 2 that the classification accuracy from combined analysis was comparable to that from the single mRNA expression because of the redundancy added. The classification performance was significantly improved after we used CS method, indicating that CS may reduce redundancy [26] in the combined datasets and thus improve the classification accuracy.

Informative features/biomarkers selected in this study have also been validated to be associated with GBM and have been reported in the literatures. In the combined data analysis, the 121 features/probes selected (shown in Additional file 2), the 3 miRNA expression probes and 118 mRNA expression probes are listed. Two of the selected miRNAs probes that represent the same miRNA, "hsa-miR-9" (sequence "TCATACAGCTAGATAACCAA"), have been validated to have stemness potential and chemoresistance to GBM cells [27-29], and known to be specifically expressed during brain neurogenesis. In the listed mRNA expression probes, the four probes of "CD44" and the three probes of "ASCL1" are selected. Both of the genes have been validated as biomarkers in subtyping GBM in multiple genomic studies [9,30-32]. It demonstrates the significance of "CD44" and "ASCL1" in discriminating different subtypes of GBM. The three probes from "THBS1" are also selected in the 121 probes list. "THBS1" is a subunit of a disulfide-linked homotrimeric protein. This protein has been shown to play roles in platelet aggregation, angiogenesis, and tumorigenesis [33]. "THBS1" is also a major activator of "TGFB1" and the "TGFB1" expression is associated with GBM [34]. Moreover, it has been found that "TbRII", a receptor of "TGFB1", has a strong relationship with human malignant glioblastoma cells [35]. There are biomarkers listed in Additional file 2 that have not been reported yet. However, they may be potential biomarkers for GBM, deserving further study.

We also performed Gene Ontology (GO) analyses to determine that these genes were enriched in specific GO terms (biological processes). The GO term "antigen processing" and presentation "lymphocyte mediated immunity" ($p = 1.78 \times 10^{-6}$), and several GO terms related to wounding healing [e.g. "response to wounding" ($p = 1.26 \times 10^{-8}$); "wound healing" ($p = 2.44 \times 10^{-6}$)], and cell adhesion [e.g. "biological adhesion" ($p = 6.53 \times 10^{-7}$); "cell adhesion" ($p = 6.41 \times 10^{-7}$)] showed highly significant enrichment for our selected genes. These results were expected. Taking

"lymphocyte mediated immunity"-related GO categories as an example, lymphocyte-mediated cellular responses play a critical role in the body's ability to generate an antitumor immune response, and activation status of lymphocytes is an important determinant of sensitivity to tumor-mediated apoptosis [36]. In addition, according to previous studies, the miRNAs we identified are related to glioblastoma. For example, it was found that "has-miR-9" inhibit differentiation of glioblastoma stem cells, and the calmodulin-binding transcription activator 1 (CAMTA1) as "has-miR-9" target is a tumor suppressor in glioblastoma [37].

To test the stability of the classification results, the samples in training and testing were randomly rearranged ten more times. The number of samples from each subtype in training and testing was maintained the same as in the description in the section "Data collection". The overall classification rate has an average value of 87.1% with a standard deviation of 4.5%, indicating that the results are rather robust.

In summary, we have developed a CS-based technique for combining multiple genomic data to subtype glioblastoma more accurately. The biomarkers identified with our approaches have also been validated or reported in some existing literatures, indicating that the integrated approach can provide comprehensive information for better disease diagnosis.

Additional files

Additional file 1: Calculation of *U* for the MCSD.

Additional file 2: List of 121 selected features.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This study was supported by the NIH grant R21 LM010042, NSF Advances in Biological informatics (ABI) grant and The Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning. The authors deeply appreciate the anonymous referees for their valuable suggestions.

Author details

[1]Department of Biomedical Engineering, Tulane University, New Orleans, LA, USA. [2]Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, USA. [3]Center for Systems Biomedicine, Shanghai University for Science and Technology, Shanghai, China.

Received: 28 December 2011 Accepted: 6 December 2012
Published: 14 January 2013

References

1. *Leukemia-Topic Overview*, http://www.webmd.com/cancer/tc/leukemia-topic-overview
2. KY Yeung, RE Bumgarner, AE Raftery, Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics **21**, 2394–2402 (2005)
3. K Seungchan, ER Dougherty, I Shmulevich, KR Hess, SR Hamilton, JM Trent, GN Fuller, W Zhang, Identification of combination gene sets for glioma classification. Mol. Cancer Ther. **1**, 1229–1236 (2002)
4. J Lapointe, C Li, JP Higgins, M Rijn, E Bair, K Montgomery, M Ferrari, L Egevad, W Rayford, U Bergerheim, P Ekman, AM DeMarzo, R Tibshirani, D Botstein, PO Brown, JD Brooks, JR Pollack, Gene expression profiling identifies clinically relevant subtypes of prostate cancer. PNAS **101**, 811–816 (2003)
5. H Ohgaki, P Kleihues, Epidemiology and etiology of gliomas. Acta Neuropathol. **109**, 93–108 (2005)
6. R Benjamin, J Capparella, A Brown, Classification of glioblastoma multiforme in adults by molecular genetics. Cancer J. **9**, 82–90 (2003)
7. CL Nutt, DR Mani, RA Betensky, P Tamayo, JG Cairncross, C Ladd, U Pohl, C Hartmann, ME McLaughlin, TT Batchelor, PM Black, AV Deimling, SL Pomeroy, TR Golub, DN Louis, Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. **63**, 1602–1607 (2003)
8. H Noushmehr, DJ Weisenberger, K Diefes, HS Phillips, K Pujara, BP Berman, F Pan, CE Pelloski, EP Sulman, KP Bhat, RGW Verhaak, KA Hoadley, DN Hayes, CM Perou, HK Schmidt, L Ding, RK Wilson, DV Den Berg, H Shen, H Bengtsson, P Neuvial, LM Cope, J Buckley, JG Herman, SB Baylin, PW Laird, K Aldape, The cancer genome atlas research network. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell **17**, 510–522 (2010)
9. RGW Verhaak, KA Hoadley, E Purdom, V Wang, Y Qi, MD Wilkerson, CR Miller, L Ding, T Golub, JP Mesirov, G Alexe, M Lawrence, MO Kelly, P Tamayo, BA Weir, S Gabriel, W Winckler, S Gupta, L Jakkula, HS Feiler, JG Hodgson, CD James, JN Sarkaria, C Brennan, A Kahn, PT Spellman, RK Wilson, TP Speed, JW Gray, M Meyerson et al., Integrated genomic analysis identifies clinically relevant subtypes of Glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell **17**, 98–110 (2010)
10. I Bentwich, A Avniel, Y Karov, R Aharonov, S Gilad, O Barad, A Barzilai, P Einat, U Einav, E Meiri, E Sharon, Y Spector, Z Bentwich, Identification of hundreds of conserved and nonconserved human microRNAs. Nat. Genet. **37**, 766–770 (2005). doi:10.1038/ng1590
11. P Fasanaro, S Greco, M Ivan, M Capogrossi, F Martelli, MicroRNA: emerging therapeutic targets in acute ischemic diseases. Pharmacol. Ther. **125**, 92–104 (2010). doi:10.1016/j.pharmthera.2009.10.003
12. MV Iorio, M Ferracin, C-G Liu, A Veronese, R Spizzo, S Sabbioni, E Magri, M Pedriali, M Fabbri, M Campiglio, S Ménard, JP Palazzo, A Rosenberg, P Musiani, S Volinia, I Nenci, GA Calin, P Querzoli, M Negrini, CM Croce, MicroRNA gene expression deregulation in human breast cancer. Cancer Res. **65**, 7065–7070 (2005). doi:10.1158/0008-5472.CAN-05-1783
13. JA Bishop, H Benjamin, H Cholakh, A Chajut, DP Clark, WH Westra, Accurate classification of non-small cell lung carcinoma using a novel MicroRNA-based approach. Clin. Cancer Res. **16**, 610–619 (2010)
14. C Soneson, H Lilljebjörn, T Fioretos, M Fontes, Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC Bioinforma **11**, 191 (2010)
15. CB Kingsley, W-L Kuo, D Polikoff, A Berchuck, JW Gray, AN Jain, Magellan: a web based system for the integrated analysis of heterogeneous biological data and annotations; application to DNA copy number and expression data in ovarian cancer. Cancer Inf. **2**, 10–21 (2006)
16. OG Troyanskaya, K Dolinski, AB Owen, RB Altman, D Botstein, A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). PNAS **100**, 8348–8353 (2003)
17. GRG Lanckriet, TD Bie, N Cristianini, MI Jordan, WS Noble, A statistical framework for genomic data fusion. Bioinformatics **20**, 2626–2635 (2004)
18. MA Davenport, MB Wakin, RG Baraniuk, *Detection and estimation with compressive measurements. Technical Report*, 2007
19. EJ Cand`es, MB Wakin, An introduction to compressive sampling. IEEE Signal Process. Mag. **25**(2), 21–30 (2008)
20. W Tang, H Cao, J Duan, Y-P Wang, A compressed sensing based approach for subtyping of leukemia from gene expression data. J. Bioinfo. Comput. Biol. **9**, 631–645 (2011). doi:10.1142/S0219720011005689
21. *TCGA Data*, http://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp
22. DC Edwards, CE Metz, MA Kupinski, Ideal observers and optimal ROC hypersurfaces in N-class classification. IEEE Trans. Med. Imag. **23**, 891–895 (2004)
23. SA Starks, V Kreinovich, Environmentally-oriented processing of multi-spectral satellite images: new challenges for Bayesian methods, in *Proceedings of the 17th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis, Boise, Idaho*, 1998, p. 271

24. B Efron, T Hastie, I Johnstone, R Tibshirani, Least angle regression. Ann. Stat. **32**, 407–451 (2004)
25. S Volinia, GA Calin, C-G Liu, S Ambs, A Cimmino, F Petrocca, R Visone, M Iorio, C Roldo, M Ferracin, RL Prueitt, N Yanaihara, G Lanza, A Scarpa, A Vecchione, M Negrini, CC Harris, CM Croce, A microRNA expression signature of human solid tumors defines cancer gene targets. PNAS **103**, 2257–2261 (2006)
26. Y Tsaig, DL Donoho, Extensions of compressed sensing. Signal Process. **86**, 549–571 (2006)
27. H-M Jeon, Y-W Sohn, S-Y Oh, S-H Kim, S Beck, S Kim, H Kim, ID4 imparts chemoresistance and cancer stemness to glioma cells by derepressing miR-9*-mediated suppression of SOX2. Cancer Res. **71**, 3410–3421 (2011)
28. MH Ko, S Kim, W Hwang, HY Ko, YH Kim, DS Lee, Bioimaging of the unbalanced expression of microRNA9 and microRNA9* during the neuronal differentiation of P19 cells. FEBS J. **275**, 2605–2616 (2008)
29. AS Yoo, BT Staahl, L Chen, GR Crabtree, MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. Nature **460**, 642–646 (2009)
30. Y Liang, M Diehn, N Watson, AW Bollen, KD Aldape, MK Nicholas, KR Lamborn, MS Berger, D Botstein, PO Brown, MA Israel, Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. PNAS **102**, 5814–5819 (2005)
31. A Ariza, D López, JL Mate, M Isamat, E Musulen, M Pujol, A Ley, J Navas-palacios, Role of CD44 in the invasiveness of glioblastoma multiforme and the noninvasiveness of meningioma: an immunohistochemistry study. Hum. Pathol. **26**, 1144–1147 (1995)
32. HS Phillips, S Kharbanda, R Chen, WF Forrest, RH Soriano, TD Wu, A Misra, JM Nigro, H Colman, L Soroceanu, PM Williams, Z Modrusan, BG Feuerstein, K Aldape, Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell **9**, 157–173 (2006)
33. AF Galvez, L Huang, MMJ Magbanua, K Dawson, RL Rodriguez, Differential expression of thrombospondin (THBS1) in tumorigenic and nontumorigenic prostate epithelial cells in response to a chromatin-binding soy peptide. Nutr. Cancer **63**, 623–636 (2011). doi:10.1080/01635581.2011.539312
34. B Lin, A Madan, J-G Yoon, X Fang, X Yan, T-K Kim, D Hwang, L Hood, G Foltz, Massively parallel signature sequencing and bioinformatics analysis identifies up-regulation of TGFBI and SOX4 in human glioblastoma. PLoS One **5**, e10210 (2010)
35. A Wesolowska, M Sliwa, B Kaminska, Development of siRNA against TbRII blocking efficiently TGFb1 signaling pathways in glioma cells. Eur. J. Biochem. **271**, 35–58 (2004). Supplement 1 July: abstract number P2.5-05
36. A Chahlavi, P Rayman, A-L Richmond, K Biswas, R Zhang, M Vogelbaum, C Tannenbaum, G Barnett, J-H Finke, Glioblastomas induce T-lymphocyte death by two distinct pathways involving gangliosides and CD70. Cancer Res. **65**, 5428–5438 (2005)
37. J Gil-Ranedo, M Mendiburu-Elicabe, M Garcia-Villanueva, D Medina, M del Alamo, M Izquierdo, An off-target nucleostemin RNAi inhibits growth in human glioblastoma-derived cancer stem cells. PLoS One **6**, e28753 (2011)

doi:10.1186/1687-4153-2013-2
Cite this article as: Tang *et al.*: **Subtyping glioblastoma by combining miRNA and mRNA expression data using compressed sensing-based approach.** *EURASIP Journal on Bioinformatics and Systems Biology* 2013 **2013**:2.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Immediate publication on acceptance
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com