CNS Tumor Prediction Using Gene Expression Data Part II

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

In this chapter, we propose a novel algorithm for characterizing a variety of CNS tumors. The proposed algorithm is illustrated with an analysis of an Affymetrix gene expression data from CNS tumor samples (Pomeroy et al., 2002). As discussed in the previous chapter entitled: CNS Tumor Prediction Using Gene Expression Data Part I, we used an ANOVA model to normalize the microarray gene expression measurements. In this chapter, we introduce a systemic way of building tumor prototypes to facilitate automatic prediction of CNS tumors.

BACKGROUND

DNA microarrays, also known as genome or DNA chips, have become an important tool for predicting CNS tumor types (Pomeroy et al., 2002, Islam et al., 2005, Dettling et al., 2002). Several researchers have shown that cluster analysis of DNA microarray gene expression data is helpful in finding the functionally similar genes and also to predict different cancer types. Eisen et al. (1998) used average linkage hierarchical clustering with correlation coefficient as the similarity measure in organizing gene expression values from microarray data. They showed that functionally similar genes group into the same cluster. Herwig et al. (1999) proposed a variant of the K-means algorithm to cluster genes of cDNA clones. Tomayo et al. (1999) used self-organized feature maps (SOFMs) to organize genes into biologically relevant groups. They found that SOFMs reveal true cluster structure compared to the rigid structure of hierarchical clustering and the structureless K-means approach. Considering the many-to-many relationships between genes and their functions, Dembele et al. (2003) proposed a fuzzy C-means clustering technique. The central goal of these clustering procedures (Eisen et al., 1998, Herwig et al., 1999, Tomayo et al., 1999, Dembele et al., 2003) was to group genes based on their functionality. However, none of these works provide any systematic way of discovering or predicting tissue sample groups as we propose in our current work.

To identify tissue sample groups, Alon et al. (1999) proposed a clustering algorithm that uses a deterministic-annealing algorithm to organize the data in a binary tree. Alizadeh et al. (2000) demonstrated a successful molecular classification scheme for cancers from gene expression patterns by using an average linkage hierarchical clustering algorithm with Pearson's correlation as the similarity measure. However, no formal way of predicting the category of a new tissue sample is reported in (Alon et al., 1999, Alizadeh et al., 2000). Such class prediction problems were addressed by Golub et al. (1999) who used SOFMs to successfully discriminate between two types of human acute leukemia. Dettling et al. (2002) incorporated the response variables into gene clustering and located differentially expressed groups of genes from the clustering result. These gene groups were then used to predict the categories of new samples. However, none of the above-mentioned works (Dettling et al., 2002, Golub et al., 1999, Alon et al.,

1999, Alizadeh et al., 2000) considered the correlation among the genes in classifying and/or predicting tissue samples. Moreover, none of these provided any systematic way of handling the probable subgroups within the known groups. In this chapter, we consider both correlations among the genes and probable subgroups within the known groups by forming appropriate tumor prototypes. Further, a major drawback of these analyses (Dettling et al., 2002, Eisen et al., 1998, Herwig et al., 1999, Tomayo et al., 1999, Dembele et al., 2003, Golub et al., 1999, Alon et al., 1999, Alizadeh et al., 2000) is insufficient normalization. Although, most of these methods normalize the dataset to remove the array effects; they do not concentrate on removing other sources of variations present in the microarray data.

Our primary objective in this chapter is to develop an automated prediction scheme for CNS tumors, based on DNA microarray gene expressions of tissue samples. We propose a novel algorithm for deriving prototypes for different CNS tumor types, based on Affymetrix HuGeneFL microarray gene expression data from Pomeroy et al. (2002). In classifying the CNS tumor samples based on gene expression, we consider molecular information, such as the correlations among gene expressions and probable subgroupings within the known histological tumor types. We demonstrate how the model can be utilized in CNS tumor prediction.

CNS TUMOR PROTOTYPE FOR AUTOMATIC TUMOR DETECTION

The workflow to build the tumor prototypes is shown in Fig. 1. In the first step, we obtain the tumor-type-specific gene expression measures. Then, we identify the marker genes that are significantly differentially expressed among tissue types. Next, a visualization technique is used to analyze the appropriateness of the marker gene selection process. We organize the marker genes in groups so that highly correlated genes are grouped together. In this clustering process, genes are grouped based on their tumor-type-specific gene expression measures. Then, we obtain eigengene expressions measures from each individual gene group by projection of gene expressions into the first few principal components. At the end of this step, we replace the gene expression measurements with eigengene expression values that conserve correlations between strongly correlated genes. We then divide the tissue samples of known tumor types into subgroups. The centroids of these subgroups of tissue samples with eigengene expressions represent the prototype of the corresponding tumor type. Finally, any new tissue sample is predicted as the tumor type of the closest centroid. This proposed novel prediction scheme considers both the correlation among the highly correlated genes and the probable phenotypic subgrouping within the known tumor types. These issues are often ignored in the literature for predicting tumor categories. The detail of the steps up to the identification of marker genes are provided in the previous chapter entitled: CNS Tumor Prediction Using Gene Expression Data Part I. In this section, we provide the details of the subsequent steps.

Now, we discuss the creation of the tumor prototypes using the tumor-specific expression values of our significantly differentially expressed marker genes identified in the previous step. Many of the marker genes are likely to be highly correlated. Such correlations of the genes affect successful tumor classification. However, this gene-to-gene correlation may provide important biological information. Hence, the inclusion of the appropriate gene-to-gene correlations in the tumor model may help to obtain a more biologically meaningful tumor prediction. To address this non-trivial need, we first group the highly correlated genes using the complete linkage hierarchical approach wherein correlation coefficient is considered as the pair-wise similarity measure of the genes. Next, for each of the clusters, we compute the principal components (PCs) and project the genes of the corresponding cluster onto the first 3 PCs to obtain eigengene expressions (Speed, 2003). Note that the PCs and the eigengene expressions are computed separately for each cluster. Such eigengenes encode the correlation information among

Figure 1. Simplified workflow to build the tumor prototypes