Heterogeneous data fusion for brain tumor classification

VANGELIS METSIS[1], HENG HUANG[1], OVIDIU C. ANDRONESI[2,3], FILLIA MAKEDON[1] and ARIA TZIKA[2,3]

[1]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019;
[2]NMR Surgical Laboratory, Department of Surgery, Massachusetts General Hospital and Shriners
Burn Institute, Harvard Medical School; [3]Athinoula A. Martinos Center of Biomedical Imaging,
Department of Radiology, Massachusetts General Hospital, Boston, MA 02114, USA

Received December 29, 2011; Accepted February 10, 2012

DOI: 10.3892/or.2012.1931

Abstract. Current research in biomedical informatics involves analysis of multiple heterogeneous data sets. This includes patient demographics, clinical and pathology data, treatment history, patient outcomes as well as gene expression, DNA sequences and other information sources such as gene ontology. Analysis of these data sets could lead to better disease diagnosis, prognosis, treatment and drug discovery. In this report, we present a novel machine learning framework for brain tumor classification based on heterogeneous data fusion of metabolic and molecular datasets, including state-of-the-art high-resolution magic angle spinning (HRMAS) proton (^{1}H) magnetic resonance spectroscopy and gene transcriptome profiling, obtained from intact brain tumor biopsies. Our experimental results show that our novel framework outperforms any analysis using individual dataset.

Introduction

Brain tumors are the second most common cancer of childhood, and comprise approximately 25% of all pediatric cancers. Over 3,400 children are diagnosed in the US each year; of these, approximately 2,600 will be under the age of 15. Brain tumors, although more treatable in younger ages, are the leading cause of solid tumor cancer death in children; they are the third leading cause of cancer death in young adults ages 20-39. Many researchers are looking for efficient and reliable ways to early diagnose brain tumor types and detect related biomarkers through different biomedical images or biological data. The machine learning algorithms have been playing the most important role during those heterogeneous biomedical/biological datasets analysis to classify different brain tumor types and detect biomarkers. In this study we examine brain tumor samples coming from both children and adults.

Magnetic resonance spectroscopic (MRS) studies of brain biomarkers can provide statistically significant biomarkers for tumor grade differentiation and improved predictors of cancer patient survival (1). Instead of selecting biomarkers based on microscopic histology and tumor morphology, the introduction of microarray technology improves the discovery rates of different types of cancers through monitoring thousands of gene expressions in a parallel, in a rapid and efficient manner (2,3). Because the genes are aberrantly expressed in tumor cells, researchers can use their aberrant expression as biomarkers that correspond to and facilitate precise diagnoses and/or therapy outcomes of malignant transformation.

Different data sources are likely to contain different and partly independent information about the brain tumor. Combining those complementary pieces of information can be expected to enhance the brain tumor diagnosis and biomarkers detection. Recently, several studies have attempted to correlate imaging findings with molecular markers, but no consistent associations have emerged and many of the imaging features that characterize tumors currently lack biological or molecular correlates (4,5). Much of the information encoded within neuroimaging studies therefore remains unaccounted for and incompletely characterized at the molecular level (6). This report presents a computational and machine learning based framework for integrating heterogeneous genome-scale gene expression and MRS data to classify the different brain tumor types and detect biomarkers. We employ wrapper method to integrate the feature selection process of both gene expression and MRS. Three popular feature selection methods, Relief-F (RF), information gain (IG) and χ^2-statistic (χ^2), are performed to filter out the redundant features in both datasets. The experimental results show our framework using the combination of two datasets outperforms any individual dataset on sample classification accuracy that is the standard validation criterion in cancer classification and biomarker detection. Our data fusion framework exhibits great potential on heterogeneous data fusion between biomedical image and biological datasets and it could be extended to another cancer diseases study.

Materials and methods

Advancements in the diagnosis and prognosis of brain tumor patients, and thus in their survival and quality of life, can be achieved using biomarkers that facilitate improved tumor typing. In our research, we apply state-of-the-art, high-resolution magic

Correspondence to: Dr Vangelis Metsis, Heracleia Human-Centered Computing Laboratory, Department of Computer Science and Engineering, University of Texas at Arlington, 500 UTA Blvd, Arlington, TX 76019, USA
E-mail: vmetsis@uta.edu

Key words: data fusion, gene selection, bioinformatics, magnetic resonance spectroscopy

angle spinning (HRMAS) proton (^{1}H) MRS and gene transcriptome profiling to intact brain tumor biopsies, to evaluate the discrimination accuracy for tumor typing of each of the above methods separately and in combination. We used 46 samples of normal (control) and brain tumor biopsies from which we obtained *ex vivo* HRMAS ^{1}H MRS and gene expression data respectively. The samples came from tissue biopsies taken from 16 different people. Out of the 46 biopsies that were analyzed, 9 of them were control biopsies from epileptic surgeries and the rest 37 were brain tumor biopsies. The tumor biopsies belonged to 5 different categories: 11 glioblastoma multiforme (GBM); 8 anaplastic astrocytoma (AA); 7 meningioma; 7 schwanoma; and 5 from adenocarcinoma.

HRMAS ^{1}H MRS. Magnetic resonance spectroscopic (MRS) studies of brain biomarkers can provide statistically significant biomarkers for tumor grade differentiation and improved predictors of cancer patient survival (1). *Ex vivo* high-resolution magic angle spinning (HRMAS) proton (^{1}H) MRS of unprocessed tissue samples can help interpret *in vivo* ^{1}H MRS results, to improve the analysis of micro-heterogeneity in high-grade tumors (7). Furthermore, two-dimensional HRMAS ^{1}H MRS enables more detailed and unequivocal assignments of biologically important metabolites in intact tissue samples (8). In Fig. 1, an *ex vivo* HRMAS ^{1}H MR spectrum of a 1.9 mg anaplastic ganglioglioma tissue biopsy is shown together with metabolites values that correspond to each frequency of the spectrum. Please see more detailed information in ref. 9.

Microscale genomics. A major focus in cancer research is to identify genes, using DNA-microarrays, that are aberrantly expressed in tumor cells, and to use their aberrant expression as biomarkers that correspond to and facilitate precise diagnoses and/or therapy outcomes of malignant transformation (10). In our study, the Affymetrix GeneChip U133Plus® DNA microarray of the complete human genome was used to perform transcriptome profiling on each specimen for two different experimental conditions, minus or plus previous HRMAS NMR analysis. The raw expression data were analyzed for probe intensities using the Affymetrix GeneChip expression analysis manual procedures; and the data were normalized using current R implementations of RMA algorithms (11).

Combining MRS and genomic data. While several studies have utilized MRS data or genomic data to promote cancer classification, to date these two methods have not been combined and cross-validated to analyze the same cancer samples. Herein, we implement a combined quantitative biochemical and molecular approach to identify diagnostic biomarker profiles for tumor fingerprinting that can facilitate the efficient monitoring of anticancer therapies and improve the survival and quality of life of cancer patients. The MRS and genomic data strongly correlate, to further demonstrate the biological relevance of MRS for tumor typing (9). Also, the levels of specific metabolites, such as choline containing metabolites, are altered in tumor tissue, and these changes correspond to the differential expression of Kennedy cycle genes responsible for the biosynthesis of choline phospholipids (such as phosphatidylcholine) and suggested to be altered with malignant transformation (12). These data demonstrate the validity of our combined approach to produce and utilize MRS/genomic biomarker profiles to type brain tumor tissue.

Classification and feature selection methods. Classification aims to build an efficient and effective model for predicting class labels of unknown data. In our case the aim is to build a model that will be able to discriminate between different tumor types given a set of gene expression values or MRS metabolite values or a combination of them. Classification techniques have been widely used in microarray analysis to predict sample phenotypes based on gene expression patterns. Li *et al* have performed a comparative study of multiclass classification methods for tissue classification based on gene expression (13). They have conducted comprehensive experiments using various classification methods including SVM (14) with different multiclass decomposition techniques, Naive Bayes (15), K-nearest neighbor and decision trees (16).

Since the main purpose of this study is not to assess the classification performance of different classification algorithms but to evaluate the potential gain of combining more than one type of data for tumor typing, we only experimented with Naïve Bayes (NB) and support vector machines (SVM) with RBF kernel.

Another related task is *feature selection* that selects a small subset of discriminative features. Feature selection has several advantages, especially for the gene expression data. First, it reduces the risk of over fitting by removing noisy features thereby improving the predictive accuracy. Second, the important features found can potentially reveal that specific chromosomal regions are consistently aberrant for particular cancers. There is biological support that a few key genetic alterations correspond to the malignant transformation of a cell (17). Determination of these regions from gene expression datasets can allow for high-resolution global gene expression analysis to genes in these regions and thereby can help in focusing investigative efforts for understanding cancer on them.

Existing feature selection methods broadly fall into two categories, wrapper and filter methods. Wrapper methods use the predictive accuracy of predetermined classification algorithms, such as SVM, as the criteria to determine the goodness of a subset of features (18). Filter methods select features based on discriminant criteria that rely on the characteristics of data, independent of any classification algorithm (19). Filter methods are limited in scoring the predictive power of combined features, and thus have shown to be less powerful in predictive accuracy as compared to wrapper methods (20). In our experiments we used feature selection method from both major categories. We experimented with Relief-F (RF), information gain (IG), and χ^2-statistic (χ^2), filter methods and we also used wrapper feature selection for each of the two types of classification algorithms.

The basic idea of *Relief-F* (21) is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. Specifically, it tries to find a good estimate of the following probability to assign as the weight for each feature f.

$$w_f = P(\text{different value of } f \mid \text{different class}) - P(\text{different value of } f \mid \text{same class})$$

Information gain (IG) (22) measures the number of bits of information obtained for class prediction by knowing the value of a

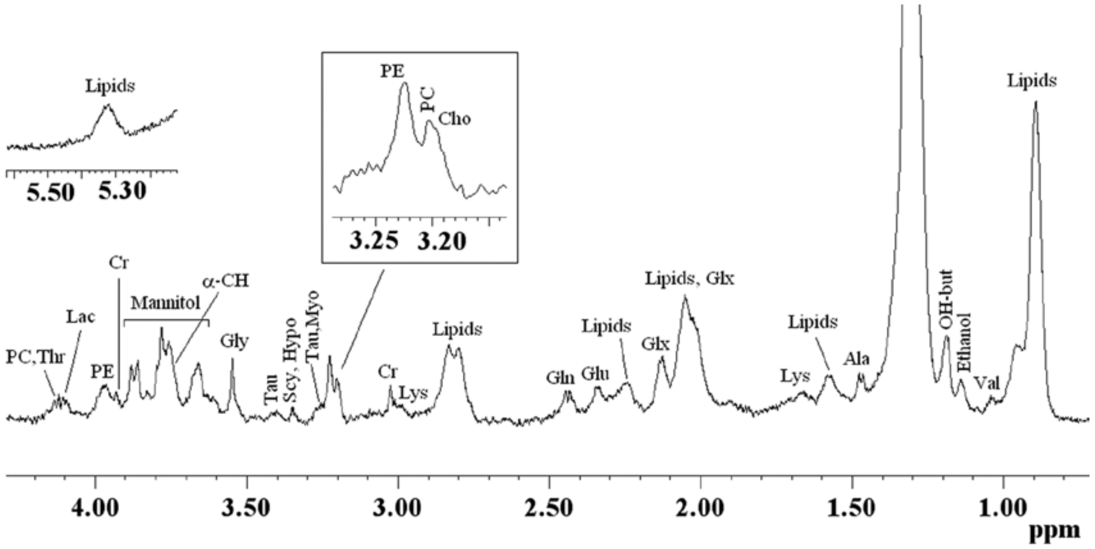

Figure 1. *Ex vivo* HRMAS ^{1}H MR spectrum of a 5.8-mg glioblastoma multiforme (GBM) tissue biopsy. Val, valine; OH-but, OH-butyrate; Lac, lactate; Ala, alanine; Lys, lysine; Glx, β-CH$_2$ of glutamine and glutamate; Glu, glutamate; Gln, glutamine; Cr, creatine; Tau, taurine; Myo, myo-inositol; Hypo, hypotaurine; Scy, scyllo-inositol; Gly, glycine; α-CH of aliphatic amino acids; PE, phosphoetanolamine; Thr, threonine; PC, phoshocholine; Cho, choline. The insert shows the choline containing compounds region.

feature. Let $\{c_i\}_{i=1}^m$ denote the set of classes. Let *V* be the set of possible values for feature *f*. The information gain of a feature f is defined to be:

$$G(f) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + \sum_{\upsilon\in V}\sum_{i=1}^{m} P(f=\upsilon)P(c_i \mid f=\upsilon)\log P(c_i \mid f=\upsilon)$$

The *χ^2-statistic (χ^2)* (23) measures the lack of independence between f and c. It is defined as follows:

$$\chi^2(f) = \sum_{\upsilon\in V}\sum_{i=1}^{m} \frac{(A_i(f=\upsilon) - \mathrm{E}_i(f=\upsilon))^2}{\mathrm{E}_i(f=\upsilon)}$$

where *V* is the set of possible values for feature *f*, $A_i(f = \upsilon)$ is the number of instances in class c_i with$f = \upsilon$ $E_i(f = \upsilon)$ is the expected value of $A_i(f = \upsilon)$. $E_i(f = \upsilon)$ is computed with $E_i(f = \upsilon) = P(f = \upsilon) P(c_i)N$, where *N* is the total number of instances.

Results

Initially we aimed at evaluating how well the classifiers would perform when applying them to each of our datasets separately. For that purpose we performed 10-fold cross validation over our 46 samples by using a combination of feature selection and classification methods.

Table I shows the classification accuracy of Naïve Bayes (NB) and SVM classifiers when using all 16 metabolites and when using a feature selection method. Clearly the wrapper feature selection method gives the better accuracy across all classifiers, followed by the case where we use all metabolites for classification. The SVM classifier using RBF kernel consistently shows the best performance in this type of data. The decision of keeping the top 6 metabolites when using the filter feature selection methods was based on the fact that that was the best number of features that were selected by using the wrapper feature selection method for each classification algorithm.

For the problem of the multiclass classification using gene expression data only, we followed a hybrid feature selection method combining filter and wrapper approaches. Using wrapper approach to select a few top genes starting from an initial number of thousands of genes is computationally prohibiting, and using filter approach to select less than 100 genes does not give good classification accuracy because the final set of selected genes contains genes that are highly correlated to each other, thus giving a redundant set of genes. In our approach, first we selected the top 100 genes using filter feature selection and then we used wrapper feature selection to further reduce the number of genes to be used resulting usually in a number between 5 and 15 genes.

The experimental results (Table II) show that in this type of data the Naïve Bayes was by far the best classification

Table I. Classification accuracy for the 6-class problem using MRS data only.

	NB (%)	SVM (%)
All metabolites	70.21	72.34
χ^2 (top 6)	46.81	51.06
IG (top 6)	46.81	51.06
RF (top 6)	63.83	68.09
Wrapper	72.34	**78.72**

Table II. Classification accuracy using gene expression data only.

	NB (%)	SVM (%)
χ^2 + wrapper	**82.98**	46.81
IG + wrapper	80.85	61.70
RF + wrapper	61.70	57.44

Table III. Classification accuracy using a combination of features from gene expression and MRS datasets.

	NB (%)	SVM (%)
Metabolite selection: wrapper Gene selection: χ^2 + wrapper	85.11	74.46
Metabolite selection: wrapper Gene selection: IG + wrapper	**87.23**	80.85
Metabolite selection: wrapper Gene selection: RF + wrapper	78.72	76.59

algorithm obtaining a maximum accuracy of 82.98% accuracy when combined with χ^2 and wrapper feature selection.

Finally, we tested the classification accuracy of our methods by using a combination of features from both gene expression and MRS data. For the MRS data we tested the wrapper feature selection method which performed best in our previous experiments. For the gene expression data we used the feature selection method that we described above, i.e., combination of filter and wrapper feature selection. After completing the feature selection stage separately for each of the datasets we combined the selected features by putting them in the same feature vector space and using that space for classification. Table III shows the classification accuracy results of our experiments. In most cases, the combination of features sets from the two datasets yield significantly better accuracy than each of them separately. In general Naïve Bayes gives the best performance with a maximum accuracy of 87.23% when using wrapper feature selection for metabolites and a combination of information gain and wrapper feature selection for genes.

Discussion

In this report, we propose a machine learning based data fusion framework which integrates heterogeneous data sources to type different brain tumors. Our method employs real biomedical/biological MRS and genomic data and applies a combination of popular feature selection and classification methods to evaluate the tumor type discrimination capabilities of the two datasets separately and together. The feature selection process identifies a number of biomarkers from each dataset which are subsequently used as features for the classification process. The experimental results show that our data fusion framework outperforms each individual dataset in the brain tumor multi-class classification problem. Since our framework is a general method, it can also be applied to any other biomedical and biological data fusion for sample classification and biomarker detection.

Acknowledgements

This study was supported in part by discretional funds to A. Aria Tzika and Filia Makedon. The authors would like to thank Dr Peter Black for providing us with the biopsies and Dr Dionyssios Mintzopoulos for organizing the MRS data.

References

1. Astrakas LG, Zurakowski D and Tzika AA: Noninvasive magnetic resonance spectroscopic imaging biomarkers to predict the clinical grade of pediatric brain tumors. Clin Cancer Res 10: 8220-8228, 2004.
2. Rich JN, Hans C, Jones B, Iversen ES, McLendon RE, Rasheed BK, Dobra A, Dressman HK, Bigner DD and Nevins JR: Gene expression profiling and genetic markers in glioblastoma survival. Cancer Res 65: 4051-4058, 2005.
3. Watson S, Meng F, Thompson R and Akil H: The Chip as a specific genetic tool. Biol Psychiatry 48: 1147-1156, 2000.
4. Carlson MRJ, Pope WB, Horvath S, Braunstein JG, Nghiemphu P, Tso CL, Mellinghoff I, Lai A, Liau LM and Mischel PS: Relationship between survival and edema in malignant gliomas: role of vascular endothelial growth factor and neuronal pentraxin 2. Clin Cancer Res 13: 2592-2598, 2007.
5. Hobbs SK, Shi G, Homer R, Harsh G, Atlas SW and Bednarski MD: Magnetic resonance image-guided proteomics of human glioblastoma multiforme. J Magn Reson Imaging 18: 530-536, 2003.
6. Diehn M, Nardini C, Wang DS, *et al*: Identification of noninvasive imaging surrogates for brain tumor gene-expression modules. Proc Natl Acad Sci USA 105: 5213-5218, 2008.
7. Cheng LL, Anthony DC, Comite AR, Black PM, Tzika AA and Gonzalez RG: Quantification of microheterogeneity in glioblastoma multiforme with ex vivo high-resolution magic-angle spinning (HRMAS) proton magnetic resonance spectroscopy. Neurooncology 2: 87-95, 2000.
8. Morvan D, Demidem A, Papon J, De Latour M and Madelmont JC: Melanoma tumors acquire a new phospholipid metabolism phenotype under cystemustine as revealed by high-resolution magic angle spinning proton nuclear magnetic resonance spectroscopy of intact tumor samples. AACR 62: 1890-1897, 2002.
9. Tzika AA, Astrakas L, Cao H, *et al*: Combination of high-resolution magic angle spinning proton magnetic resonance spectroscopy and microscale genomics to type brain tumor biopsies. Int J Mol Med 20: 199-208, 2007.
10. Nutt CL, Mani DR, Betensky RA, *et al*: Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. AACR 63: 1602-1607, 2003.
11. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264, 2003.
12. Podo F: Tumour phospholipid metabolism. NMR Biomed 12: 413-439, 1999.
13. Li T, Zhang C and Ogihara M: A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics 20: 2429-2437, 2004.
14. Vapnik V: Statistical Learning Theory. Wiley, New York, NY, 1998.
15. Metsis V, Androutsopoulos I and Paliouras G: Spam filtering with naive bayes-which naive bayes. Proceedings of the Third Conference on Email and Anti-Spam (CEAS), Mountain View, CA, 17: 28-36, 2006.
16. Tan PN, Steinbach M and Kumar VL: Introduction to Data Mining. Addison Wesley, Boston, MA, 2006.
17. Renan MJ: How many mutations are required for tumorigenesis? Implications from human cancer data. Mol Carcinog 7: 139-146. 1993.
18. Guyon I and Elisseeff A: An introduction to variable and feature selection. J Machine Learning Res 3: 1157-1182, 2003.
19. Ding C and Peng H: Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol 3: 185-205, 2005.
20. Chai H and Domeniconi C: An evaluation of gene selection methods for multi-class microarray data classification. ECML/PKDD-04: 3-10, 2004.
21. Kononenko I: Estimating Attributes: Analysis and Extensions of Relief. ECML-94: 171-182, 1994.
22. Mitchell TM: Machine Learning. McGraw Hill, New York, NY, 1997.
23. Liu H and Setiono R: Chi2: feature selection and discretization of numeric attributes. In: Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, Herndon, VA, USA: 388-391, 1995. http://dx.doi.org/10.1109/TAI.1995.479783