Computers in Biology and Medicine 43 (2013) 1863–1869

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Sparse Manifold Clustering and Embedding to discriminate gene expression profiles of glioblastoma and meningioma tumors ☆

Juan M. García-Gómez [a,c,*], Juan Gómez-Sanchis [b], Pablo Escandell-Montero [b], Elies Fuster-Garcia [a], Emilio Soria-Olivas [b]

[a] Biomedical Informatics Group (IBIME-ITACA), Universitat Politècnica de València. Camí de Vera s/n, Building 8G, 1st Floor. 46022, València, Spain
[b] IDAL, Intelligent Data Analysis Laboratory, Electronic Engineering Department, University of Valencia, Avd Universitat s/n, 46100, Burjassot, Valencia, Spain
[c] GIBI230(Grupo de Investigación Biomédica en Imagen), Instituto de Investigación Sanitaria (IIS) hospital la Fe, Spain.

ARTICLE INFO

Article history:
Received 1 June 2013
Accepted 31 August 2013

Keywords:
Manifolds
Automatic classification
Microarray data analysis
Medical applications
Bioinformatics

ABSTRACT

Sparse Manifold Clustering and Embedding (SMCE) algorithm has been recently proposed for simultaneous clustering and dimensionality reduction of data on nonlinear manifolds using sparse representation techniques. In this work, SMCE algorithm is applied to the differential discrimination of Glioblastoma and Meningioma Tumors by means of their Gene Expression Profiles. Our purpose was to evaluate the robustness of this nonlinear manifold to classify gene expression profiles, characterized by the high-dimensionality of their representations and the low discrimination power of most of the genes. For this objective, we used SMCE to reduce the dimensionality of a preprocessed dataset of 35 single-labeling cDNA microarrays with 11500 original clones. Afterwards, supervised and unsupervised methodologies were applied to obtain the classification model: the former was based on linear discriminant analysis, the later on clustering using the SMCE embedding data. The results obtained using both approaches showed that all (100%) the samples could be correctly classified and the results of all repetitions but one formed a compatible cluster of predictive labels. Finally, the embedding dimensionality of the dataset extracted by SMCE revealed large discrimination margins between both classes.

1. Introduction

Brain tumors are[1] growths of abnormal cells in the tissues of the brain. Brain tumors are the second fastest growing cause of cancer death among people older than 65 years [1], in addition, they are also the second leading cause of cancer death (after leukemia) in children under fifteen years and young adults up to the age of thirty-four.

The brain tumors are classified in grades based on their malignancy characteristics. On one hand, low-grade tumors have low proliferative potential and possibility of cure following surgical resection alone. On the other hand, high grade tumors are generally associated with a rapid *pre-* and *post-operative* evolution of the disease. Specifically, the most frequent primary brain tumor types are of glial origin (40%), 30% are derived from the meninges and 8% are located in cranial and spinal nerves [2]. In adults over 45 years, the most frequent tumors are from the meningioma and glioblastoma types. Meningioma are usually graded I tumors whereas glioblastoma are the most aggressive tumors (grade IV). Glioblastomas arise from glial cells, they are the most infiltrative tumors, and a poor prognosis is associated with them [3]. In contrast, Meningiomas arise from meningothelial cells, they usually show well defined edges and they remain at the benign stage [4].

Biomedical data that come from different biological levels offer great information for the medical decision process. New biomedical technologies go insight the origin and prognosis of the illness moving to an evidence-based medicine paradigm. Despite of the extended use of histopathology as gold standard of Primary Brain Tumours (PBTs), high throughput genome sequences and expression techniques [5] will likely allow to improve the prediction of the clinical course and the response to therapy of patients [6,7]. Microarray-based gene expression profiles simultaneously show messenger RNA expression level of genes monitored under certain condition, such as belonging to a tumor tissue.

Different technologies are available to study gene expression at the transcriptomic level [8,9]. Single-labeling cDNA microarrays are a cheap technology more flexible than any commercial product. This makes them accessible to a wide spectrum of research laboratories of molecular biology. A challenging problem of high-throughput genome techniques is its high-dimensionality, in terms of the number of variables in the profiles [10,11]. For example, the initial dimension of gene-expression profiles studied in this work is 11,500. Moreover, most of those variables have little discrimination power, and hence,

☆This work was supported by the University of Valencia through project UV-INV-AE11-41271.

* Corresponding author at: Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas, Universidad Politécnica de Valencia (UPV), Camino de Vera, s/n, Edificio G-8, Acceso B, 3ª planta, Valencia – 46022, Spain. Tel.: +34 96 387 72 78.
E-mail address: juanmig@ibime.upv.es (J.M. García-Gómez).

[1] Dictionary of Cancer Terms. National Cancer Institute. http://www.cancer.gov/dictionary/ (Online; accessed 10-3-2008).

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.08.025

they do not give relevant information for the design of predictive models for classification. Hence, robust feature reduction methodologies are required to obtain gene-expression signatures to visualize the datasets, study differential gene-expressions, design predictive models and identify new molecular subtypes.

Nonlinear manifold techniques have recently arisen as the generalization of the classical linear multivariate techniques for feature extraction and reduction. Nonlinear manifolds establish a correspondence between a high dimensional space and a lower dimensionality from topological relationships. Sparse Manifold Clustering and Embedding (SMCE) is a new nonlinear manifold algorithm proposed for simultaneous clustering and dimensionality reduction of data on nonlinear manifolds using sparse representation techniques.

Several studies have applied machine learning techniques, including manifold learning techniques, for discriminating gene expression profiles or new next generation sequencing from tumours. Fuller et al. [12] used cDNA array technology to profile with multidimensional scaling (MDS) the gene expression of 30 primary human glioma tissue samples comprising 4 different glioma subtypes: glioblastoma (GM, WHO grade IV), anaplastic astrocytoma (AA, WHO grade III), anaplastic oligodendroglioma (AO, WHO grade III), and oligodendroglioma (OL, WHO grade II). Marko et al. [13] applied different unsupervised method to integrated genomic, transcriptomic, and morphologic data to reveal molecular classification of low-grade gliomas. Zang and Zang [14] tested a supervised orthogonal discriminant projection for tumor classification using gene expression data in five public tumor datasets. Huang and Feng [15] proposed a parameter-free semi-supervised local Fisher discriminant analysis (pSELF) to map the gene expression data into a low-dimensional space for tumor classification. They tested the method in the SRBCT, DLBCL, and Brain Tumor gene expression data sets. Siu and Hing [16] applied the locally linear embedding (LLE) method to project high dimensional genomic data from the 1000 Genomes Project and a PHASE III Mexico dataset of the HapMap into low dimensional in order to identify population substructures by common and rare variants. Li et al. [17] developed and tested in documents, images, and gene expression data sets their relational multimanifold coclustering based on symmetric nonnegative matrix trifactorization.

In this work, the performance of the SMCE algorithm to classify glioblastoma and meningioma tumors by means of their gene expression profiles has been evaluated. Glioblastomas and meningiomas tumor types have been chosen because they are two diagnoses from different types of cells and with an antagonist aggressiveness behavior. This leads us to expect a good outcome of the classification results, so it gives us the opportunity to focus the attention on the capability of the method to reduce the dimensionality of the representation and to evaluate the robustness of the method with respect to changes in the samples.

For this objective, an Affymetrix-based preprocessing to a dataset of 35 single-labeling cDNA microarrays with 11,500 original clones has been performed. Next, SMCE has been applied directly to the matrix of pre-processed gene-expression values. Afterwards, a linear discriminant analysis and a clustering algorithm have been applied alternatively to the SMCE embedded data in order to produce the classification model. The evaluation based on a bootstrap strategy showed the stability of the models. Moreover, the visual inspection of the embedding dimensionality ratified the discrimination capability of the approach.

Next section shows the dataset preparation process. Afterwards, Section 4 briefly describes the SMCE method and Section 4 shows the simulation experiments. Section 5 shows the obtained results and discusses their relevance. Finally, Section 6 shows the conclusions of the paper.

2. Dataset preparation

The dataset used in this study is composed by samples extracted from 35 frozen biopsies carried out at the Hospital Prínceps d'Espanya (Bellvitge, Spain). The preparation of the samples is fully described in [18]. From the 35 samples, 18 samples were histopathologically diagnosed as Meningiomas (MM) and 17 samples as Glioiblastomas (GM). The samples consisted in single-labeling cDNA microarrays based on human CNIO oncochip, which is a 12 K cDNA Clone Set microarray that contains 11,500 cDNA clones representing 9300 loci. The total number of probes (DNA spots that contains specific DNA sequence) by microarray was 27,648.

Gene-expression data needed to be pre-processed to allow the comparison between the microarrays expression values. The pre-processing used in this study is based on an Affymetrix preprocessing pipeline, which consisted in the next steps:

1. *Pre-filtering by scanning flags*. We dropped each probe whose scanner flag was less than 0. This step avoided the use of the genes that did not agree the quality control established during the scanning of the microarray. After the pre-filtering, based on the scanning quality control, 23,652 replicates remained in the expression matrix. The criterion to drop a replicate was to have more than the 20% of control.
2. *Background correction*. We removed the non-specific hybridization contained in the foreground. For this purpose, we applied the background smoothing procedure defined by Edwards [19]. This method applies the simple subtractions when foreground is bigger enough than the background; on the other hand, when the difference is small or negative, the correction is carried out by a smooth monotonic function that is linear with respect the background intensity of the spot.
3. *Normalization*. Normalization methods must be applied before comparing arrays to attenuate the effect of systematic variations that do not correspond to differences in expression. The normalization method applied in this study was the *mean cyclic loess normalization* method [20]. This is an adaptation of the cyclic loess method for multiple one-labeled microarrays, where, the expression value of microarrays is adjusted based on the local regression of the M vs. A function, being M the difference of the log expression value of each probe of the microarray and the equivalent average probe calculated from all the microarrays in the dataset and being A the average of the log expression values instead of the difference. The quantile normalization of the Robust Multiarray Average (RMA) [20] assumes that the intensities of the chips follow a common distribution, hence, the normalization substitutes the value of each probe of each microarray by the mean of the probes that are in the same quantile for all the microarrays.
4. *Post-filtering by CNIO verification*. A post-filtering method was applied to drop the genes that were not verified by the CNIO institution by PCR evidence (single band) and sequence verification. After the background correction and the normalization steps, the post-filtering based on the CNIO verification was applied and 15,584 replicates passed the filter.
5. *Summarization*. The summarization of the replicates of each microarray obtains only one measure of the expression for each gene of each patient. In this study, we considered the simple mean to obtain the final gene-expression value. The summarization of the replicates by gene leads us to a final expression matrix of 7219 genes for the 35 samples.

This pipeline for preprocessing the Gene expression was carried out with the R language and Bioconductor libraries. As a result, each sample of the preprocessed dataset consisted in 7219 variables, which correspond to the gene expression values in arbitrary units.