The impact of gene expression analysis on the classification and prediction of patients’ medical conditions
Famili, Fazel; Phan, Sieu; Liu, Zijing; Pan, Youlian
The impact of gene expression analysis on the classification and prediction of patients' medical conditions

Fazel Famili, Sieu Phan, Ziying Liu and Youlian Pan

Knowledge Discovery Group, Institute for Information Technology, National Research Council Canada, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada

{fazel.famili, sieu.phan, ziying.liu, youlian.pan}@nrc-cnrc.gc.ca

Abstract. Proper analysis and validation of gene expression data is not a trivial task. Most of the existing approaches identify individual or group of markers/signatures where the validation only goes as far as literature and biological experimental validation. In this paper we discuss a framework that could be considered to develop and validate patterns from gene expression data and work towards future clinical test kits.

Keywords: Gene expression analysis, patient classification, medical condition prediction

1 Introduction

With rapid development of tools and support systems in life sciences, medical diagnosis and prognosis is going through an evolving paradigm. This change of paradigm is so fast that even some of the procedures that were valid as far as five years ago may no longer be fully acceptable or sufficient for today’s modern medicine. Consequently, the ability to identify the key elements related to classification of patients (where healthy and disease or its subtypes matter) has rapidly changed. Two issues related to the newly generated life science data are important to discuss here: (i) data type and its volume, (ii) data complexity and the need for integration.

High throughput data consists of several forms, most of which comes from a common starting point (e.g. samples, tissue, cell lines). Examples are genomics, proteomics, metabolomics, and of course, associated sequence data (DNA or protein sequences). The volume of these data, initially motivated by the completion of human genome, has rapidly increased over the last decade so that it is now being considered as 50-100% increase every year [1, 2, 3]. Other factors have also contributed to this, such as computational powers, data storage and understanding the real value of this data. Although some researchers have indicated that the superabundance of this data and the associated information may have created a scarcity of attention, we can
provide several examples where small or large amount of data has resulted in novel discoveries leading to applications and even tools that would be helpful for improved life sciences. Examples are: (i) biomarkers discovery and disease signatures identifications, (ii) better disease understanding, (iii) improved therapeutic (better personalized medical care) and more efficient drug discovery, and (iv) better patient monitoring and understanding the drug efficacy.

All of these have created an enormous business opportunities and life science industry applications where successful applications have been deployed or even products (such as tool kits) have been introduced. Two examples are:

- Agendia, an innovative cancer diagnostics company based on genomics, uses its knowledge and expertise in the application of clinically useful gene expression profiling to facilitate breast cancer diagnosis, prognosis and treatment. Motivated by the Human Genome Project and a genomics microarray platform, Agendia’s MammaPrint breast cancer test has been designed to determine a patient’s risk of metastases, therefore providing a more personalized approach to breast cancer therapy.

- ArticDX, a Canadian company that is developing two molecular tests, Colo Risk™, for Colorectal Cancer and Macula Risk™ for Age-related Macular Degeneration, developed genetic tests that are based on their own proprietary markers discovered by ARCTIC researchers. They are partnering with other researchers to support the translation of their gene discoveries into applications that can be of benefit to doctors and patients in the healthcare setting.

Our objective in this paper is to briefly highlight the importance and impact of high throughput data in life sciences with a focus on gene expressions data, its volume, crucial role of data analysis and potential industrial applications. We provide a brief literature review in the next section followed by a spectrum of applications. Focusing on a small part of this spectrum, we then introduce our proposed framework, explain its application, and provide some discussions that include our ongoing research in this area.

2 Related work

Gene expression analyses based on the use of microarray data allow a simultaneous investigation of the expression patterns of tens of thousands of genes, which provide a powerful approach for understanding the intricate process of tumor development and progression in a global view. Gene expression patterns are thought to be different for various tissue types, at different stages of development, disease vs. normal tissue, and among disease subtypes, which can be used to improve accuracy of diagnostics, predictive markers of treatment response and management of patients’ treatment plan.

Since the emergence of microarray techniques in 1995[4, 5], an enormous number of gene expression knowledge discovery efforts have attempted to understand the genes involved in malignant development and disease progression, to improve the accuracy of diagnosis and prognosis. Golub’s study [6] is one of the pioneers in using gene expression data to discover more efficient diagnostic approach. In their study, a
set of genetic markers of different types of leukemia were obtained for classification and diagnostics of leukemia, in which the separation of AML (acute myeloid leukemia) and ALL (acute lymphoblastic leukemia) on the basis of gene expression analyses was demonstrated. Methods called “neighborhood analysis” and “weighted vote” were applied to build class predictors. SOM clustering algorithm was applied for class discovery. Recently, Andersson et al [7] conducted gene expression analyses on 121 pediatric acute leukemia patients based on lineage and genetic subtype with high accuracies and validated the results with an independent dataset on B-lineage ALL. This study of a consecutive series of childhood leukemias confirms and extends previous reports demonstrating that global gene expression profiling provides a valuable tool for genetic and clinical classification of childhood leukemias.

Gene expression analyses have been applied in many other kinds of cancer subtype classifications. For instance, Nutt et al [8] investigated whether gene expression profiling, coupled with class prediction methodology, can classify high-grade gliomas in a way more objective, explicit, and consistent than standard pathology. Supervised learning approaches were used to build a two-class prediction model based on a subset of glioblastomas and anaplastic oligodendrogliomas with classic histology. The experimental results showed that the model provided a more accurate prediction of prognosis in these nonclassic lesions than did pathological classification. They suggested that class prediction models, based on defined molecular profiles, classify diagnostically challenging malignant gliomas in a manner that better correlates with clinical outcome than does standard pathology. Another application in GBM (Glioblastoma multiforme), is the study performed by Liang et al [9], which used gene expression analysis to perform on surgical samples of brain tumors to characterize inter-tumor variability at molecular level. They revealed molecular signatures that reflect underlying pathogenic mechanisms and identified molecular markers that associate with survival. Agglomerative hierarchical clustering was combined with a two-step algorithm to identify survival-associated genes, in which Cox regression coefficients were calculated for all clustered genes, and a moving average of these values was calculated for consecutive genes in the hierarchically clustered list. There are many other applications, include melanoma [5], ovarian cancer [10], lung cancers [11, 12, 13, 14, 15], and gastric cancers [16], colon cancer[17, 18, 19] etc.

The current clinical and histopathological criteria used today for diagnosis and prognosis of tumor patients are inadequate for personalized treatment and clinical outcome prediction. van’t Veer [20]’s work demonstrated that the gene expression profile will outperform all currently used clinical parameters in predicting disease outcome. In their study [20], van’t Veer and colleagues identified 70 gene prognosis-signatures as a predictor of survival in breast cancer. They used DNA microarray analysis on primary breast tumors of 117 young patients, and applied supervised classification to identify a gene expression signature strongly predictive of distant metastases in patients without tumor cells in local lymph nodes at diagnosis. The 70 gene prognosis profile has been further validated with 295 patient samples in van de Vijver et al’s [21] work. They concluded that the gene-expression profile is a more powerful predictor of the outcome of disease in young breast cancer patients than the standard systems based on clinical and histological criteria. Gene-expression profiling
can help distinguish between patients at high risk and those at low risk for developing distant metastases, and so identify patients for adjuvant therapy.

3 Spectrum of data and its data analysis applications

Data in today’s life sciences consists of several types. Fig.1. shows a conceptual view of this data that consists of four types:

(i) Medical records which represent information such as observations and treatment procedures, mostly in the form of text and set of attributes. This information is becoming more and more digitized with today’s electronic health records,

(ii) Clinical data, such as pathologists reports, lab reports, X-rays, CT and MRI scans, etc, also mostly digitized, from which data analysis features could be derived. This information may be linked and also integrated with medical records,

(iii) High throughput data, which is perhaps the most granular pieces of information in today’s life sciences that one could expect to generate for selected patients. This information is not as easily available and it is normally obtained through an elaborate and sometimes complex process. Examples are converting a tissue obtained through a biopsy to a microarray and its associated data.

(iv) There may still be other types of data that are not related to any of the above categories and yet may be very helpful in our data analysis. These are data that are not collected under any specific circumstances. Genetic information is one example.

The focus of this short paper is only on the genomics aspect of the high throughput biological data. Furthermore, our intention is to discuss:

- gene expression analysis
- classification and prediction
- patient’s medical conditions,

More specifically, our objective would be to investigate topics such as a person: (i) being sick or healthy, (ii) belonging to any sub-class of a particular disease, or (iii) monitoring patients when they are treated, their disease status, progression etc.

4 A proposed framework and its application

In this section, we outline a generic framework to exploit the microarray technology for the stratification of disease types or the prediction of a patient’s medical condition such as drug resistance or predisposition to the development of certain illnesses.
The proposed framework consists of 3 basic steps. First from the gene expression data, we identify sets of highly discriminatory genes that can differentiate the different medical conditions under consideration. This process is summarized in Fig. 2. Along with many other researchers, we have developed methodologies to first identify list of differentially-expressed gene lists and then to narrow down these lists to sets of highly discriminatory gene sets.
The second step is to feed the gene sets identified in Step 1 to a machine learning (ML) module that would discover more reliable models. This would be the core of our reasoning module. The role of the reasoning module is to predict patients’ outcome based on gene expression intensity of the diagnostics panel and patients’ clinical information. One can then construct clinical test kits that could be used by clinical laboratories or at doctors’ offices to perform the more accurate diagnosis.

To illustrate the applicability of the framework, we have experimented the first 2 steps of the framework with a case study of patients’ response to chemotherapy. The study is based on the dataset published in [22]. A group of AML patients were treated with intensive cytarabine-based induction and consolidation. 33 patients were selected for the study: 22 had good response and 11 had poor response. Preliminary analysis indicated promising results. The process identified a prognostics gene panel of 51 genes and constructed a companion reasoning module consisting of 34 models. The overall sensitivity, specificity, positive and negative predictive rates were all high. Detailed of the study will be reported when completed.

5 Discussion and future work

In this paper we have briefly provided an overview of data and its scope in today’s life sciences and highlighted an impact that gene expression data could make in the advancement of classification and prediction of patients’ medical conditions. We have discussed a framework that could be considered to develop and validate future clinical test kits. We have preformed a preliminary study using this framework on a group of AML patients where results were promising. We are currently investigating use of time series leukemia data to obtain models for patient response analysis. Our future work also includes an integrated approach where multiple forms of omics data will be used to generate consistent markers and sets of more biologically plausible signatures.
References


