CNS Tumor Prediction Using Gene Expression Data Part I

Atiq Islam
University of Memphis, USA

Khan M. Iftekharuddin
University of Memphis, USA

E. Olusegun George
University of Memphis, USA

David J. Russomanno
University of Memphis, USA

INTRODUCTION

Automated diagnosis and prognosis of tumors of the central nervous system (CNS) offer overwhelming challenges because of heterogeneous phenotype and genotype behavior of tumor cells (Yang et al. 2003, Pomeroy et al. 2002). Unambiguous characterization of these tumors is essential for accurate prognosis and therapy. Although the present imaging techniques help to explore the anatomical features of brain tumors, they do not provide an effective means of early detection. Currently, the histological examination of brain tumors is widely used for an accurate diagnosis; however, the tumor classification and grading based on histological appearance does not always guarantee absolute accuracy (Yang et al., 2003, Pomeroy et al., 2002). In many cases, it may not be sufficient to detect the detailed changes in the molecular level using a histological examination (Yang et al. 2003) since such examination may not allow accurate prediction of therapeutic responses or prognosis. If the biopsy sample is too small, the problems are aggravated further.

Toward achieving a more reliable diagnosis and prognosis of brain tumors, gene expression measures from microarrays are the center of attention to many researchers who are working on tumor prediction schemes. Our proposed tumor prediction scheme is discussed in two chapters in this volume. In part I (this chapter), we use an analysis of variance (ANOVA) model for characterizing the Affymetrix gene expression data from CNS tumor samples (Pomeroy et al. 2002) while in part II we discuss the prediction of tumor classes based on marker genes selected using the techniques developed in this chapter. In this chapter, we estimate the tumor-specific gene expression measures based on the ANOVA model and exploit them to locate the significantly differentially expressed marker genes among different types of tumor samples. We also provide a novel visualization method to validate the marker gene selection process.

BACKGROUND

Numerous statistical methods have evolved that are focused on the problem of finding the marker genes that are differentially expressed among tumor samples (Pomeroy et al., 2002, Islam et al., 2005, Dettling et al., 2002, Boom et al., 2003, Park et al., 2001). For example, Pomeroy et al. (2002) uses student t-test to identify such genes in embryonal CNS tumor samples. Because of the non-normality of gene expression measurements, several investigators have adopted the use of nonparametric methods, such as the Wilcoxon Sum Rank Test (Wilcoxon, 1945) as a robust alternative to the parametric procedures. In this chapter, we investigate a Wilcoxon-type approach and adapt the resulting procedures for locating marker genes.

Typically, statistical procedures for microarray data analysis involve performing gene specific tests. Since the number of genes under consideration is usually large, it is common practice to control the potentially large number of false-positive conclusions and family-wise error rates (the probability of at least one false positive statement) through the use of P-value adjustments. Pollard et al. (2003) and Van der Laan et al. (2004a, 2004b, 2005c) proposed methods to control family-wise error rates based on the bootstrap resampling technique of Westfall & Young (1993). Benjamini & Hochberg (1995), Efron et al. (2001) and Storey et al. (2002, 2003a, 2003b, 2004) introduced various techniques for controlling the false discovery rate (FDR), which is defined as the expected rate of falsely rejecting the null hypotheses of no differential gene expression. These adjustment techniques have gained prominence in statistical research relating to microarray data analysis. Here, we use FDR control because it is less conservative than family-wise error rates for adjusting the observed P-values for false discovery. In addition, we propose a novel marker gene visualization technique to explore appropriate cutoff selection in the marker gene selection process.

Before performing formal analysis, one should identify the actual gene expression levels associated with different tissue groups and discard or minimize other sources of variations. Such an approach has been proposed by Townsend & Hartl (2002) who use a Bayesian model with both multiplicative and additive small error terms to detect small, but significant differences in gene expressions. As an alternative, an ANOVA model appears to be a natural choice for estimating true gene expression (Kerr et al., 2000, Pavlidis et al., 2001, Wolfinger et al. 2001). In the context of cDNA microarray data, the ANOVA model was first proposed by Kerr et al. (2000).

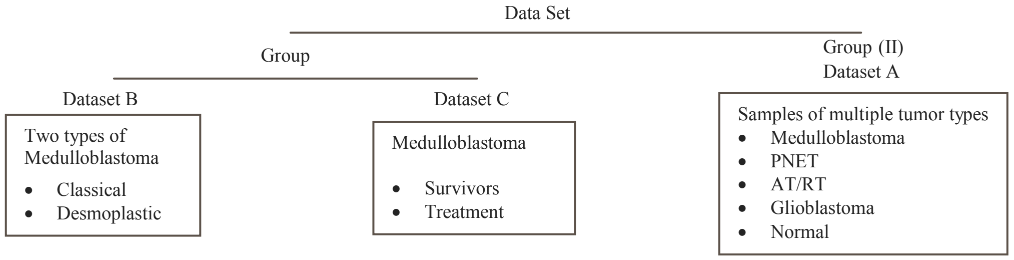
TUMOR-SPECIFIC GENE EXPRESSION ESTIMATION AND VISUALIZATION TECHNIQUES

To illustrate our procedure, we use the a microarray data set by Pomeroy et al. (2002) of patients with different types of embryonal tumors. The patients include 60 children with medulloblastomas, 10 young adults with malignant gliomas, 5 children with AT/RTs, 5 with renal/extra-renal rhabdoid tumors, and 8 children with supratentorial PNETs. First, we preprocess the data to remove extraneous background noise and array effects. To facilitate our analysis, we divide the dataset into groups as shown in Fig. 1. We rescale the raw expression data obtained from Affymetrix's GeneChip to account for different chip intensities.

Microarray data typically suffer from unwanted sources of variation, such as large-and-small-scale intensity fluctuations within spots, non-additive background, fabrication artifacts, probe coupling and processing procedures, target and array preparation in the hybridization process, background and over-shining effects, and scanner settings (McLachlan & Ambroise, 2005). To model these variations, a number of methods have been reported in the literature (Kerr et al., 2000, Lee et al., 2000, Pavlidis, 2001, Wolfinger et al., 2001, Ranz et al., 2003, Townsend, 2004, Tadesse et al., 2005). An ANOVA model similar to the one used by Kerr el al. (2000) is adopted in our current work and facilitates obtaining the tumor-specific gene expression measures from the preprocessed microarray data. Our two-way ANOVA model is given as:

$$y_{jgk} = \mu + \alpha_g + \beta_j + \gamma_{jg} + \varepsilon_{jgk} \qquad (1)$$

Figure 1. Dataset grouping


Data Set
Group
Group (II)
Dataset A
Dataset B
Dataset C
Two types of Medulloblastoma
• Classical
• Desmoplastic
Medulloblastoma
• Survivors
• Treatment
Samples of multiple tumor types
• Medulloblastoma
• PNET
• AT/RT
• Glioblastoma
• Normal