Biological specifications for a synthetic gene expression data generation model

Francesca Ruffino[1], Marco Muselli[2] and Giorgio Valentini[1]

[1] DSI - Dipartimento di Scienze dell'Informazione,
Università degli Studi di Milano, 20135 Milano, Italy,
{ruffino,valentini}@dsi.unimi.it
[2] IEIIT, Istituto di Elettronica, Ingegneria dell'Informazione
e delle Telecomunicazioni, Consiglio Nazionale delle Ricerche, Genova, Italy
muselli@ice.ge.cnr.it

Abstract. An open problem in gene expression data analysis is the evaluation of the performance of gene selection methods applied to discover biologically relevant sets of genes. The problem is difficult, as the entire set of genes involved in specific biological processes is usually unknown or only partially known, making unfeasible a correct comparison between different gene selection methods. The natural solution to this problem consists in developing an artificial model to generate gene expression data, in order to know in advance the set of biologically relevant genes. The models proposed in the literature, even if useful for a preliminary evaluation of gene selection methods, did not explicitly consider the biological characteristics of gene expression data. The main aim of this work is to individuate the main biological characteristics that need to be considered to design a model for validating gene selection methods based on the analysis of DNA microarray data.

1 Introduction

Analysis of gene expression data may be performed at different levels, ranging from the analysis of differential expression of genes, to unsupervised and supervised analysis of sets of genes and tissues [1].

An important related problem is to determine the subset of genes involved in the biological process under examination. Such problem is generally referred to as gene selection and several statistic and machine learning techniques have been proposed in literature to face with it [2–4].

Unfortunately, the entire set of genes involved in specific biological processes is usually unknown or only partially known, and as a consequence the evaluation of the real effectiveness of gene selection methods is very difficult and in many cases unfeasible.

Several models have been proposed to simulate gene expression data, in order to make available synthetic gene expression data for classification, clustering and gene selection problems [5, 6]. Even if these models may be in principle helpful to test gene selection methods, their main limitation consists in a drastic

modelling simplification, without sufficiently taking into account the biological characteristics of gene expression data.

In this paper we address the problem of the analysis of the specifications needed to properly model the biological characteristics of gene expression data. In particular, the main concerns of this work are the relationships between the biological and modelling issues involved in the design of a flexible tool to generate synthetic gene expression data. To this end we performed an analysis of the gene expression literature to individuate structural commonalities in gene expression data. The design and the implementation of an artificial model will allow us to properly evaluate the performance of clustering and gene selection methods, as all subsets of the simulated genes involved in simulated biological processes will be known in advance.

2 Profiles and expression signatures

The main goal of gene selection methods is to find sets of genes significantly related to a specific functional status (e.g. diseased vs. healthy). In the biomolecular literature sets of biologically relevant and differentially expressed genes are named *expression signatures* [7–11].

To our knowledge the term *expression signature* has been introduced by Alizadeh et al. [7], to characterize gene expression patterns found by gene expression profiling. More precisely this term refers to a group of genes coordinately expressed in a given set of specimens and in a specific physiological or pathophysiological condition.

The correlation between the mRNA levels of the genes is due to the underlying regulatory system, by which the same set of transcription factors and binding sites may be directly or indirectly shared by the genes belonging to the same expression signature. Hence a gene expression signature indicates a cluster of coordinately expressed genes, whose coordination reveals the fact that they participate to the same biological process (and hence they are controlled by the same set of regulation factors); indeed they are usually named by either the cell type in which their component genes are expressed, or by the biological process in which their component genes are known to function.

From this standpoint the overall *expression profile* of a patient can be interpreted as a collection of gene expression signatures that reveal different biological features of the analyzed sample [7].

2.1 Gene expression signatures in human diseases

Expression signatures has been mainly discovered and analyzed in gene expression profiles of diseases. For instance, the expression profiling of B-cell malignancies through hierarchical clustering, revealed expression signatures related to cell-proliferation, to lymph-nodes, T-cells, Germinal Centre B-cells (GCB) and others [7].

Independent Component Analysis performed on gene expression data from ovarian cancer tissues found gene expression signatures representing potential pathophysiological processes in ovarian tissue samples [8]. Expression profiling of rhabdomyosarcoma (RMS), the most common soft tissue sarcoma in children, identified two signatures associated with metastatic RMS, responsible for most of the fatal outcome of this disease [11], while two way hierarchical clustering analysis identified several expression signatures expressed in different types of bladder carcinoma [9].

Expression signatures have been also identified in species other than humans and in contexts not related to tumoral differentiation. For instance comparative functional genomics based on shared patterns of regulations across orthologous genes identified shared expression signatures of aging in orthologous genes of *D. melanogaster* and *C. elegans* [10].

Summarizing, *expression profiles* and *expression signatures* seem to be well-established biological structures that characterize gene expression data.

2.2 Characteristics of gene expression signatures.

In this section we discuss the main characteristics of gene expression signatures.

Differential expression and co-expression. Differential expression analysis of single genes, even if it may be useful to identify specific genes involved in biological processes [12], cannot capture the complexity of tightly regulated processes, crucial for the proper functioning of a cell.

Correlations between gene expression levels have been observed [13, 7], reflecting the fact that in most biological processes genes are co-regulated. As recently observed, not all changes in co-regulation are manifested by up or down regulation of individual genes, and we need to explicitly consider interactions between genes to discover patterns in the data [14].

Hence, we need sets of co-regulated genes, that is expression signatures, to reveal functional relationships between genes.

Gene expression signatures as a whole rather than single genes contain predictive information. Many times is the signature taken as a whole that seems to contain predictive information for a biologically meaningful identification of tissue samples. For instance, it was found an expression signature of 8 upergulated and 9 downregulated genes associated with metastasis in different types of adenocarcinoma: none of these genes represents a marker, but it is the signature as a whole that represents a "collective marker" of tumor metastasis [15].

In other works [15, 14] it has been shown that in some cases relevant differences are subtle at the level of individual genes but coordinate in gene expression groups.

Genes may belong to different gene expression signatures at the same time. Many genes may be involved in a number of distinct behaviours, depending on the specific conditions of the tissue. From this standpoint they may belong to different expression signatures [16]. Indeed each gene may be influenced by several transcription factors, each of which influences several genes [8]. Moreover many underlying conditions in a given sample may concur to define a gene expression signature (e.g. tumorigenesis, angiogenesis, apoptosis) [17].

Expression signatures may be independent of clinical parameters. An expression signature of 153 genes can be used to correctly classify hepatocellular carcinoma (HCC) intra-hepatic metastasis from metastatic-free HCC [18]. This expression signature, that embeds high predictive information, has been shown to be independent of tumor size, tumor encapsulation and patient age, and also very similar to that of their corresponding metastases.

Several other works showed that a bio-molecular characterization of tumours can discover different subtypes of malignancies, not detectable with traditional morphological and histopathological features (see e.g. [7, 2]).

Different gene expression profiles may share signatures and may differ only for few signatures . It has been shown that gene expression signatures may be shared and partially expressed in different gene expression profiles [7, 15, 18].

For instance, it has been shown that Diffuse Large B-Cell Lymphoma (DLBCL) subgroups (GCB-like and activated B-like DLBCL) share most of the expression signatures but they differ mainly for two signatures (GCB and activated B-cell signatures) partially expressed respectively in germinal centre B-cell and activated peripheral blood B cell [7].

Moreover, hierarchical clustering, in the space of a 128 genes signature of metastatic adenocarcinoma nodules of diverse origin, showed two clusters of primary tumors that were highly correlated with metastatic ones: this fact, together with a differential overall survival in primary adenocarcinoma tumors showed that this gene expression signature is present in subpopulation of primary tumors [15].

Hence gene expression profiles of functionally different tissues may share expression signature and differ only for a subset of expression signatures. These expression signatures may be also partially expressed (that is, not all the genes belonging to the expression signature are over-expressed or under-expressed), reflecting functional alterations in diseased patients.

3 Biological and modelling issues

In light of the characteristics of gene expression signatures (Sect. 2.2), in this section we discuss the relationships between the biological and modelling issues we need to consider to design an artificial model for gene expression data synthesis. Schematically, we identified the following main items:

1. Expression profiles may be characterized as a set of gene expression signatures. A set of gene expression signatures defines a *functional group* of samples. The model should allow us to define expression profiles in terms of expression signatures, with a large flexibility with respect to the number and gene composition of the synthetic expression signatures.
2. Expression signatures are interpreted in the literature as a set of coexpressed genes. These genes may be overexpressed and underexpressed with respect to the other genes and with respect to a particular condition. Accordingly, in the model, each expression signature should be defined as a set of overexpressed or underexpressed genes, that is genes with gene expression levels above or below a given threshold. The model should define a signature *active* if its genes are coordinately over(under)expressed.
3. Expression signatures may be defined either by the overall available knowledge about bio-molecular processes (e.g. by Gene Ontology categories) or may be discovered through statistical and machine learning methods. The model should permit to define arbitrary signatures, in order to allow us a large range of applications in different biological contexts.
4. Genes may belong to different signatures at the same time. As a consequence the model should allow us to assign the same gene to different signatures.
5. The model should permit to select from few few units to few hundreds of genes for each gene expression signatures, as the number of genes within a signature usually vary within this range.
6. Apart from technical variation (that in principle should be detected and canceled by proper design and implementation of bio-technological experiments and suitable pre-processing procedures [19]), gene expression is biologically variable also within functional classes (conditions) [20]. The model should reproduce the variation of gene expression data. Variation of single genes may be simulated sampling from a predefined distribution. Our preliminary analysis of gene expression data showed that gene expression values are close to be normally distributed, but it would be useful to analyze a larger number of gene expression data to properly evaluate this item.
7. Not always expression signatures show large variations of gene expression levels: some signatures may present modest but coordinate variation. The model should be sufficiently flexible to allow small variations of coexpressed genes, and to this end it should include tunable parameters of the gene distributions.
8. Not all the genes within a signature may be expressed in all samples. Moreover gene expression variation between individuals may introduce variation into expression signatures. The model should permit to introduce flexibility in the number of genes that can be underexpressed or overexpressed, as well as to introduce individual variability within a functional group.
9. Different expression profiles may differ only for few signatures, that is different functional groups may share the same (or very similar) expression signatures. The model should allow to define an expression profile as a set of signatures and to define other functional groups in terms of subsets of previously defined signatures, eventually modifying or adding new signatures.

10. Some signatures may be only partially expressed within a particular expression profile. The model should be sufficiently flexible to allow us to define an expression profile in several ways: (a) a set of active signatures; (b) a set of randomly active signatures; (c) a set of randomly active signatures with a set of "mandatory" active signatures.

4 Conclusions

In this paper we analyzed the biological issues underlying the modelling of an artificial system for simulating gene expression data.

We identified the expression signatures as a major common biological structure in gene expression data and we provided the biological specifications to develop an artificial model for gene expression data synthesis.

The next step of this works consists in developing and implementing a biologically motivated gene expression data generation model, to properly evaluate the performance of gene selection methods.

References

1. Baldi, P., Hatfield, G.: DNA Microarrays and Gene Expression. Cambridge University Press, Cambridge, UK (2002)
2. Golub, T., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science **286** (1999) 531–537
3. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning **46** (2002) 389–422
4. Muselli, M.: Gene selection through Switched Neural Networks. In: NETTAB-2003, Workshop on Bioinformatics for Microarrays, Bologna, Italy (2003)
5. Weston, J. et al.: Use of the zero-norm with linear models and kernels methods. Journal of Machine Learning Research **3** (2003) 1439–1461
6. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics **19** (2003) 1090–1099
7. Alizadeh, A. et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403** (2000) 503–511
8. Martoglio, A., Miskin, J., Smith, S., MacKay, D.: A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. Bioinformatics **18** (2002) 1617–1624
9. Dyrskjøt, L. et al.: Identifying distinct classes of bladder carcinoma using microarrays. Nature Genetics **33** (2003) 90–96
10. McCarroll, S. et al.: Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nature Genetics **36** (2004) 197–204
11. Yu, Y. et al.: Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homoprotein Six-1 as key metastatic regulators. Nature Medicine **10** (2004) 175–181
12. Cui, X., Churchill, G.: Statistical tests for differential expression in cDNA microarray experiments. Genome Biology **4** (2003)
13. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS **95** (1998) 14863–14868

14. Kotska, D., Spang, R.: Finding disease specific alterations in the co–expression of genes. Bioinformatics **20** (2004) i194–i199
15. Ramaswamy, S., Ross, K., Lander, E., Golub, T.: A molecular signature of metastasis in primary solid tumors. Nature Genetics **33** (2003) 49–54
16. Gasch, P., Eisen, M.: Exploring the conditional regulation of yeast gene expression through fuzzy k-means clustering. Genome Biology **3** (2002)
17. Ihmels, J., Bergmann, S., Barkai, N.: Defining transciption modules using large-scale gene expression data. Bioinformatics (2004)
18. Ye, Q. et al.: Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. Nature Medicine **9** (2003) 416–423
19. Chen, J., et al.: Analysis of variance components in gene expression data. Bioinformatics **20** (2004) 1436–1446
20. Cheung, V., et al.: Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genetics **33** (2003) 422–425