Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

A GMM-IG framework for selecting genes as expression panel biomarkers

Mingyi Wang [a], Jake Y. Chen [a,b,c,*]

[a] *School of Informatics, Indiana University, 535 W. Michigan Street, Indianapolis, IN 46202, USA*
[b] *Department of Computer and Information Science, Purdue University School of Science, Indianapolis, IN 46202, USA*
[c] *Indiana Center for Systems Biology and Personalized Medicine, 719 N. Indiana Ave, WK Suite #190, Indianapolis, IN 46202, USA*

ARTICLE INFO

Article history:
Received 28 August 2008
Received in revised form 29 June 2009
Accepted 2 July 2009

Keywords:
Gene selection
Data integration
Microarray data
Lung cancer
Gaussian mixture model
Information gain

ABSTRACT

Objective: The limitation of small sample size of functional genomics experiments has made it necessary to integrate DNA microarray experimental data from different sources. However, experimentation noises and biases of different microarray platforms have made integrated data analysis challenging. In this work, we propose an integrative computational framework to identify candidate biomarker genes from publicly available functional genomics studies.

Methods: We developed a new framework, Gaussian Mixture Modeling-Coupled Information Gain (GMM-IG). In this framework, we first apply a two-component Gaussian mixture model (GMM) to estimate the conditional probability distributions of gene expression data between two different types of samples, for example, normal versus cancer. An expectation-maximization algorithm is then used to estimate the maximum likelihood parameters of a mixture of two Gaussian models in the feature space and determine the underlying expression levels of genes. Gene expression results from different studies are discretized, based on GMM estimations and then unified. Significantly differentially-expressed genes are filtered and assessed with information gain (IG) measures.

Results: DNA microarray experimental data for lung cancers from three different prior studies was processed using the new GMM-IG method. Target gene markers from a gene expression panel were selected and compared with several conventional computational biomarker data analysis methods. GMM-IG showed consistently high accuracy for several classification assessments. A high reproducibility of gene selection results was also determined from statistical validations. Our study shows that the GMM-IG framework can overcome poor reliability issues from single-study DNA microarray experiment while maintaining high accuracies by combining true signals from multiple studies.

Conclusions: We present a conceptually simple framework that enables reliable integration of true differential gene expression signals from multiple microarray experiments. This novel computational method has been shown to generate interesting biomarker panels for lung cancer studies. It is promising as a general strategy for future panel biomarker development, especially for applications that requires integrating experimental results generated from different research centers or with different technology platforms.

1. Introduction

In the last decade, gene expression profiles that simultaneously measures the expression level of tens of thousands genes is increasingly being applied to the diagnosis, prognosis, and treatment selection of many diseases (e.g. [1–6]). A common task for these biomedical applications involves the selection of relevant genes for sample classification (e.g., different phenotypes, subclasses of cancers, normal or cancerous). This task, known as feature selection in machine learning, begins with available class-labeled data and aims to determine which feature(s) best discriminate the classes. However, in typical gene expression profiling studies, researchers can only afford to collect dozens or at most hundreds of samples, primarily due to the laborious process and high cost in obtaining biological samples in human. This phenomena, also known as "large *p*, small *n*" (*p* for genes and *n* for array samples), has made it very challenging to obtain statistically significant results due to far more variables than available observations [7]. This often leads to unstable and irreproducible selected gene lists when they are compared with results obtained from other studies [8]. With the increasing availability of DNA

* Corresponding author at: Department of Computer and Information Science, Purdue University School of Science, Indianapolis, IN 46202, USA. Tel.: +1 317 278 7604; fax: +1 317 278 9201.

E-mail addresses: mingy.wang@gmail.com (M. Wang), jakechen@iupui.edu (J.Y. Chen).

0933-3657/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2009.07.006

microarray data accumulated from different labs with similar experimental study goals, it has become essential bioinformatics research topic to integrate heterogeneous study-specific gene expression data to improve accuracy and consistency of single-study microarray gene expression experiments, particularly for emerging biomarker panel applications [9].

Meta-analysis [10,11] and transformation methods [12,13] are two major current computational strategies to combine information from different microarray experiments. The meta-analysis approach treats each gene in each study independently during the combination and different studies are integrated for each gene first into a "meta microarray result" before significant gene lists are selected. The major drawback is relevant genes may be missed from individual studies due to a small sample size. The transformation approach translates gene expression measurements from different studies into a common scale before the unification of these studies. However, the heterogeneities of microarray experiments, for example, different sample processing protocols, different microarray platforms, different absence/presence calling techniques, and the lack of cross-comparison standards for microarray experiments have made scaling of microarray experiments impractical for experiments performed from other research labs, even if the experimental study goals and samples are similar. As of today, there is little consensus or guidelines as to how to perform such a data transformation effectively.

In this paper, we propose a new framework based on Gaussian Mixture Modeling-coupled Information Gains (GMM-IG) to address the challenging integrated microarray data analysis problem, using biomarkers identified from multiple lung cancer data sets as a case study. The framework overcomes the weakness of conventional meta-analysis by assuming inherently different models for different groups of genes expressed under disease versus normal conditions. It also does not assume the availability of global scaling functions that map results from one gene expression experiment to another. Instead, the unification of results is based on unions of genes with discrete class-label values. This new framework enables the expansion of microarray studies by including gene expression results performed under diverse microarray study conditions, therefore relieving concerns from the intractable "large *p*, small *n*" problem. Results obtained from the GMM-IG framework are shown to be both reliable and accurate for the top candidate biomarkers identified, therefore showing encouraging potentials for future biomarker discovery applications.

2. Methods

In this section, we describe the methodology used to select genes after integrating different microarray data sets. To summarize briefly, our GMM-IG framework consists of two phases of data analysis. In the first phase, we use Gaussian Mixture Modeling (GMM) to quantify gene expression measurements of the raw data and provide initial assessment of the features. In the second phase, we use an Information Gain (IG) method to select and rank important genes from integrated data sets.

2.1. Gaussian mixture modeling

Statistical mixture models have been used before, primarily in the clustering of microarray gene expression profiles [14] and identifications of differentially expressed genes [15,16]. In this paper, we apply this computational method to a new area: integrative cross-study microarray data analysis. Here, we make a useful empirical assumption about the activity of genes and hence their varied expression level across a set of microarrays, i.e., that the genes generally assume two distinct biological states each (either "on" or "off"). The combination of such binary pattern from multiple genes determines the sample phenotype, such as, normal or cancerous. Collectively, the measurements for a particular gene from different studies follow a complex distribution, which we could model as the mixture of two simpler Gaussian distributions in normal and cancerous samples respectively, one corresponding to the lack of up-regulation or down-regulation and the other to up-regulation. Then, we estimate the parameters of these two distributions from the original gene expression values before we binarize each gene according to its modeled distributions. We further test if the underlying binary state of a gene varies between the two classes in order to determine the discriminative value of the gene as candidate biomarkers for disease versus normal classifications. If the underlying binary state of a gene does not significantly vary between the two classes, then this gene is not discriminative for the classification and should discard. To do so, we use a heuristic procedure that measures the separability of the mixture components. Although expression values of a gene with high separability may be highly variable among different data sources, the underlying mixture of two distributions should remain the same for this gene, unless most studies used are determined unreliable—extremely rare if one can select only trustworthy study results into the integrative analysis. In fact, preliminary assessment of "trustworthiness" of each study to be integrated can be performed with conventional meta-analysis and data sets that provide minimal overlaps of results with the majority studies may be filtered prior to the GMM-IG analysis. This assumption provides greater ease of conceptualization and is extremely valuable for us to cancel out the disparities caused by different sources.

The following provides additional details to the GMM-IG framework. Computationally, microarray data can be represented as a matrix of expression levels. For N microarray experiments (corresponding to N tissue samples), where we measure the expression levels of P genes in each experiment, the results can be represented by $P \times N$ matrix. Mixture models are a type of density models which comprise a number of component functions, usually Gaussian. The distribution of a gene across N samples, $X \in \mathbf{R}^N$ is a mixture of K Gaussians if its density function is

$$f(X|\theta) = \sum_{j=1}^{K} \alpha_j \Phi(X; \mu_j, \sigma_j^2) \tag{1}$$

and

$$\Phi(X; \mu_j, \sigma_j^2) = \frac{1}{\sqrt{(2\pi)|\Sigma_j|}} \exp\left\{ -\frac{1}{2}(X-\mu_j)^T \sum_{j}^{-1} (X-\mu_j) \right\} \tag{2}$$

where $\mu_j \in \mathbf{R}^N$ and Σ_j are the mean vector and $N \times N$ covariance matrix of Gaussian j, respectively. The parameter set $\theta = \{\alpha_j, \mu_j, \Sigma_j\}$ consists of

$$\alpha_j > 0, \quad \sum_{j=1}^{K} \alpha_j = 1 \tag{3}$$

where, α_j is the prior probability for Gaussian j.

Learning a GMM is essentially an unsupervised clustering task. The expectation-maximization (EM) algorithm [17] is used to determine the maximum likelihood parameters of a mixture of K Gaussians in the feature space. Given a set of feature vectors $x_1, \ldots, x_N$, the maximum likelihood estimation of θ is

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, f(x_1, \ldots, x_N|\theta) \tag{4}$$

The EM algorithm is an iterative method to obtain θ_{ML}. The first step in applying the EM algorithm to the problem at hand is to