Published in final edited form as:
Adv Genet Eng. 2012 January 16; 1: 101–. doi:10.4172/2169-0111.1000101.

Subtyping of Gliomaby Combining Gene Expression and CNVs Data Based on a Compressive Sensing Approach

Wenlong Tang[1], **Hongbao Cao**[1], **Ji-Gang Zhang**[2], **Junbo Duan**[1], **Dongdong Lin**[1], and **Yu-Ping Wang**[1,2,*]

[1]Department of Biomedical Engineering, Tulane University, New Orleans, USA

[2]Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, USA

Abstract

It is realized that a combined analysis of different types of genomic measurements tends to give more reliable classification results. However, how to efficiently combine data with different resolutions is challenging. We propose a novel compressed sensing based approach for the combined analysis of gene expression and copy number variants data for the purpose of subtyping six types of Gliomas. Experimental results show that the proposed combined approach can substantially improve the classification accuracy compared to that of using either of individual data type. The proposed approach can be applicable to many other types of genomic data.

Keywords

Gene Expression; CNVs data; Compressive Sensing; Glioma; Classification; Combined Analysis

Introduction

In recent years, the development of bio-techniques allows researchers to collect different types of data from an experiment, such as gene expression data, SNP data, and Copy Number Variations (CNVs) data. A better result could be generated based on combining multiple types of data than using any individual data. Combined analysis with different data types of genome-wide measurements is not a new concept, but how to combine them efficiently for biological discovery is always challenging. A web based platform, called Magellan, was developed for the integrated analysis of DNA copy number and expression data in ovarian cancer [1]. The significant correlation between gene expression and patient survival has been found by Magellan. Troyanskaya et al. [2] developed a Bayesian framework to combine heterogeneous data sources for predicting gene function. Improved accuracy of the gene groupings has been achieved compared with microarray analysis alone. Kernel-based statistical learning algorithms were also used in the combine analysis of multiple genome-wide datasets [3]. Some combined analysis methods need the datasets to have the same distribution [4]; one has to transform the datasets to be the same distribution

[*]**Corresponding author:** Department of Biomedical Engineering, School of Science and Engineering, Tulane University, 534 Lindy Boggs Building, New Orleans, LA 70118, USA, Tel: 504-865-5867, wyp@tulane.edu.

before the analysis. Recently, an integrative approach combining linkage, gene expression, and association has been reported to identify candidate genes regulating BMD [5]. The combined analysis approach proposed in this work has no specific requirement for the data types or data distributions. In order to test the effectiveness of our approach, we applied it to the subtyping of gilomas.

Gliomas are tumors that start in the brain or spine and arise from glial cells [6]. Gliomas are the most common type of primary brain tumors in adults [7]. The classification of gliomas can be based on cell type, grade and location. For instance, gliomas can be classified into low-grade and high-grade determined by pathologic evaluation of the tumor. In this study, we define the subtypes based on genetic and molecular signatures according to the reference [7].

The classification of glioma subtypes has attracted a lot of attentions and has been investigated by many research groups. Most of the works have been based on gene expression data. It was reported that four subtypes of gliomas, oligodendroglioma, anaplastic oligodendroglioma, anaplastic astrocytoma and glioblastomamultiforme, can be distinguished by only two-gene or three-gene combinations [8]. Nutt et al. [9] built a k-nearest model with 20 features to classify 28 glioblastomas and 22 anaplastic oligodendrogliomas. It was claimed that class distinctions according to the model were significantly associated with survival outcome (P=0.05). Chakraborty et al. [10] considered several Bayesian classification methods to classify gliomas with gene expression data. A Bayesian variable selection scheme was also proposed for gene selection. Noushmehr et al. [12] found a distinct subset of samples in The Cancer Genome Atlas (TCGA) glioma samples displaying concerted hypermethylation at a large number of loci. They took it as evidence that a glioma-CpG island methylator phenotype exists. Verhaak et al. [12] classified glioma into four subtypes: Proneural, Neural, Classical, and Messenchymal, based on gene expression data. MRI data have also been used in the classification of gliomas [13,14]. However, to the authors' best knowledge; few researchers have combined two or more than two types of data to improve the gliomas classification.

Therefore, a novel approach that can combine multiple data sets is needed for improved classification. Compressed Sensing (CS), also called compressive sampling, has been developed recently in statistics and signal processing, and becomes a powerful tool for data analysis. We recently used CS method to classify chromosomes from Multicolor Fluorescence In-Situ Hybridization (M-FISH) images [15], as well as integrated analysis of copy number data and gene expression data for identifying gene groups susceptible to cancers [16]. In these studies, we demonstrated the advantages of the CS methods in compact representation of combined genomic data, resulting in higher classification accuracy.

The work described in this work is to develop a CS based integration and classification methods and apply them to identify the subtyping of gliomas. The results demonstrate that the proposed methods can significantly improve the classification accuracy of gliomas compared to individual gene expression or CNVs data analysis.

Data Collection

The data in this study is publicly available from the website of National Cancer Institute (https://caintegrator.nci.nih.gov/rembrandt/home.do). Two unsupervised methods had been used to analyze the six glioma subtypes based on the gene expression data of the patients [7]. In our study, we classify the six Glioma subtypes by integrated analysis of both gene expression and CNVs data. The overview of the six hierarchically nested subtypes of gliomas is shown in Figure 1.

We collected a dataset that has 56 samples (patients) with both gene expression data (54675 genes for each sample) and CNV data (758 probes for each sample). Eight samples belong to the Oligodendroglioma-rich (O) main type that has 4 OAs and 4 OBs. For the rest 48 samples, Glioblastoma-rich (G), we have 27 GAs (10 GA1s and 17 GA2s) and 21 GBs (13 GB1s and 8 GB2s).

Methods

According to the structure of the six subtypes described in Figure 1, we used divisive (TOP-DOWN) algorithm to subtype each of the 6 classes. At the top level, the data are classified into two main types O and G; then those two subtypes are further classified, until 6 subtypes are obtained at the bottom level. Sparse Representation Clustering (SRC) method proposed by usis applied to select Informative Variables (IVs) (genes or probes) for the subtyping of gliomas in the analysis. SRC algorithm was obtained from Compressed Sensing (CS) theory, which aims to approximate a sparse solution of $y = \boldsymbol{A}x$ in a given underdetermined matrix $\boldsymbol{A}$.

Feature selection

To distinguish the two groups (*e.g.*, O and G), it is helpful to extract significant features from the overall gene expression and CNVs data, respectively. For each gene or probe, we extracted 4 features: the standard deviation of each group (Std_1 and Std_2), the absolute value of the mean difference of the two groups (*MD*), and the Pearson's linear correlation coefficient (*Corr*). Thus for *i-th* variable, we have a 4 dimensional feature vector as follows:

$$V_i = \{Std_{i1}, Std_{i2}, MD_i, Corr_i\} \in \mathbb{R}^4 \quad (1)$$

Where $i=1,2,\ldots N$, and N is the number of genes/CNVs. Each feature is normalized by its overall maximum value so that each entry of $V_i \in [0,1]$. A number of *M*IVs can be selected accordingly, yielding $M \ll N$. The detail of the feature selection can be found in reference [17]. After the normalization, we get the feature dataset as the input of SRC algorithm for the selection of significant genes with small Std_1 and Std_2 , high *MD* and *Corr*.

SRC algorithm

In CS theory, if a signal $x \in \mathbb{R}^n$ is sparse, it could be recovered stably by its measurements $\boldsymbol{A}x = y$. This can be formulated as solving the following optimization problem:

$$(P0) \quad \hat{x}=argmin\|x\|_0, \text{ subject to } \boldsymbol{A}x=y \quad (2)$$

Where $\|x\|_0$ is *l*-0 norm, $y \in \mathbb{R}^k$, $k \ll n$, $\boldsymbol{A} \in \mathbb{R}^{k\times n}$. This is an NP hard problem by traversing all possible entries for *x*. The *l*-1normis used instead by minimizing the nonzero numbers, which can be considered as a linear programming problem:

$$(P1) \quad \hat{x}=argmin\|x\|_1, \text{ subject to } \boldsymbol{A}x=y \quad (3)$$

Where $\|x\|_1=\sum_{i-1}^{n}|x|$ The solution path of this problem has a piecewise-linear-property [18], and can be solved with *k*-steps when *x* is sparse enough, and ***A*** is under certain condition [19,20].

The basic problem in SRC is to use labelled training samples (included in ***A***) from distinct classes to correctly determine the class to which a new test sample belongs. If we design $\boldsymbol{A}=\{a_i\} \in \mathbb{R}^{k\times n}$, is a positive unite vector with $\|a_i\|=1$, a new unclassified sample $y \in \mathbb{R}^k$ will result in an estimate of sparse solution $\hat{x}$, whose non-zero entries correspond to a particular cluster.

Sparse Representation-based Clustering (SRC):

1. Input characteristic matrix ***A*** with vectors of sdifferent clusters and a test sample *y* $\in R^{k\times 1}$.
2. Normalize the columns of ***A*** to have unit *l*-2 norm.
3. Solve the *l*-1 norm minimization problem (P1) defined by Equation (3).
4. Calculate the vector angle $\theta_i(y, \boldsymbol{A}\delta i\ (x))$, $i\in\{1,2,\ldots, s\}$, where $\delta_i\ (x)$ is a mask function that maps x to a sparse vector, with non-zero entries in the *i*-th group.
5. *Identity*(*y*) = arg $\min_i\ (\theta_i)$

The details of the SRC algorithm can be found in reference [16].

Transformation matrix design

To design a transformation matrix ***A***, several methods have been proposed [19], such as incoherent matrices, random projection matrices, *etc*. In this study, we propose a method of designing ***A*** by considering all possible classes in a subtyping work.

If *m* number of features is used for clustering, there will be $c = 2^m - 1$ possible groups, with characteristic matrix $\boldsymbol{A} = \{\boldsymbol{A}_k\}$, and $k = 1,\ldots,c$. We label each group with a column vector $\boldsymbol{V}_k \in R^{m\times 1}$ being given a binary value, designating different combinations of 1 and 0. Then we design characteristic matrix of the *k-th* group $\boldsymbol{A}_k = \{\boldsymbol{V}_{ki}\} \in R^{m\times n_i}$, where $i = 1,\ldots, n_i$, and $n_i > m$; $\boldsymbol{V}_{k_i}=\boldsymbol{V}_k+\boldsymbol{V}_0$, and $\boldsymbol{V}_0$ is a random vector with small amplitude; $\boldsymbol{V}_{k_i}$ and $\boldsymbol{V}_k$ have the

relation of $\theta(V_{k_i}, V_k) < \frac{1}{2} min_{k \neq j} \theta(V_k, V_j)$, which guarantees that each column vector is corresponding to its groups only.

To guarantee a sample vector belonging to the *k* -th group be represented by characteristic matrix $\boldsymbol{A}_k = \{\boldsymbol{V}_{ki}\}$, it requires that $rank(\boldsymbol{A}_k = m$, where .In addition to the requirements mentioned above, a valid *v* for the SRC based classifier should have a sparse solution *x* whose non-zero entries concentrate mostly on one group, while that of an invalid vector with non-zero entries spread evenly over all groups. To quantify this observation, the Sparsity Concentration Index (SCI) [21] shown in Eq. (4) is introduced to validate $\boldsymbol{A}$ to measure how concentrated the feature vectors are on a particular class in the dataset.

$$SCI(x) = \frac{s \times max_i \frac{\|\delta_i(x)\|_1}{\|x\|_1} - 1}{s - 1} \in [0, 1], \quad (4)$$

Wheres is the number of classes, is a mask function that that maps x to a sparse vector, with non-zero entries in the *i*-th group. For a solution $\hat{x}$ found by the SRC algorithm, if $SCI(\hat{x}) = 1$, the feature vector *y* is represented using vectors only from a single class; if $SCI(\hat{x}) = 0$, the sparse coefficients are spread evenly over all classes. We choose a threshold $\tau \in [0,1]$ and accept a vector as valid if $SCI(\hat{x}) > \tau$; otherwise, reject it as invalid.

Compressed sensing based classifier

The training of transformation matrix can be formulated as a sparse representation problem as shown in Eq. (5),

$$\boldsymbol{Y} = \varnothing \boldsymbol{S} + \boldsymbol{n}_0 \quad (5)$$

Where $\boldsymbol{Y} = \{\boldsymbol{y}_i\} \in \mathbb{R}^{M \times c}$ are the gene expressions of selected genes for the total samples/patients; $n_0 \tilde{\mathcal{N}}\left(0, \sigma_0^2 I_N\right)$ is *i.i.d.* Gaussian noise; $\mathbf{S} = \{\boldsymbol{s}_i\} \in \mathbb{R}^{N \times c}$ are the gene expressions of all the genes for the total samples/patients, and $M \ll N$. The matrix $\varnothing \in \mathbb{R}^{M \times N}$ is a sparse transformation matrix. The linear system given by (5) is an underdetermined sparse system, which can be solved by using L-1 norm minimization algorithm.

A CS based classifier is developed to classify the glioma subtypes. To testify whether a given vector belongs $\boldsymbol{y} \in \mathbb{R}^M$ to a known signal $\boldsymbol{s}_i \in \mathbb{R}^N$ or not, we set the hypothesis as follows [22]:

$$\begin{aligned} \tilde{H}_0 &: \boldsymbol{y} = \varnothing \boldsymbol{n}, \\ \tilde{H}_1 &: \boldsymbol{y} = \varnothing (\boldsymbol{s}_i + \boldsymbol{n}) \end{aligned} \quad (6)$$

From (6), we have $\boldsymbol{y} \sim$ N (0, $\sigma^2 \varnothing^T$) under $\tilde{H}_0$, $\boldsymbol{y} \sim$ N ($\varnothing \mathbf{s}_i$, $\sigma^2 \varnothing \varnothing^T$)under $\tilde{H}_1$, which gives:

$$f_0(y) = \frac{exp\left(-\frac{1}{2}\mathbf{y}^T\left(\sigma^2 \varnothing\varnothing^T\right)^{-1}\mathbf{y}\right)}{\left|\sigma^2\varnothing\varnothing^T\right|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \quad (7)$$

and

$$f_1(y) = \frac{exp\left(-\frac{1}{2}(\boldsymbol{y}-\varnothing\boldsymbol{s}_i)^T\left(\sigma^2\varnothing\varnothing^T\right)^{-1}(\boldsymbol{y}-\varnothing\boldsymbol{s}_i)\right)}{\left|\sigma^2\varnothing\varnothing^T\right|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \quad (8)$$

Thus, the likelihood ratio test is: if $\frac{f_1(y)}{f_0(y)} < 1$, $\boldsymbol{y}$ is under $\tilde{H}_0$; otherwise, $\boldsymbol{y}$ is under $\tilde{H}_1$. The likelihood ratio test can be simplified by taking a logarithmand the compressive classification of $\boldsymbol{y} \in \mathbb{R}^M$ can be derived as follows.

Define compressive detector $\tilde{\boldsymbol{t}}$ as:

$$\tilde{\boldsymbol{t}} := \boldsymbol{y}^T\left(\varnothing\varnothing^T\right)^{-1}\varnothing\boldsymbol{S} \quad (9)$$

Where $\tilde{\boldsymbol{t}} = \{\tilde{t}_i\} \in \mathbb{R}^{1\times c}$, $\boldsymbol{S} = \{s_i\} \in \mathbb{R}^{N\times c}$, i=1,2, …, c.

It has been proven by reference [18] that under the condition of $\tilde{H}_0$:

$$\tilde{\boldsymbol{t}}_i \mathbb{N}\left(0, \quad \sigma^2 \boldsymbol{s}_i^T \varnothing^T\left(\varnothing\varnothing^T\right)^{-1}\varnothing\boldsymbol{s}_i\right) \quad (10)$$

While under the condition of :

$$\tilde{\boldsymbol{t}}_i \tilde{\mathbb{N}}\left(\boldsymbol{s}_i^T\varnothing^T\left(\varnothing\varnothing^T\right)^{-1}\varnothing\boldsymbol{s}_i, \quad \sigma^2\boldsymbol{s}_i^T\varnothing^T\left(\varnothing\varnothing^T\right)^{-1}\varnothing\boldsymbol{s}_i\right). \quad (11)$$

We then calculate the differences of the standard score of $\tilde{t}_i$ $\left(ds\tilde{t}_i\right)$ under the two conditions:

$$ds\tilde{t}_i = \frac{|\tilde{t}_i|}{\sigma_i} - \frac{|\tilde{t}_i - \mu_i|}{\sigma_i} \quad (12)$$

Where $\sigma_i = \left(\sigma^2 s_i^T \varnothing^T\left(\varnothing\varnothing^T\right)^{-1}\varnothing s_i\right)^{1/2}$, and $\mu_i = s_i^T\varnothing^T\left(\varnothing\varnothing^T\right)^{-1}\varnothing s_i$.

We assign a class ID label to the vector $\boldsymbol{y}$:

$$\text{Identity } (\boldsymbol{y}) = argmax_i\left(ds\tilde{t}_i\right) \quad (13)$$

If *Identity*(***y***) is within the range of 1to c_1, ***y*** belongs to class1; otherwise, ***y*** belongs to class 2. Obviously, our proposed approach can be extended to the classification of multiple classes.

It can be seen that by introducing the sparse transformation matrix $\varnothing$, we projected the original signal $s_i \in \mathbb{R}^{N\times 1}$ to a very smaller dimensional signal $\varnothing s_i \in \mathbb{R}^{M\times 1}$. In the following process, instead of dealing with the original signal, we only used $\varnothing s_i \in \mathbb{R}^{M\times 1}$ and $\varnothing\varnothing^T \in \mathbb{R}^{M\times M}$ in the construction of the compressive detector $\tilde{\boldsymbol{t}}$ and calculation of σ_i and μ_i , leading to a fast classification.

Cross validation and experiment design

A cross validation method, Leave One Out (LOO) [23], is widely used in evaluating the detection accuracy of different classes of subjects. It was employed here to evaluate the efficiency of feature selection and the performances of compressive detector. To find the best LOO accuracy for each subtyping, we calculated the classification accuracy by LOO, based on from 5 to 200 IVs, in three cases: subtyping based on gene expression data, CNVs data and their combinations.

Results

The SRC approach was used to select different numbers of IVs, while the CS based classifier was employed to classify the subtypes of gliomas. Finally, the classification accuracy was calculated by the LOO method. Figure 2 plots the classification accuracies of O and G on the top level of the hierarchical structure for suptyping gliomas. Three results are compared in the Figure 2: classification accuracies calculated by CNVs data, by gene expression data and by the combination of the two types of data. The combined result was calculated by fixing the number of IVs from CNVs data as the one that achieves the highest accuracy and iterating the number of IVs from gene expression data from 5 to 200 genes. In this specific case, combined analysis doesn't show any significant advantages compared to the gene expression analysis only.

On the second level of the subtyping, the results shown in Figure 3 are for the classification of GA and GB. The performance of the combined analysis is obviously much better than either individual analysis. The highest classification accuracy of the combined analysis achieves 77.1%, which is higher than 70.8% from the gene expression. In Figure 4, the combined classification rate can beas high as 100% for the subtyping of OA and OB compared to the highest classification rate of 87.5% from the gene expression data, individully.

On the bottom level, in Figure 5, it can be seenthat the combined data analysis for clasifying GA1 and GA2 has the same highest classification rate, 85.2%, as the individual analysis of gene expression data, but with less IVs.The combined analysis used only 15 IVs to achieve the highest classification accuracy, 50 IVs less compared to the individual gene expression analysis. Figure 6 shows the comparison of the combined analysis and individual analysis for the classification of GB1 and GB2 subtypes. It is shown that the combined analysis

yields higher classification accuracy (90.5%) than either individual analysis, 81.0% for gene expression and 76.2% for CNVs.

It can be found that combined analysis performs better than either individual analysis in the classification of OA and OB, GA and GB, GB1 and GB2 in Figure 7. For the other two classification, O and G, GA1 and GA2, the combined method has comparable classification rate with the gene expression data analysis. The classification rate based on the CNVs data is the lowest.

Conclusion and Discussion

This paper proposed a CS based method to subtype gliomas by combining gene expression data and CNVs data from the same subject of patients. Experiments have been performed to compare the results of combined analysis with individual analysis. The calculation results show that the combined analysis achieves a significant improvement of the classification accuracy than using the individual analysis except for the subtyping of O and G types of gliomas.

The different types of genomic data used in this study, e.g., gene expression data and CNVs data have different resolutions and provide different information. Gene expression reveals how the functions of genes change, while CNVs indicate where these functional changes occur in the genome, Thus, the information from different sources could be complementary , which can be used to improve the accuracy of classifying diseases. In the human genome, about half of the detected CNVs overlap with regions which code proteins [24]. Therefore, CNV loci encompassing genes may potentially affect gene expression and subsequently relevant phenotypes [25,26]. They exert their influence by modifying the expression of genes mapping within and close to the rearranged region [27]. The information both from CNV and gene expression shed light on the pathogenesis underlying the complex diseases.The proposed integrated data analysis approach provided an appealing solution to subtype gliomas.

The sample size of the Oligodendroglioma-rich (O) is relatively small (4 OAs and 4 OBs). That could influence the reliability of the result. That is probably the reason why we cannot obtain improved accuracy classification of O and G, even with combined data analysis. It could also explain why the classification results in Figure 4 oscillate, with the accuracies up and down. Those problems could be avoided by increasing the sample size of O subtype. In summary, the combined analysis method proposed in this work provides an improved way of subtyping gliomas than using an individual data. It has the potential to improve the diagnostic accuracy in the clinical practice.

Acknowledgments

This work has been supported by the NIH grant R21 LM010042 and NSF grant. The authors thank Dr. Aiguo Li and Dr. Howard A. Fine from National Cancer Institute for their great help.

References

1. Kingsley CB, Kuo W-L, Polikoff D, Berchuck A, Gray JW, et al. 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 Inform. 2007; 2:10–21. [PubMed: 19458754]
2. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). ProcNatlAcadSci U S A. 2003; 100:8348–8353.
3. Lanckriet GR, De Bie T, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics. 2004; 20:2626–2635. [PubMed: 15130933]
4. Berger JA, Hautaniemi S, Mitra SK, Astola J. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. IEEE/ACM Trans ComputBiolBioinform. 2006; 3:2–16.
5. Farber CR, van Nas A, Ghazalpour A, Aten JE, Doss S, et al. An integrative genetics approach to identify candidate genes regulating BMD: combining linkage, gene expression, and association. J Bone Miner Res. 2009; 24:105–116. [PubMed: 18767929]
6. Mamelak AN, Jacoby DB. Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601). Expert Opin Drug Deliv. 2007; 4:175–186. [PubMed: 17335414]
7. Li A, Walling J, Ahn S, Kotliarov Y, Su Q, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res. 2009; 69:2091–2099. [PubMed: 19244127]
8. Kim S, Dougherty ER, Shmulevich I, Hess KR, Hamilton SR, et al. Identification of combination gene sets for glioma classification. Mol Cancer Ther. 2002; 1:1229–1236. [PubMed: 12479704]
9. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 2003; 63:1602–1607. [PubMed: 12670911]
10. Chakraborty S, Mallick BK, Ghosh D, Ghosh M, Dougherty E. Gene Expression-Based Glioma Classification Using Hierarchical Bayesian Vector Machines. Sankhya. 2007; 69:514–547.
11. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell. 2010; 17:510–522. [PubMed: 20399149]
12. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17:98–110. [PubMed: 20129251]
13. Guillamo JS, Monjour A, Taillandier L, Devaux B, Varlet P, et al. Brainstem gliomas in adults: prognostic factors and classification. Brain. 2001; 124:2528–2539. [PubMed: 11701605]
14. Morris M, Greiner R, Sander J, Murtha A, Schmidt M. Learning a classification-based glioma growth model using MRI data. Journal of Computers. 2006; 1:21–31.
15. Cao, H.; Wang, Y-P. M-Fish image analysis with improived adaptive fuzzy C-Means clustering based segmentation and sparse representation classification. 3rd International Conference on Bioinformatics and Computational Biology (BICoB); New Orleans, Louisiana, USA. 2011.
16. Cao, H.; Wang, Y-P. Integrated analysis of gene expression and copy number data using sparse representation based clustering model. 3rd International Conference on Bioinformatics and Computational Biology (BICoB); New Orleans, Louisiana, USA. 2011.
17. Tang W, Cao H, Duan J, Wang YP. A compressed sensing based approach for subtyping of leukemia from gene expression data. J BioinformComputBiol. 2011; 9:631–645.
18. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Statist Soc B. 1996; 58:267–288.
19. Donoho DL, Tsaig Y. Fast solution of l1-norm minimization problems when the solution may be sparse. IEEE Trans Inf Theory. 2008; 54:4789–4812.
20. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Statist. 2004; 32:407–499.

21. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell. 2009; 31:210–227. [PubMed: 19110489]

22. Davenport MA, Wakin MB, Baraniuk RG. Detection and estimation with compressive measurements. Technical Report. Jan 24.2007

23. Efron, B.; Tibshirani, RJ. An introduction to the bootstrap. Chapman and Hall; New York: 1993.

24. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, et al. Large-scale copy number polymorphism in the human genome. Science. 2004; 305:525–528. [PubMed: 15273396]

25. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science. 2007; 315:848–853. [PubMed: 17289997]

26. Schuster-Böckler B, Conrad D, Bateman A. Dosage sensitivity shapes the evolution of copy-number varied regions. PLoS One. 2010; 5:9474.

27. Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet. 2009; 18:1–8. [PubMed: 18815198]

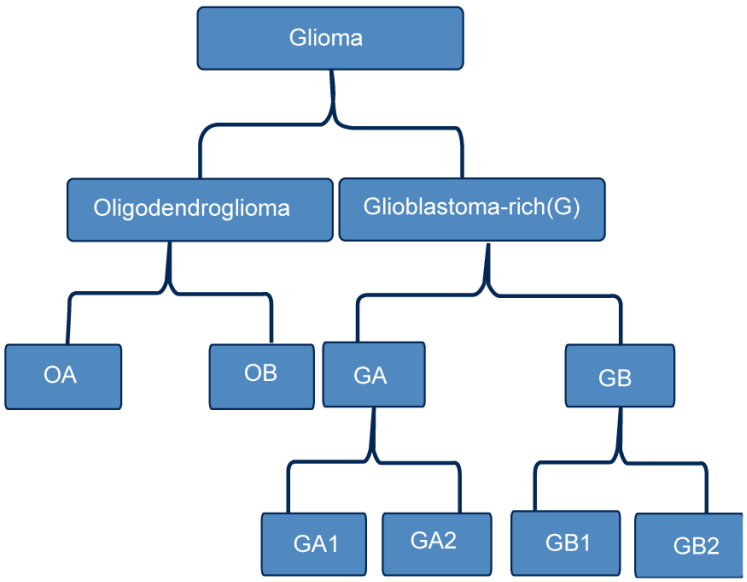

Figure 1.
The hierarchical structure of the six subtypes of gliomas.

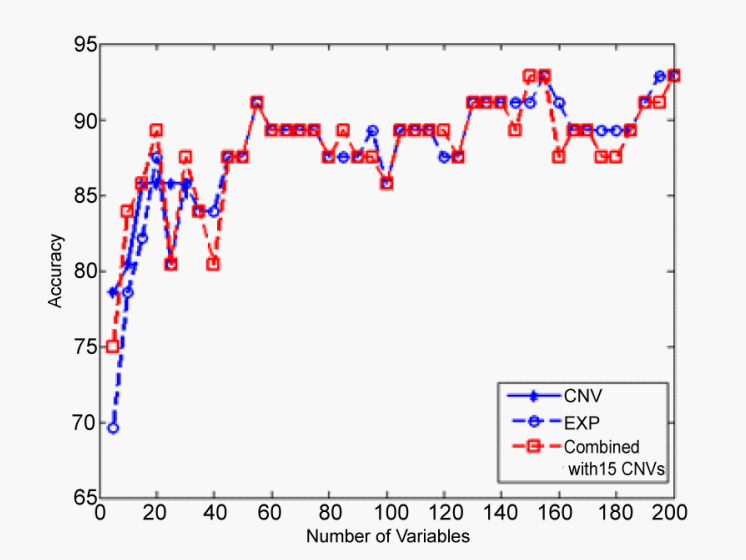

Figure 2.

The classification accuracies of O and G subtypes by using gene expression data only (circle) and the combined data (square), corresponding to different numbers of IVs from 5 to 200. For the CNVs data (star), the maximum number of IVs that can be reached is 30 due to the limitation of sample size. In this specific case, combined analysis doesn't show any significant advantages compared to the gene expression analysis only.

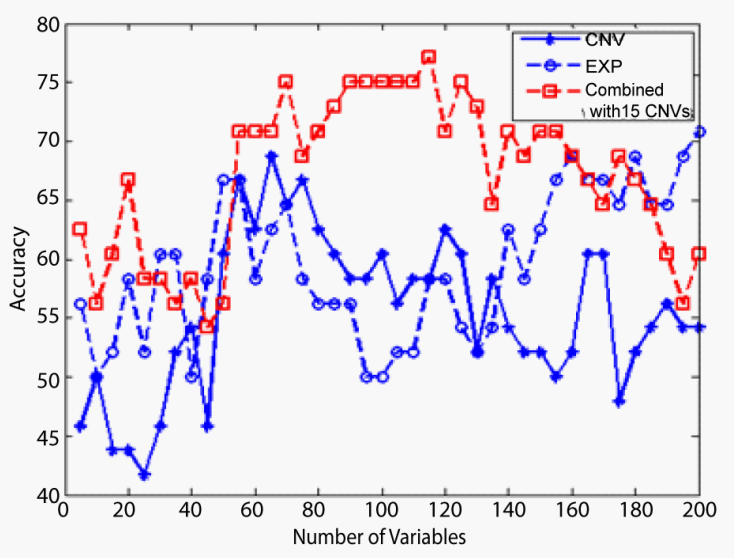

Figure 3.

The classification accuracies of GA and GB subtypes by using CNVs data (star), gene expression data (circle) and the combined data of the two (square), corresponding to different numbers of IVs from 5 to 200. Note that the combined analysis can reach higher accuracies than either individual analysis.

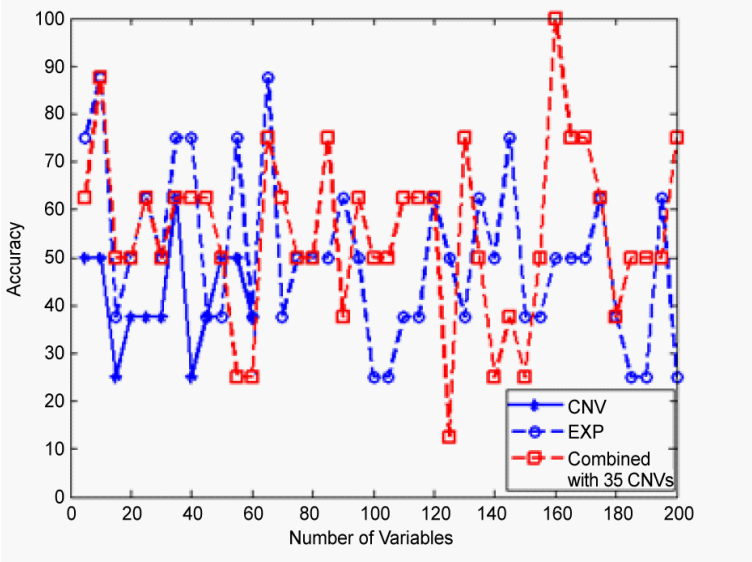

Figure 4.

The classification accuracies of OA and OB subtypes by using gene expression data (circle) and the combined data of the two (square), corresponding to different numbers of IVs from 5 to 200. For the CNVs data (star), the maximum number of IVs that can be reached is 60 due to the limitation of the sample size. Note that the combined analysis can reach higher accuracies (the highest 100%) than either individual analysis.

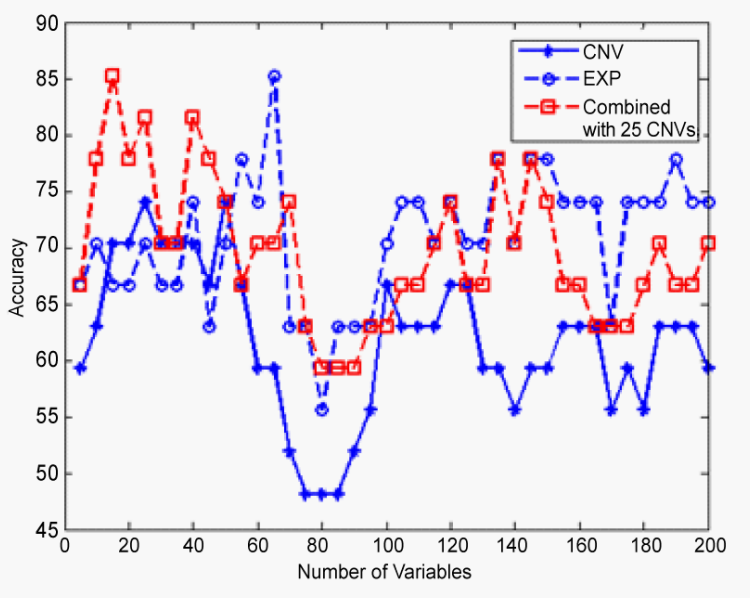

Figure 5.

The classification accuracies of GA1 and GA2 subtypes by using CNVs data (star), gene expression data (circle) and the combined data of the two (square), corresponding to different numbers of IVs from 5 to 200. Note that the combined analysis can have the same highest accuracy as individual gene expression analysis but with less IVs.

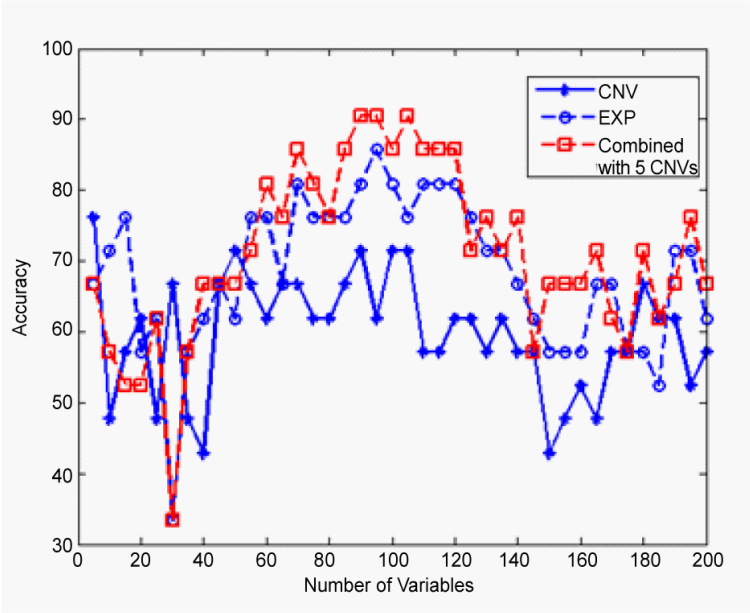

Figure 6.

The classification accuracies of GB1 and GB2 subtypes by using CNVs data (star), gene expression data (circle) and the combined data of the two (square), corresponding to different numbers of IVs from 5 to 200. Note that the combined analysis can reach higher accuracies than either individual analysis.

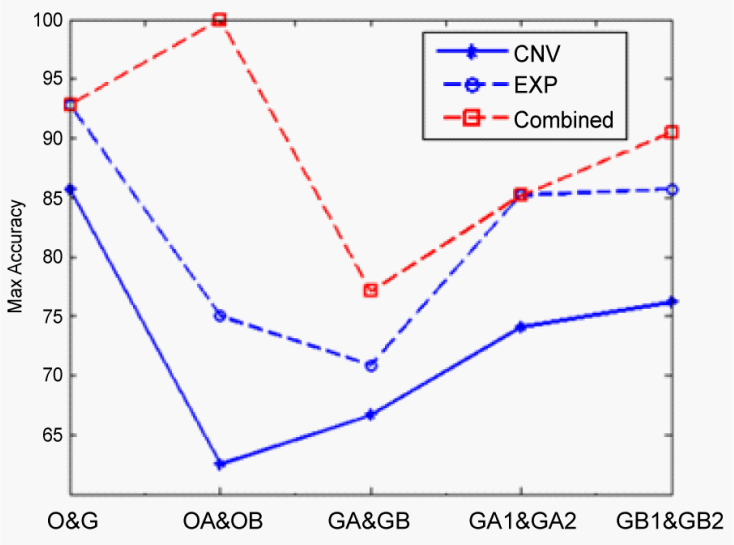

Figure 7.

The maximum classification accuracies for the five binary classifications. Note that the classification accuracy of using the CNVs data individual analysis is the worst; the accuracy using gene expression data individual analysis is better and the accuracy of using combined analysis is the best among the three cases.