The class of Microarray games and the relevance index for genes[1,2]

Stefano Moretti[3,4], Fioravante Patrone[5] and Stefano Bonassi[4]

Abstract: Nowadays, microarray technology is available to generate a huge amount of information on gene expression. This information must be statistically processed and analyzed, in particular to identify those genes which are useful for the diagnosis and prognosis of specific diseases. We discuss the possibility of applying game-theoretical tools, like the Shapley value, to the analysis of gene expression data.

Via a "truncation" technique, we build a coalitional game whose aim is to stress the relevance ("sufficiency") of groups of genes for the specific disease we are interested in. The Shapley value of this game is used to select those genes which deserve further investigation. To justify the use of the Shapley value in this context, we axiomatically characterize it using properties with a genetic interpretation.

Key-words: coalitional game, Shapley value, power index, gene expression, microarray

1 Introduction

Proteins are the structural constituents of cells and tissues and may act as necessary enzymes for biochemical reactions in biological systems. Most genes contain the information for making a specific protein. This information is coded in genes by means of the deoxyribonucleic acid (DNA). *Gene expression* occurs when genetic information contained within DNA is *transcripted* into messenger ribonucleic acid (mRNA) molecules and then *translated* into the proteins.

[1] The authors are grateful to two anonymous referees for their extremely helpful comments.

[2] An earlier version of this paper was presented at the VI Spanish Meeting on Game Theory and Practice, July 12-14, 2004, Elche, Spain.

[3] This author gratefully acknowledges the financial support of the EU project NewGeneris, European Union 6^{th} FP (FOOD-CT-2005-016320).

[4] Unit of Molecular Epidemiology, National Cancer Research Institute, Largo R. Benzi 10, 16132 Genoa, Italy. E-mails: stefano.moretti@istge.it; stefano.bonassi@istge.it

[5] DIPTEM, University of Genova, P.le Kennedy - Pad D, 16129, Genoa, Italy. E-mail: patrone@diptem.unige.it

Nowadays, the microarray technology allows for the quantification of the expression (*i.e.* the amount of mRNA) for genes under the same biological condition (for instance, a tumor). A microarray works by exploiting the ability of a given mRNA molecule to bind specifically to, or hybridize to, the DNA template from which it originated. By using an array containing many DNA samples, it can be determined, in a single experiment, the expression levels of hundreds or thousands of genes within a cell by measuring the amount of mRNA bound to each site on the array.

There are several different experimental platforms based on microarray technology (see, for instance, Parmigiani *et al.* (2003)). However, a common objective of gene expression microarrays is to consistently generate a matrix of expression data, in which the rows (possibly thousands) index the genes and the columns (usually in the order of tens) index the study samples. Numbers in the matrix represent gene expression values which quantify the level of expression of genes in the samples.

The aim of this work is to address the problem of quantifying the relative relevance of genes in a complex scenario - such as the pathogenesis of a genetic disease - on the basis of the information provided by microarrary experiments, taking into account the *level of interaction* among the genes.

Complex experimental artifacts associated with microarray data collection emphasize the need for pre-processing analysis of the data (for instance, the design of the arrays, the quality assessment of the rough data, the normalization procedures), with the goal to reduce systematic errors arising from several experimental procedures. Despite the reduction of experimental bias has been the objective of several works on microarray analysis in the last few years (see, for example, Dudoit *et al.* (2001); Smith and Speed (2003); Parmigiani *et al.* (2003)), in practice the problem of completely removing the experimental variability is still unsolved and a statistical treatment of the data provided by microarrays is required.

For this reason, in our approach we refer to the *observed average* level of interaction of a group of genes, *i.e.*, the average number of tumor samples in which such a group of genes can be considered responsible, according to a pre-defined causality principle, for the onset of the tumor: the higher the number of samples observed, the lower the probability that chance could affect the inferences provided by the model.

The basic idea of this model comes from the theory of coalitional games. In particular we consider the framework of simple games, which have been widely applied to the analysis of the power of players in interaction situations as Councils, Parliament, etc. (Shapley and Shubik (1954), Banzhaf (1965); see Owen (1995) for a general introduction to this topic and a summary of these results). We adopt the same formal language of coalitional games for modelling the interaction among genes, considered as players, in connection with a biological condition of interest, *e.g.* the pathogenesis of a genetic disease or tumor. The game we consider origins from the comparison of two matrices of gene expression data; one from tumor samples and the other one from normal DNA (referent healthy subjects). We first use a discriminant method on each sample to split the whole set of genes in two sets, *i.e.*, those genes showing an expression ratio largely different from normal samples, and those with expression levels corresponding to normal DNA samples. At this preliminary stage of the model, for each single gene, as in detail explained in Section 2, we use the interval boundaries containing most data in the normal distribution of that gene as cut-offs for discrimination (Becquet *et al.* (2002)). We then introduce a causality relation (also called *sufficiency principle*) which directly determines the characteristic function of the game. An interpretation of the biological meaning of a relevance index, used for measuring the "power" of each gene in inducing the tumor, is given and it turns out to coincide with the Shapley value of the game considered.

We start with some preliminary notations in the next section. In Section 3 the class of microarray games is introduced starting from the general notion of the *sufficiency principle*, and some basic properties and examples of such games are reported. In Section 4 an axiomatic characterization of the Shapley value is given by means of five properties suitable to genetic interpretation of this index. Section 5 concludes with some considerations on related works and future research.

2 Preliminary notations

Let $N = \{1, 2, \dots, n\}$ be a set of n genes, $S_R = \{s_1^R, s_2^R, \dots, s_r^R\}$ be a set of r reference samples, *i.e.* the set of cells from normal tissues and let $S_D = \{s_1^D, s_2^D, \dots, s_d^D\}$ be the set of cells from tissues with a genetic disease.

The goal of a microarray experiment is to associate to each sample $j \in S_R \cup S_D$ an *expression profile* $A(j) = (A_{ij})_{i \in N}$, where $A_{ij} \in \mathbb{R}$ represents the *expression value* of the gene i in sample j. Globally, such expression values will be indicated as the *data set* of the microarray experiment. In the following we will refer to the data set resulting from the pre-processed method usually called normalization (Dudoit *et al.* (2001), Smith and Speed (2003)), which allows for comparison among expression intensities of genes from different samples. The data set can be expressed in the form of two real valued expression matrices $\mathbf{A}^{S_R} = (A_{ij}^{S_R})_{i \in N, j \in S_R}$ and $\mathbf{A}^{S_D} = (A_{ij}^{S_D})_{i \in N, j \in S_D}$. In summary, we will denote as a *microarray experimental situation* (MES) the tuple $E = \langle N, S_R, S_D, \mathbf{A}^{S_R}, \mathbf{A}^{S_D} \rangle$.