Game Theory applied to gene expression analysis

Stefano Moretti

Promotor: Prof. Fioravante Patrone

Copromotor: Dr. Stefano Bonassi

University of Genoa
Department of Mathematics
Doctorate in Mathematics and Applications
MAT/09

To Alessandra and Giovanna

Acknowledgements

Three years as a Ph.D student involved in a multidisciplinary field at the Department of Mathematics and at the Unit of Molecular Epidemiology gave me many opportunities to exchange opinions with a lot of people from completely different research area. Here I would like to take the opportunity to thank some persons that have been very important for the accomplishment of this thesis.

First of all, I want to thank my friend and promotor Fioravante Patrone, for his enthusiastic, conscious and inspired supervision. At the time when I started my doctorate, I had already been co-worker of Fioravante for few enjoyable years, and I am not able to fully express my gratitude to him for his superb guidance in doing research in Game Theory, applied Mathematics and many other topics.

My thanks also to Stefano Bonassi, for his continuing support and for having provided a tangible opportunity to focus my efforts on the application of Game Theory to gene expression analysis.

I am grateful to people of the Unit of Translational Paediatric Oncology of the National Institute for Cancer Research (IST), in particular to Paola Scaruffi and Gian Paolo Tonini for their cooperation and assistance in explaining me many biological aspects on which microarray technology is based.

My sincere appreciation to Franco Fragnelli, Roberto Lucchetti and Stef Tijs, for profitable discussions on game theoretical topics of interest for my work.

Furthermore, I want to express my gratitude to all my colleagues for creating a nice environment for doing research.

I am grateful to my parents for their unceasing support, and to have given me the opportunity to enjoy Science in a loving environment when I was a young student.

Last, but for sure not least, I thank my wife Alessandra - computer scientist and mother - for her efforts in giving me all her strength all the time, also in the most difficult moments, successfully fulfilling her difficult job, and never subtracting a carefulness to our daughter Giovanna and me. She is without any doubt the most astonishingly talented person I have ever known.

Stefano Moretti
April 2006
Genoa

Contents

Chapter 1

Introduction

Nowadays, microarray technology is available for taking 'pictures' of gene expressions. Within a single experiment of this sophisticated technology, the level of expression of thousands of genes can be estimated in a sample of cells under a given condition. This monograph deals with the discussion and the application of a methodology based on Game Theory for the analysis of gene expression data. Roughly speaking, the starting point is the observation of a 'picture' of gene expressions in a sample of cells under a biological condition of interest, for example a tumor. Then, Game Theory plays a primary role to quantitatively evaluate the relevance of each gene in regulating or provoking the condition of interest, taking into account the observed relationships in all subgroups of genes.

To fully understand the methodology introduced in this thesis, some prerequisites both on Game Theory and on microarray data analysis are required. In order to create a common background for readers who approach for the first time Game Theory or microarray data analysis or both of them, I suggest to look at Sections 1.1 and 1.2, aimed, respectively, to give a basic introduction to cooperative Game Theory and to the statistical analysis of microarray expression data. The contents of those sections are fundamental to understand the objectives of this thesis, which are described in Section 1.3. Finally, in order to understand the theory behind the game theoretical model applied to gene expression data, the preliminary definitions introduced in Section 1.4 and in Section 1.5 are required.

1.1 Introduction to Game Theory

Game Theory is a mathematical theory dealing with models for studying interaction among decision makers (which are called *players*). Dealing with decision makers interaction, the reader should be aware that a decision problem that involves only one decision maker is not properly in the domain of application of Game Theory.

Since the seminal book by John von Neumann and Oskar Morgenstern (1944) "Theory of Games and Economic Behavior", it is usual to divide Game Theory into two main groups of interaction situations (which are called *games*), *non-cooperative* and *cooperative* games. Non cooperative games deal with conflict situations where non binding agreements among the players can be made. In cooperative games all kinds of agreement among the players are possible.

In non cooperative games, each player will choose to act in his own interest keeping into account that the outcome of the game depends on the actions of all the players involved. Actions by players can be simultaneous (for instance the 'stone, paper, scissors' game or the 'matching pennies' game) or at several points in time (for instance the game of chess).

Cooperative games deal with situations where groups of players (which are called *coalitions*) coordinate their actions with the objective to end up in joint profits which often exceed the sum of individual 'profits'[1].

Another important classification in Game Theory about the goals of the analysis performed using its tools. A game, both non-cooperative and cooperative, can be analyzed with the objective to indicate what players should do in the game to maximize their profits (usually this goal is referred to as the 'normative approach'). Another reason for using Game Theory, is to predict the outcome of game, i.e. whether or not players optimize their profits (usually referred to as the 'predictive approach').

In this dissertation I focus on the application of cooperative game theory to the analysis of gene expression data from microarray experiments. If it is quite obvious that I do not plan to give advice to genes on how they should behave inside a biological cell, on the other hand it is not so straightforward to figure out how to describe the behavior of genes making them to play a certain game,

[1] For game theorists, *utility value* would be more correct than the term profit. As for the ordinary language, I use for the moment the term profit with reference to something that is in the interest of the decision maker to be maximized.

and then use it as a tool to predict which genes obtain the maximum profit.

First, it is not obvious at all what is the meaning of 'profit' in this context. In the following sections I will extensively introduce and discuss this topic. However, the possibility to extend the concept of profits, benefits, savings or whatever could be in the interest of each decision maker to be maximized on her/his own count, is a well known feature of Game Theory applications. In Game Theory, the term 'profit' usually is more correctly replaced by utility value of a *rational* player. I do not want to enter here the discussion of how an utility function is defined and why it is a numerical representation of the *preferences* of a rational decision maker. For introductions to this problem see for instance the books by Kreps (1990) and Osborne and Rubistein (1994).

I will simply note that, sometimes, reasonable considerations bring game theorists to assume that the players preferences are nicely represented just by money, and so money will be the profits to be considered in the game. For example, one can describe a situations using cooperative games in coalitional form where the players are willing to join bigger coalitions in order to have extra monetary benefits or extra monetary savings thanks to the effects of cooperation. For instance, consider a cooperative game in coalitional form with three players, 1, 2 and 3, and with a *characteristic function* $v : \mathcal{P}(\{1,2,3\}) \rightarrow \{0,1\}$, where $\mathcal{P}(\{1,2,3\})$ is the set of all possible subsets of $\{1,2,3\}$, and such that each coalition with at least two players get 1 euro, and all the remaining coalitions get 0 euro (i.e. all the single player coalitions and the empty coalition get 0). Formally, we are considering the cooperative game in coalitional form $(\{1,2,3\}, v)$ such that $v(\{1,2,3\}) = v(\{1,2\}) = v(\{1,3\}) = v(\{2,3\}) = 1$ and $v(\{1\}) = v(\{2\}) = v(\{3\}) = v(\emptyset) = 0$ (for this kind of problems see books by Owen (1993), Tijs (2003), Young (1995)).

Other times, preferences of players are not addressed to things that have a monetary counterpart. This is the case, for example, of decisions in a parliament. Assume that there are three parties, A, B and C, which share the seats in parliament by 45%, 40%, and 15%. The preferred outcome for a party or a coalition of parties is intended as the ability to force a decision. In this case, I will say that the coalition is a winning one. Suppose that decisions are made by simple majority. No one of single parties will profit from missing the cooperation with others, in the sense that all parties alone are loosing coalitions. On the contrary, all coalitions with more than one party inside will be a winning coalition.

This parliament situation can be properly represented by a cooperative game in coalitional form, where players are the three parties A, B and C and the value of each sub-set of players (coalition) is the label of winning or loosing coalition. Consider 1 as label for winning coalitions, and 0 as label for loosing coalitions. So, only coalitions with at least two players get 1 and the remaining coalitions get 0. We are indeed considering the game $(\{A, B, C\}, w)$ such that $w(\{A, B, C\}) = w(\{A, B\}) = w(\{A, C\}) = w(\{B, C\}) = 1$ and $w(\{A\}) = w(\{B\}) = w(\{C\}) = w(\emptyset) = 0$. Note that this game has precisely the same structure of the game $(\{1, 2, 3\}, v)$ introduced before. In both cooperative games in coalitional form there are three players (different names, in this case, are not essential), and in both games only coalitions with at least two players get 1, and the others get 0.

What is basically changed, making the 'same' game suitable for the description of such completely different situations, is just the definition of the objectives of each coalition in relation to the preferences of its players. In game $(\{1, 2, 3\}, v)$, it has been assumed that the objective of the players is to maximize their rewards; in terms of preferences it has been assumed that each player prefers 1 euro to nothing. In game $(\{A, B, C\}, w)$, it has been assumed that the objective of the players is to force a decision in the parliament, so players prefer to have the ability to force a decision than not to have it. Concerning this kind of models, there are many other important aspects that cannot be taken up in a basic introduction on Game Theory. But I think that these very preliminary considerations are already sufficient to give a first insight on the extreme flexibility of the formal definition of cooperative game in coalitional form in representing completely different interaction situations.

Now, some words on what it is possible to predict using cooperative games in coalitional form.

Consider again the example of the parliament. Since the decision rule was the simple majority it seems not very likely that the distribution of power, however defined, coincides with the distribution of seats for parties A, B and C. In order to discuss issues related to the problem of assigning power to the players of similar cooperative games, and understand how the power distribution changes when the number of seats or the decision rule change, classical analytical tools developed in the Game Theory framework are *power indices* (see for instance

Felsenthal and Machover (1998) for a formal discussion of the problem; Owen (1993) for some political applications). The most popular, widely applied to many political institutions (e.g USA President Elections, ONU Council, EU Parliament etc.) are the Shapley-Shubik power index (Shapley and Shubik (1954)) and Banzhaf-Coleman power index (Banzhaf (1965)). Surprisingly, most power indices are nothing else that well known *solution concepts* for cooperative games in coalitional forms. This means that the same method can be used to allocate among the players the profits of the big coalition in games where the value of each coalition represents, for example, monetary rewards.

Which arguments can support the application of the same solution concept to so different interaction situations and their consequent alternative interpretations?

The answer to this question is rooted in the *property driven* approach[2]. If the quantification of power is the goal of the analysis, the property driven approach suggests to postulate discriminating properties which a power measure has to satisfy in order to qualify as an appropriate measure. If the cooperative game concerns monetary profits and the objective is to *fairly* allocate the total reward of cooperation, of course the basic properties to be postulated can be different and their interpretation must be appropriate to the context. On the strength of the property driven approach, it often happens that a solution concept satisfies sound properties in completely different situations (see for instance the volume by Roth (1988) for different applications of the Shapley value). The strong connection with the property driven approach is in my opinion one of the main reasons of success of applied Game Theory, success which is widely manifested by the several applications of Game Theory to different scientific fields, especially in Economics, Political Science, Social Science and Evolutionary Biology. Next, I will try to convince the reader that it can also be successfully applied to gene expression analysis.

1.2 Introduction to microarray data analysis

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

[2] In Game Theory this approach is also known as the *axiomatic* method

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.

Nowadays, a revolutionary technique, i.e., the microarray technology, allows for the collection of huge amount of information concerning the function of human genes. This approach provides a quantitative measure of gene expression (the amount of mRNA in a cell sample) for thousands of genes in the same experiment. The crucial step of this procedure is the *hybridization*: many DNA regions immobilized on a small glass, plastic or nylon matrix (probes), bind to a complementary sequence from the sample under study (sampled mRNA itself or cDNA obtained by inverse transcription of sampled mRNA), labelled with fluorescent dyes that flag their presence when exposed to a specific wavelength of light. A separate experiment takes place in each of many individual spots arrayed as a regular pattern on the matrix, whence the name array (Parmigiani *et al.* (2003)).

There are several microarray based technologies, which involve different experimental procedures (see for instance Schena (2003), 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 units or tens) index the study samples. Numbers in the matrix represent gene expression ratios which quantify the relative expression of genes in one target sample with respect to a given reference sample.

Complex experimental artifacts associated with microarray data collection have been described, emphasizing the need for statistical treatment of data during all stages of the experiment. This includes the design of the slide, the quality assessment, the normalization process (Dudoit *et al.* (2001); Smith and Speed (2003), Amaratunga and Cabrera (2004)) and other pre-processing data analysis (Amaratunga and Cabrera (2004), Parmigiani *et al.* (2003)) with the objective of removing systematic variation in microarray experiments. In the following of this paper I will assume to work on a matrix of gene expression values that have been already pre-processed.

Many models for data analysis have been presented in the literature for inferring, from a matrix of gene expression data, the role of genes, their interactions and their behavior when changes in condition of the biological system occur

(Moler *et al.* (2000), Su *et al.* (2003)).

So far, classical statistical techniques used for extracting information from gene expression microarrays can be classified in three main groups: *inferential statistical methods* used for identifying genes that are regulated by different conditions of interest, e.g., to find single genes or groups of genes which show a statistically significant difference in the expression levels under two or more conditions of interest (Fujarewicz and Wiench (2003), Storey and Tibshirani (2003)); *unsupervised analysis techniques*, used as a method to identify groups of genes with similar patterns in the expression data (Golub *et al.* (1999), Alon *et al.* (1999)); *class prediction tools*, where selected genes are used to classify samples into known categories of morphology, known biological features, clinical outcomes, or other condition of interests according to gene expression patterns. It is mostly aimed at supporting early diagnosis in new samples (Dudoit and Fridlyand (2003), Golub *et al.* (1999), Dudoit *et al.* (2002b)).

In order to give a slightly more accurate idea about how these classical statistical methods have currently been applied to microarray data analysis, I follow the essential outline of the presentation of the methods in what I consider one of the most complete books on microarray analysis at the moment, i.e. the book by Amaratunga and Cabrera (2004).

Concerning the inferential statistical methods, the main task of these methods is usually accomplished by mean of *statistical hypothesis testing.* The result of an hypothesis testing on a gene expression matrix is its decision among two possible options: to reject the conjecture (*null hypothesis*) that there is no differences in terms of gene expression between two conditions of interest or not to reject the null hypothesis and declare that there is insufficient evidence to detect a difference of gene expression between the two conditions. In order to select or develop a good test for a particular microarray data-set, it is necessary to make assumptions about that microarray data-set. Different assumptions for the same situation will generally lead to quite different tests and perhaps even quite different test results. In general it is important to consider assumptions carefully, but this is a very difficult task on microarray analysis where the biological knowledge that could be used as diagnostics to check the assumptions is still vague and strongly dependent from the biological conditions of interest.

Unsupervised analysis techniques, also known as *pattern discovery* or *cluster analysis*, has as a main objective to produce evidences for correlated patterns of gene expression displayed by genes behaving jointly, such as genes performing similar functions or genes operating along a genetic pathway. Based on the quantifications of similarity between observations, most of these methods depend on either a dissimilarity or similarity measure, which quantifies how far, or how close, two observations (for example vectors of gene expressions across different samples) are from each other. Dissimilarity measures which have been employed in microarray analysis are classical distances like the *Euclidean distance* (Coco *et al.* (2005)). It is matter of fact that different definitions of dissimilarity measures bring to different clusters of similar genes. The notion of similarity or dissimilarity used, however, should reflect an a priori selected attribute for joint gene behavior that it is expected to be informative with respect to the biological condition under investigation. So far, it is not clear which analytical instruments should be used to evaluate the meaning of a given dissimilarity or similarity measure, and the choice of a metric is still almost completely arbitrary.

Finally, few words on supervised analysis, also called *class prediction*. To better understand the main characteristics of this kind of analysis, I found more explanatory to refer to the biological conditions of interest directly as tumors. In fact, the tumors are known to be of various different classes and a microarray gene expression data-set can be extracted from samples collected from different tumors. Now it is likely that different genes are expressed in the cells of different tumor classes. Therefore it can be conjectured that it ought to be possible to differentiate among the tumors classes by studying and contrasting their gene expression profiles, that is developing a classification rule to discriminate them. The great potential of these methods is that the classification rule could be exploited to predict the class of a new tumor sample of unknown class based on its gene expression profile. Another advantage of these methods is the easy way to evaluate their performance, as the proportion of misclassifications on the gene expression matrix where the original tumor class of samples is known (*training set*), i.e. the *misclassification rate*. On the other hand, from the mathematical point of view, the biggest problem in the applications of supervised methods to gene expression data-set is the number of genes much greater than the number of samples. By retaining such a large number of genes, it is incredibly easy for supervised methods to find good-looking but non-reproducible and meaningless classification rules, with low misclassification rate on the training set and very high misclassification rate on the gene expression data where the information on tumor classes is unknown (*test set*). From this follows the necessity to find a strategy to reduce the number of genes, for example, performing the supervised procedure only on those genes which result differentially expressed on the basis of the application of inferential statistical methods. Besides the problems concerning the assumptions which affect the statistical inferential methods as I mentioned before, this filtering approach encounters also other disadvantages. Some genes retained could be false positive, and even so produce good performance as classifiers, performance that of course are not reproducible on other data-set. Even worst, it may exist a set of genes that together acts as a classifier, but each individual gene in the set does not, making them good candidate for being filtered out all together. Moreover, many retained genes could show the same pattern of expression, determining a redundancy in the information.

1.3 Objectives and overview of the thesis

The criterium for the choice of one particular statistical method should be based on the (justified) claim that such method is able to select genes covering the most relevant role in the mechanisms which provoke a biological condition or response of interest (e.g. a tumor). Unfortunately, the big difficulty in taking the decision is that classical statistical methods are not directly related with a biologically sound and operative definition of genes *relevance* in this context.

Consequently, different sets of genes may be selected depending on the application of different statistical methods (Jaeger *et al.* (2003)). Since usually there exists a limit on the number of genes to choose, a researcher might not be able to include all relevant genes in deserving further investigations.

For example, an extremely difficult question to answer is whether a group of genes which are individually differentially expressed between two different conditions are more or less relevant in regulating the mechanisms governing these conditions than another group of genes able to characterize the two conditions only jointly. Differently stated, similarly to the considerations done for the classification problem, it may exist a set of genes A that together have a characteristic expression pattern under each condition, but each individual gene in the set has not. On the contrary, it may exist a set of genes B where each individual gene is differentially expressed under the two conditions. So, the problem is: how to make a quantitative comparison of the roles played by the two respective sets A and B in regulating or provoking the condition of interest?

Another very hard practical problem faced when attempting to use a classical statistical method in quantifying genes relevance, is that genes relevance index should take into account the interaction links among genes in the mechanisms which determine the biological condition of interest. This would imply the application of the statistical method to each possible subgroup of thousand of genes, which is often a procedure computationally too costly.

A completely different approach, based on a cooperative game in coalitional form where the players are genes, has been proposed in this thesis.

In my opinion, the novelty of the approach with respect to the classical statistical methods is essentially twofold. First, the class of cooperative games used, called the class of microarray games, provides the effective opportunity to describe the association between the global expression of each coalition of genes and a biological condition of interest and, as a consequence, to incorporate in the successive analysis all possible genes interaction ties related with the biological condition. For example, it is possible to describe the association between the over-expression or the under-expression properties of genes in each coalition and the tumor or the effect of a treatment in samples.

Even considering all possible subsets of genes, which means increasing a lot the level of complexity of the analysis, no strong assumptions on the expression probability distributions have been done. In fact, the characteristic function of a microarray game relays completely on the observed experimental gene expression matrix. The very relevant assumption in this context, is the definition of the causality relation (also called *sufficiency principle*) which incorporates the criterium used to establish whether the expression levels of genes in a coalition are associated or not with the biological condition of interest.

All the information on genes associations stored in the characteristic function of a microarray game can be successively exploited to quantitatively resume the role of each gene in each possible coalition by means of the application of solution concepts for cooperative games. The second novelty of the approach presented in this thesis is based on this idea of application of solution concepts to microarray games, and on the strong connection between game theory and the property driven approach commonly used for studying the properties of solution concepts. As I pointed out in the general introduction on cooperative game theory, the property driven characterization of solution concepts has abundantly been used in Game Theory, attempting to investigate the real extent of the theory and to contextualize its potential applications.

Usually, the interpretation of the results obtained by classical statistical procedures are strongly dependent from the theoretical model used for the analysis or from strong assumptions about the reference population from which the samples are collected. The property driven approach offers the possibility to overturn this view: only weak assumptions on the population are needed and what is strongly outlined *a priori* are the boundaries for a plausible interpretations of the results. In the game theoretical approach, the result is the outcome of a solution concept applied to a microarray game built on a gene expression matrix. Its interpretation is contextualized ex-ante by means of sound basic properties, that have to be satisfied by a numerical representation of the role played by each gene in associating the expressions of coalitions with the condition of interest. This view is particular valuable in the genomic field, which is still a relatively young research topic, and the evidences to support strong hypothesis on the reference populations or the application of sophisticated mathematical models are still far from to be clear. These considerations are, in my opinion, effectively resumed by the following sentence, in Stöltzner (2004): if a field is still provisional in its basic concepts, and experience with models is fragmentary, the property driven method is able to act as a controlling instance and steering device for further exploration.

On the other hand, it is not possible to neglect the fact that gene expression is a stochastic, or "noisy", process (Elowitz (2002), Swain (2002)). Besides the biological noise, microarrays data, as any other experimental process, are subject to random experimental noise. As a consequence, since a microarray game is inferred from gene expression data, a microarray game itself follows a stochastic law. Therefore, I felt the necessity to introduce microarray games in an alternative way, supported by inference arguments, with the objective to assess the effects of the random variability on the observed results of the game theoretical analysis.

Summing up, the class of microarray games and the methods for their analysis, constitute the core of this work. Their intrinsic simple structure was the main reason that convinced me to focus my efforts on their analysis. On the other hand, it is possible that some aspects of the real phenomenon that I was going to investigate were missed due to the same reason. To catch such aspects, one possibility could be to make the models a bit more sophisticated, and the preliminary study on new classes of gene expression based games is the present direction of my work and the conclusion of my dissertation.

1.3.1 A brief summary of the following chapters

Next sections 1.4 and 1.5 introduce some preliminaries on cooperative games and on microarray data analysis, respectively.

Chapter 2 is based on Moretti *et al.* (2004), where the class of *microarray games* has been introduced. Via a dichotomization technique applied to gene expression data, it is constructed a game whose characteristic function takes values on the interval $[0, 1]$. The objective of such a game is to stress the relevance ('sufficiency') of groups of genes in relation to a specific biological condition or response of interest (e.g. a disease of interest). It has been discussed the possibility of applying game-theoretical tools that can take into account the relationships which exist among genes, like the Shapley value. The highest Shapley values of the game should point to the most influential genes, so that it could be useful as a hint for pointing at the genes that mostly deserve further investigation. A property driven characterization of the Shapley value with a genetic interpretation is also provided in order to contextualize and justify the use of the Shapley value as relevance index for genes.

Chapter 3 is based on Moretti (2006). It has been presented a statistical framework aimed at estimating the accuracy of the observed genes relevance index and a procedure to test the null hypothesis of no differences in terms of relevance index for genes studied in samples regulated by different biological conditions. The first goal of Chapter 3 is to answer the question on how accurate are the relevance estimates provided by the Shapley value applied on games introduced in Chapter 2. That question is the prelude for the second subject of this chapter, i.e. comparing the relevance of genes under different biological conditions or responses.

Chapter 4 is still in a germinal form and contains many directions on which I am presently working. In Section 4.1, an alternative model based on minimum cost spanning tree representation of gene expression data has been introduced. One of the main characteristics of this model is the possibility to avoid the dichotomization technique required for microarray games introduced in Chapter

2. In Section 4.2, the connections between microarray games and the class prediction problem have been also presented. Finally, in Section 4.3, it has been introduced an overview of analysis performed on gene expression data of neuroblastoma samples that is still in progress and that I am doing using the game theoretical tools presented in the previous chapters.

Finally, note that all the algorithms presented in this dissertation and other procedures used in the analysis of expression data have been implemented using the statistical programming language R (R Development Core Team (2004)), and available on request.

1.4 Preliminary notations on cooperative games

Now, let us introduce some basic game theoretical notations. A *cooperative game with transferable utility* or TU-game, also known as *coalitional game with transferable payoff*, is a pair (N, v), where N denotes the finite set of *players* and $v : 2^N \to I\!R$ the *characteristic function*, with $v(\emptyset) = 0$. Often we identify a TU-game (N, v) with the corresponding characteristic function v. A group of players $T \subseteq N$ is called a *coalition* and $v(T)$ is called the *value* of this coalition. A TU-game (N, w) such that $w : 2^N \to [0, 1]$ is called a $[0, 1]$-*game*. We will denote the class of all $[0, 1]$-games as $\mathcal{W}$, with $\mathcal{W} \subset \mathcal{G}$, being $\mathcal{G}$ the class of all TU-games (N, v).

Let $\mathcal{C} \subseteq \mathcal{G}$ be a subclass of TU-games. Given a set of players N, we denote by $\mathcal{C}^N \subseteq \mathcal{G}$ the class of TU-games in $\mathcal{C}$ with N as set of players.

The *unanimity game* (N, u_R) based on the unanimity set $R \subseteq N$ is the game described by $u_R(T) = 1$ if $R \subseteq T$ and $u_R(T) = 0$, otherwise. Every TU-game (N, v) can be written as a linear combination of unanimity games in a unique way, i.e. $v = \sum_{S \subseteq N, S \neq \emptyset} \lambda_S(v) u_S$ (see for instance Owen (1995)). The coefficients $(\lambda_S(v))_{S \in 2^N \setminus \{\emptyset\}}$ are called *unanimity coefficients* or *dividends* of the game (N, v).

A TU-game (N, v) is *monotonic* if for all $S, T \subseteq N$, $S \subseteq T$ implies that $v(S) \leq v(T)$.

Let $i \in N$. For each $S \subseteq N \setminus \{i\}$, the quantity $m_i(v, S) = v(S \cup \{i\}) - v(S)$ is the marginal contribution of player i to coalition S. A TU-game (N, v) is *convex* if for all $i \in N$ and all $S, T \subseteq N \setminus \{i\}$, $S \subseteq T$ implies that

$$m_i(v, S) \leq m_i(v, T)). \tag{1.1}$$

An *allocation* $(x_i)_{i \in N}$ of a TU-game (N, v) is a vector in $I\!R^N$ describing the payoffs of the players, where player $i \in N$ receives x_i.

An *one-point solution* for a class $\mathcal{C}$ of TU-games is a function ψ that assigns a payoff vector $\psi(v)$ to every TU-game in the class, that is $\psi : \mathcal{C}^N \to I\!R^N$.

The most famous one-point solution in the theory of cooperative games with transferable utility is the *Shapley value*, introduced by Shapley (1953). To have a basic idea about the Shapley value, suppose that all the players are arranged in some order, all orderings being equally likely. The Shapley value ϕ_i of the game $(N, v) \in \mathcal{G}^N$, for each $i \in N$, is defined as the expected marginal contribution, over all orderings, of player i to the set of players who precede him. Since for each $S \subseteq N \setminus \{i\}$ there are precisely $\frac{(s-1)!(n-s)!}{n!}$ orderings in which players in S precede player i, than the Shapley value ϕ_i applied to game $(N, v) \in \mathcal{G}^N$ can be calculated by the general formula

$$\phi_i(v) = \sum_{S \subseteq N : i \in S} \frac{(s-1)!(n-s)!}{n!} m_i(v, S)) \tag{1.2}$$

for each $i \in N$, where $s = |S|$ and $n = |N|$ are the cardinality of coalitions S and N, respectively.

An alternative representation of the Shapley value can be given in terms of the unanimity coefficients $(\lambda_S(v))_{S \in 2^N \setminus \{\emptyset\}}$ of a game (N, v), that is:

$$\phi_i(v) = \sum_{S \subseteq N : i \in S} \frac{\lambda_S(v)}{s} \tag{1.3}$$

for each $i \in N$.

Another one-point solution for cooperative games with transferable utility is the *Banzhaf value*, introduced by Banzhaf (1965). The Banzhaf value $\beta_i(v)$ of a game $(N, v) \in \mathcal{G}^N$, is defined as follows

$$\beta_i(v) = \sum_{S \subseteq N : i \in S} \frac{1}{2^{n-1}} m_i(v, S), \tag{1.4}$$

for each $i \in N$.

A common characteristics of the Banzhaf value and of the Shapley value of a game (N, v) is that both one-point solutions belong to the class of allocations which can be obtained via the general formula

$$\epsilon_i(v) = \sum_{S \subseteq N : i \in S} p_i(S) m_i(v, S), \tag{1.5}$$

where $p(S)$, for each $S \in 2^N \setminus \{\emptyset\}$, is the probability that a player $i \in S$ joins the other players in $S \setminus \{i\}$ to form coalition S. So, $\epsilon_i(v)$ is the average marginal contribution of player $i \in N$ with respect to all the possible coalitions in which player i can enter. If $p(S)$ is assumed to be the same for each coalition $S \in 2^N \setminus \{\emptyset\}$, then $p(S) = \frac{1}{2^{n-1}}$, and the definition of the Banzhaf value by formula (1.4) is obtained. If $p(S)$ is assumed to be dependent from S, one choice could be assume that $p(S) = \frac{(s-1)!(n-s)!}{n!}$, and the definition of the Shapley value by formula (1.2) is obtained. Of course, other probability distributions on the set of all coalitions can be used in order to define different one-point solutions.

Finally, a particular set, possibly empty, of allocations of a TU-game (N, v) is the *core*, which is defined as follows:

$$core(v) = \{x \in I\!R^N | \sum_{i \in S} x_i \geq v(S) \ \forall S \in 2^N \setminus \{\emptyset\}; \sum_{i \in N} x_i = v(N)\}.$$

1.5 Preliminary notations on microarray data analysis

Let $G = \{1, 2, \ldots, n\}$ be a set of n genes, $S_R = \{1, 2, \ldots, r\}$ be a set of r reference samples, i.e. the set of cells from normal tissues and, finally, let $S_D = \{1, 2, \ldots, d\}$ be the set of d cells from tissues with a biological condition or response of interest (e.g. a disease).

The goal of a microarray experiment is to associate to each sample $j \in S_R \cup S_D$ an *expression profile* $(a_{ij})_{i \in G}$, i.e. $a_{ij} \in I\!R$ represents the *relative expression value* of the gene i in sample j with respect to the reference sample. 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 expression matrices $A^{S_R} = (A^j)_{j \in S_R}$ and $A^{S_D} = (A^j)_{j \in S_D}$, where the index here represents a column, i.e. a sample, where the column A^j is the expression profile on G of sample j. In summary, we will denote as a *microarray experimental situation* (MES) the tuple $E =< G, S_R, S_D, A^{S_R}, A^{S_D} >$.

As the first step of our analysis, we are interested in understanding whether genes in each sample in S_D are abnormally expressed with respect to the expression values showed in S_R according to a certain discriminative criterium. For example, we could refer to the set of abnormally expressed genes in a sample as the set of over (under) expressed genes in that sample, or the union of over expressed and under expressed genes in that same sample .

We need to introduce useful notation to deal with abnormally expressed genes. Note that gene $i \in G$ which results abnormally expressed on a sample $j \in S_D$ can be represented setting to 1 the value of a boolean variable b_{ij}. We call *abnormal expression profile* the vector $\mathbf{B}^j = (b_{ij})_{i \in G}$. A *discriminant method* can be expressed as a map m assigning to each expression profile from tumor samples a corresponding abnormal expression profile. Hence, all the information on the differences of gene expression of sample in S_D from the ones of sample in S_R can be represented via an *abnormal expression matrix* $\mathbf{B}^{E,m} \in \{0,1\}^{G \times S_D}$.

Since for our purposes the relevant information is contained in the abnormal expression matrix $\mathbf{B}^{E,m}$, in the sequel we identify the MES E and the discriminant method m with the matrix $\mathbf{B}^{E,m}$. Sometimes, unless otherwise clear from the context, we will also refer to a boolan matrix $\mathbf{B} \in \{0,1\}^{G \times S_D}$ as an abnormal expression matrix which has been calculated applying some discriminant method m to some MES E.

Example 1 Consider an MES $E =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ such that A^{S_R} is reported in the following table

	sample 1	sample 2	sample 3	sample 4
gene 1	0.4	0.2	0.3	0.6
gene 2	12	10	4	5
gene 3	8	13	20	9
gene 4	0	-0.5	1.4	1.1

and A^{S_D} is given in the following one

	sample 1	sample 2	sample 3
gene 1	0.9	0.4	0.7
gene 2	4.6	15	18
gene 3	7	21	12
gene 4	0.1	-0.4	1.6

Note that also negative values are possible. This is due to the fact that, usually, in literature the data set of a microarray experiment is presented in terms of the logarithm of the relative gene expression ratios, i.e, gene expression in the target sample / gene expression in the reference sample. Consequently, a positive number indicates a higher gene expression in the target sample than in the reference one, whereas a negative number indicates a lower expression in the target sample.

Now consider a very naive discriminant method m for the two classes 1 and 0, where 1 labels abnormally expressed genes and 0 labels normally genes and such that

$$(m(A^j, A^{S_R}))_i = \begin{cases} 1 & \text{if } A_i^j \geq max_{j \in \{1,\dots,|S_R|\}} A_{ij}^{S_R} \text{ or } A_i^j \leq min_{j \in \{1,\dots,|S_R|\}} A_{ij}^{S_R} \\ \\ 0 & \text{otherwise.} \end{cases}$$

Then the corresponding abnormal expression matrix is the following

$$\mathbf{B}^{E,m} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Chapter 2

The class of Microarray games and the relevance index for genes

2.1 Introduction

Aim of this chapter 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 microarray experiments, and taking into account the interaction level of each subgroup of genes.

In analyzing gene-gene relationships in microarray data, the main difficulty is the impossibility to obtain, trough pre-processing data analysis, a total elimination of the technical and biological bias. For this reason, in our approach we refer to the observed average interaction level 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 is the number of samples observed, the lower is the probability that chance could affect the inferences provided by the model. The basic idea of this model comes from the theory of cooperative games with transferable utility (TU-games). In particular we considered 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. (Owen (1995), Shapley and Shubik (1954), Banzhaf (1965)). We adopted the same formal language of TU-game for modelling the interaction among genes, considered as players, in relation to the pathogenesis of a genetic disease, e.g., a tumor. The game we considered origins from the comparison of two matrixes of gene expression data; one from tumor samples and the other from normal DNA (referent healthy subjects). We first used 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 1.5, we used the interval boundaries containing most data in the normal distribution of that gene as cut-offs for discrimination (Becquet et al.(2002)). We then introduced 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, has been given and it turned out to coincide with the Shapley value of the game considered.

In Section 2.2 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 2.3 an axiomatic characterization of the Shapley value is given by means of five properties suitable to genetic interpretation of this index. Section 2.4 concludes with some considerations on related works and future research.

2.2 Interaction among genes

In this phase of the analysis we assume that the abnormal expression profile $\mathbf{B}^j$, for each sample $j \in S_D$, is a sufficient conditions for the onset of the disease (or another biological condition or response of interest) in individuals from which samples in S_D are collected (*sufficiency principle for groups of genes*). Stated differently, a group of genes $A \subseteq G$ which are abnormally expressed in a sample of S_D (according to a discriminant method m applied to the reference expression matrix A^{S_R}) implies that an individual whose sample has at least all (possibly many more, due to biological and technical bias affecting the data set) genes in G abnormally expressed (again on the basis of m and A^{S_R}) should have the disease.

One could wonder why a microarray experiment can show -as it usually happens- different groups of abnormally expressed genes in different tumor samples.

We attempt to provide an answer to such a question with arguments coming from different directions.

One is dealing with biology: it is in fact likely that early stages of carcinogenesis involve metabolic paths which are controlled by different groups of genes.

Another reason is technical: a microarray experiment can be affected by many sources of noise (Parmigiani et al.(2003), Smith and Speed (2003)) and this unwanted variability can affect the measurement of expression values. Despite the reduction of variability in microarray experiments has been the objective of several works in the last few years, in practice the likely misclassification of some genes considered as abnormally expressed cannot be avoided, due to technical uncertainty.

The arbitrariness of methods used for the discriminant analysis should also be considered, i.e. the structure of the sufficient groups can be easily biased by a bad choice of the discriminant method.

The aim of this work is to give an answer to the following questions: how much relevant for the onset of a tumor are the genes which are abnormally expressed inside the sample S_D? Is it possible to provide a measure of the power of genes in determining the onset of the tumor in an individual, on the basis of the information collected via samples S_D and S_R and the discriminant method m used?

Consider for instance a MES $\bar{E} =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ and a discriminant method m such that the corresponding abnormal expression matrix is

$$\mathbf{B}^{\bar{E},m} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.1}$$

On the basis of matrix (2.1) it seems very reasonable to affirm that on the basis of the information collected A^{S_R} and the discriminant method used m, all the genes abnormally expressed have the same power in causing the tumor, assuming the principle of sufficiency for groups of abnormally expressed genes introduced before.

On the other hand, it could be reasonable to expect experimental situations where there are many abnormal expression profiles inside the sample S_D, like in the abnormal expression matrix of Example 1 and Example 2.

Example 2 Consider again the MES E of Example 1 and a more conservative discriminant method $\bar{m}$ such that

$$(\bar{m}(A^j, A^{S_R}))_i = \begin{cases} 1 & \text{if } A_i^j \leq p_i^{25\%} \text{ or } A_i^j \geq p_i^{75\%} \\ \\ 0 & \text{otherwise.} \end{cases}$$

The resulting abnormal expression matrix is the following

$$\mathbf{B}^{E,\bar{m}} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

where $p_i^{25\%}$ and $p_i^{75\%}$ are the 25^{th} and the 75^{th} percentiles of the expression distribution of gene i (i.e. the i^{th} row) in the reference expression matrix A^{S_R}, for each $i \in G$.

How to deal with these situations?

Given an MES $E =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ and a discriminant method m, first we determined the average number of individuals with the tumor due to the abnormal expression of a given group of genes. Of course we calculated such average values on the basis of the information provided by the pair $< E, m >$, that is, for each group $A \subseteq G$, we looked at the number of groups of abnormal expressed genes in $\mathbf{B}^{E,m}$ that are included in A. We formalize such a concept via the following definitions (in the following $\mathbf{B}^{E,m}(j)$ will be the column j, $j \in \{1, \ldots, |S_D|\}$, of the abnormal expression matrix $\mathbf{B}^{E,m}$).

Definition 1 *Let $v \in \{0,1\}^n$, $n \in \{1,2,\ldots\}$. We define the* support *of v denoted by $sp(v)$ the set*

$$sp(v) = \{i \in \{1, \ldots, n\} \mid v_i = 1\}.$$

Example 3 Consider the abnormal expression matrix $\mathbf{B}^{E,m}$ of Example 1. Then $sp(\mathbf{B}^{E,m}(1)) = \{1,3\}$, $sp(\mathbf{B}^{E,m}(2)) = \{2,3\}$ and $sp(\mathbf{B}^{E,m}(3)) = \{1,2,4\}$.

Definition 2 *Let $E =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ be an MES and let m be a discriminant method. We define the* average number of individuals with tumor determined by the genes in G, *for each $T \in 2^G \setminus \{\emptyset\}$ as the value*

$$v(T) = \frac{|\Theta(T)|}{|S_D|} \tag{2.2}$$

where $|\Theta(T)|$ is the cardinality of the set

$$\Theta(T) = \{k \in \{1, \dots, |S_D|\} \mid sp(\mathbf{B}^{E,m}(k)) \subseteq T,\ sp(\mathbf{B}^{E,m}(k)) \neq \emptyset\} \tag{2.3}$$

and $v(\emptyset) = 0$.

Now, the definition of the corresponding TU-game should be clear:

- the set of players is the set of genes G;
- the characteristic function is the average number of individuals with tumor determined by the genes T, for each $T \in 2^G \setminus \{\emptyset\}$.

More formally

Definition 3 *Let $E =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ be an MES and let m be a discriminant method. We define the corresponding* microarray game *as the TU-game (G, v), where v is defined as in Definition 2.*

Remark 1 *Condition $sp(\mathbf{B}^{E,m}(k)$ in relation (2.3) is due to practical considerations concerning the interpretation of the sufficiency principle for groups of genes on samples where genes do not show any abnormal expression properties. We are assuming that the contribution of such a sample in increasing the level of association between the abnormal expression of genes in S and the disease (or another condition of interest) is null, for each coalition $S \subseteq N$.*

The class of microarray games will be denoted with the symbol $\mathcal{M}$. Let $E =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ be an MES and let m be a discriminant method. According to equality (2.2), an equivalent way to calculate the corresponding microarray game v is as a sum of unanimity games as follows

$$v = \frac{1}{|S_D|} \sum_{j \in \{1, \dots, |S_D|\}} u_{sp(\mathbf{B}^{E,m}(j))}, \tag{2.4}$$

where $u_{sp(\mathbf{B}^{E,m}(j))}$ is the unanimity game on $sp(\mathbf{B}^{E,m}(j)) \subseteq G$, for each $j \in \{1,\ldots,|S_D|\}$.

Alternatively, it is possible to rewrite equation (2.4) in terms of the unanimity coefficients of a microarray game v. Let $\bar{\lambda}_S \in \{0,1,2,\ldots\}$ be the number of occurrences of the coalition S as support in the abnormal expression matrix $\mathbf{B}^{E,m}$. In formula

$$v = \frac{1}{|S_D|} \sum_{S\subseteq N: S\neq\emptyset} \bar{\lambda}_S u_S. \tag{2.5}$$

where $\bar{\lambda}_S = |\{k \in \{1,\ldots,|S_D|\} \text{ s.t. } sp(\mathbf{B}^{E,m}(k)) = S,\ sp(\mathbf{B}^{E,m}(k)) = \emptyset\}|$.

Example 4 Consider again the abnormal expression matrix $\mathbf{B}^{E,m}$ of Example 1 By equation 2.4 the corresponding microarray game $(\{1,2,3,4\},v)$ is such that

$$v = \frac{1}{3}\big(u_{\{1,3\}} + u_{\{2,3\}} + u_{\{1,2,4\}}\big).$$

It follows that $v(\emptyset) = v(\{1\}) = v(\{2\}) = v(\{3\}) = v(\{4\}) = v(\{1,2\}) = v(\{1,4\}) = v(\{2,4\}) = v(\{3,4\}) = 0$; $v(\{1,3\}) = v(\{2,3\}) = v(\{1,3,4\}) = v(\{2,3,4\}) = v(\{1,2,4\}) = \frac{1}{3}$; $v(\{1,2,3\}) = \frac{2}{3}$, $v(\{1,2,3,4\}) = 1$.

It is easy to check that microarray games are $[0,1]$-games.

At this point, the major fundamental question addressed by our work can be formulated in the following terms: is it possible to employ the standard theory of TU-games to measure the expected relevance of each gene in determining the onset of tumor on the basis of the microarray experimental situation and the discriminant method used?

For instance, we can calculate the Shapley value of a microarray game. In the last fifty years, many studies have addressed the goal of evaluating the power of players (e.g. members of councils, voters in an electoral systems, parties in a parliament etc.) (see for instance Owen (1995)), which are TU-games whose characteristic function can only assume values 1 (for winning coalitions, i.e. coalitions which are able to force the endorsement of a motion) or 0 (for loosing coalitions). In such contexts, the idea was to evaluate the amount of power of players according to the role covered by each of them in supporting the goal of each possible coalition. There, the Shapley value (Shapley (1953), Shapley and Shubik (1954)), as well as many other solutions for TU-games, have been interpreted as power index for players (Shapley and Shubik (1954), Banzhaf (1965)).

On the other hand, even if the Shapley value has been proved to be very meaningful in political applications, it cannot be taken for granted the same significance in the microarray context.

The next examples show the behavior of the Shapley value on some particular instances of microarray.

Example 5 The Shapley value of the microarray game in Example 4 is $(\frac{5}{18}, \frac{5}{18}, \frac{1}{3}, \frac{1}{9})$. This means that on the basis of the corresponding MES E and the discriminant method m the Shapley value of the microarray game states that the most important attribute in determining the tumor onset - on the average - is gene 3, followed by genes 1 and 2 with the same score and gene 4.

Example 6 The Shapley value of the microarray game corresponding to the abnormal expression matrix in Example 2 is $(\frac{2}{9}, \frac{1}{3}, \frac{2}{9}, \frac{2}{9})$. On the basis of the considerations detailed in Example 5, we obtain that the most important gene in determining the tumor onset, on the average, is gene 2, followed by gene 1, 3 and 4 with the same score.

Example 7 Consider again the abnormal expression matrix (2.1). The Shapley value of the corresponding microarray game is $(\frac{1}{2}, 0, \frac{1}{2}, 0)$.

Example 8 We introduce here a preliminary application of our model on a real MES $E_c =< G, S_D, S_R, A^{S_D}, A^{S_R} >$ where A^{S_D} and A^{S_R} represent the tumor/normal data set (freely obtainable on the web site[1]) containing expression levels of a set G of 2000 genes measured using *Affymatrix oligonucleotide* microarrays for a set S_D of 40 tumor samples and a set S_R of 22 normal samples of colon tissues. After the preprocessing stage performed by the Bioconductor specific software for microarray analysis (Gentleman et al.(2004)), we applied the discriminant method m introduced in Example 1 in order to provide the abnormal expression matrix $\mathbf{B}^{E_c,m}$, which finally produces the corresponding microarray game (G, v_c).

In the following Table, the first ten genes with highest Shapley value [2] on the microarray game (G, v_c) have been indicated.

[1] http://microarray.princeton.edu/oncology/affydata/index.html

[2] We computed the Shapley value of the microarray game (G, v_c) by means of the procedure suggested by equation (1.3), implemented in the programming language R (R Development Core Team (2004)). Also the discriminant methods and other procedures for the management of data sets used in this application have been implemented using the language and environment R.

Gene Number	Gene Name	Shapley $\times(10^{-3})$
Z50753	H.sapiens mRNA for GCAP-II/ uroguanylin precursor	3.83
H17434	NUCLEOLIN (HUMAN)	3.56
H06524	GELSOLIN PRECURSOR, PLASMA (HUMAN)	3.34
H72234	DNA-(APURINIC OR APYRIMIDINIC SITE) LYASE (HUMAN)	3.33
M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.	3.23
U06698	Human neuronal kinesin heavy chain mRNA, complete cds.	3.21
H61410	PLATELET GLYCOPROTEIN IV (H. sapiens)	3.14
R39209	HUMAN IMMUNODEFICIENCY VIRUS TYPE I ENHANCER-BINDING PROTEIN 2 (H. sapiens)	3.13
M58050	Human membrane cofactor protein (MCP) mRNA, complete cds.	3.09
H08393	COLLAGEN ALPHA 2(XI) CHAIN (H. sapiens)	3.01

The complete distribution of the Shapley value on the genes is depicted in Figure 1.

Some of the genes selected were previously observed to be associated with the colon cancer (Fujarewick and Wiench (2003)): the vasoactive intestinal peptide (VIP), has been suggested to promote the growth and proliferation of tumor cells; the membrane cofactor protein (MCP) represents a possible mechanism of the ability of the tumor to evade destruction by the immune system (tumor *escape*); gelsolin is protein which acts as both a regulator and an effector of *apoptosis*, i.e. the mechanism responsible for the physiological deletion of cells. DNA-apurinic or apyrimidinic site lyase protein plays an important role in DNA repair and in resistance of cancer cells to radiotherapy (Moler et al.(2000)).

For comparison, we computed on the corresponding microarray game also another very famous solution: the Banzhaf value (Banzhaf (1965)). The common genes in the top ten of Shapley value and Banzhaf value have been indicated in bold.

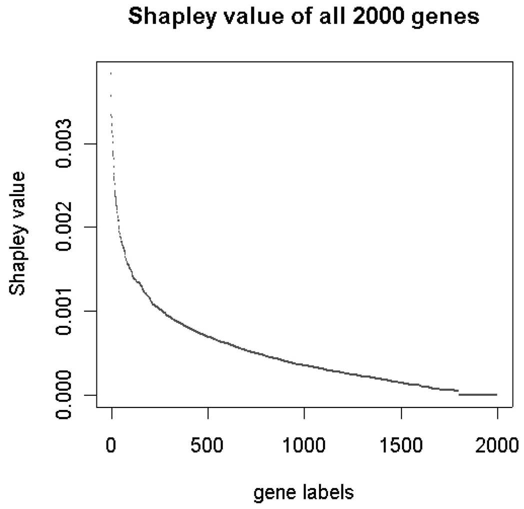

Figure 2.1: Shapley value of genes in a real MES.

The previous examples show a reasonable behavior of the Shapley value in measuring the relevance of each gene in determining the tumor onset. Furthermore, to support the idea that the Shapley value is a good estimator of the relevance of each gene, in the next section we provide a new axiomatic characterization of this solution satisfying properties which have a nice interpretation in the gene scenario.

We end this section with some properties of microarray games.

Proposition 1 *Let $< G, S_D, S_R, A^{S_D}, A^{S_R} >$ and m be an MES and a discriminant method, respectively, and let v be the corresponding microarray game in $\mathcal{M}^G$. Then v is a super-additive, monotone and convex TU-game.*

Proof Super-additivity and monotonicity follow directly from the fact that unanimity games are super-additive and monotone and by equation (2.5) microarray games are positive linear combination of unanimity games.

It is easy to check that Convexity follows analogously from convexity of unanimity games. First, note that for an unanimity game u_S, $S \subseteq N$, the marginal contribution $u_S(T) - u_S(T \setminus \{i\})$ can be 0 or 1, for each $i \in N$ and each $T \in 2^N \setminus \{\emptyset\}$ such that $i \in T$. If $i \in N \setminus S$, then $u_S(T) - u_S(T \setminus \{i\}) = 0$. On the other hand, by definition of unanimity game, if $i \in S$ then the following statement:

$$u_S(T) - u_S(T \setminus \{i\}) = 1 \Rightarrow u_S(R) - u_S(R \setminus \{i\}) = 1$$

holds for each R, T such that $T \subseteq R \subseteq N$ and $i \in T$. Hence, it remains to prove equation (1.1) on game u_S.

Again, convexity of v follows immediately by equation 2.5, since for each $T \subseteq R \subseteq N \setminus \{i\}$ and each $i \in N$

$$\begin{aligned}
&v(T \cup \{i\}) - v(T) = \\
&\tfrac{1}{|S_D|} \textstyle\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_V u_V(T \cup \{i\}) - \tfrac{1}{|S_D|} \sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_V u_V(T) = \\
&\tfrac{1}{|S_D|} \textstyle\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_V \big(u_V(T \cup \{i\}) - u_V(T)\big) \leq \\
&\tfrac{1}{|S_D|} \textstyle\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_V \big(u_V(R \cup \{i\}) - u_V(R)\big) = \\
&v(R \cup \{i\}) - v(R),
\end{aligned}$$

where $\bar{\lambda}_S = |\{k \in \{1, \ldots, |S_D|\} \text{ s.t. } sp(\mathbf{B}^{E,m}(k)) = S\}|$. ■

2.3 An axiomatic characterization of the Shapley value with genetic interpretation

In order to characterize the Shapley value by means of properties with genetic interpretation, the definition of partnership of genes takes a basic role.

Definition 4 *Let $v \in \mathcal{M}^N$. A coalition $S \in 2^N \setminus \{\emptyset\}$ such that for each $T \subsetneq S$ and each $R \subseteq N \setminus S$*

$$v(R \cup T) = v(R)$$

is a partnership of genes *in the microarray game v.*

The worth $v(S)$ of a partnership of genes S represents the maximum average number of onsets of the tumor that genes in the partnership are able to determine in the population, whatever the interaction of its genes with the others outside the partnership may be. Note that the concept of partnership in TU-games has been introduced in Kalai and Samet (1988) in a general context not involving genes.

Remark 2 Let $v \in \mathcal{M}^N$ and let $S \in 2^N \setminus \{\emptyset\}$ be a partnership in v. Then it is trivial to prove that each $T \subseteq S$ is a partnership itself.

Let $v \in \mathcal{M}^N$. A *maximal partnership* $S \in 2^N \setminus \{\emptyset\}$ in v is a maximal subset of N with the property to be a partnership in v.

We denote by $\mathcal{P}(v)$ the set of all the maximal partnerships in v. Note that, by Definition 4, it immediately follows that all one player coalitions are partnerships in v. One easily obtains that the collection of maximal partnerships in v forms a partition of N. For instance, in the microarray game $(\{1, 2, 3, 4\}, v)$ of Example 4 $\mathcal{P}(v) = \{\{1\}, \{2\}, \{3\}, \{4\}\}$ and coincides with set of all the partnerships in v; whereas in the microarray game of Example 7 $\mathcal{P}(v) = \{\{1, 3\}, \{2, 4\}\}$.

Some interesting properties for solutions of microarray games, which are related to the concept of partnership of genes, are the following.

Let $F : \mathcal{M}^N \to I\!\!R^N$ be a solution on the class of microarray games.

Property 1 Let $v \in \mathcal{M}^N$. The solution F has the *Partnership Rationality* (PR) property, if

$$\sum_{i \in S} F_i(v) \geq v(S)$$

for each $S \in 2^N \setminus \{\emptyset\}$ such that S is a partnership of genes in the game v.

The PR properties determines a lower bound of the power of a partnership, i.e. the total relevance of a partnership of genes in determining the onset of the tumor in the individuals should not be lower than the average number of cases of tumor enforced by the partnership itself.

Property 2 Let $v \in \mathcal{M}^N$. The solution F has the *Partnership Feasibility* (PF) property, if

$$\sum_{i \in S} F_i(v) \leq v(N)$$

for each $S \in 2^N \setminus \{\emptyset\}$ such that S is a partnership of genes in the game v.

On the contrary of PR, the PF properties determines an upper bound of the power of a partnerships, i.e. the total relevance of a partnership of genes in determining the tumor onset in the individuals should not be greater than the average number of cases of tumor enforced by the grand coalition, which is always 1.

Property 3 Let $v \in \mathcal{M}^N$. The solution F has the *Partnership Monotonicity* (PM) property, if

$$F_i(v) \geq F_j(v)$$

for each $i \in S$ and each $j \in T$, where $S, T \in 2^N \setminus \{\emptyset\}$ are partnerships of genes in v such that $S \cap T = \emptyset$, $v(S) = v(T)$, $v(S \cup T) = v(N)$, $|S| \leq |T|$.

The PM property is very intuitive: consider two disjoint partnerships of genes enforcing the same average number of cases of tumor in the set of samples. If the genes outside the union of those two partnerships are irrelevant - that is they do not contribute in increasing the average number of tumors - then genes in the smaller partnership should receive a higher relevance index than genes in the bigger one.

The next two properties do not involve the concept of partnership of genes.

Property 4 [3] Let $v_1, \ldots, v_r \in \mathcal{M}^N$. The solution F has the *Equal Splitting* (ES) property, if

$$F(\frac{\sum_{i=1}^r v_i}{r}) = \frac{\sum_{i=1}^r F(v_i)}{r}.$$

Remark 3 Note that $\frac{\sum_{i=1}^r v_i}{r} \in \mathcal{M}^N$.

Moreover, if $\mathbf{B}_1^{E_i,m}, \ldots, \mathbf{B}_r^{E_r,m}$ are r abnormal expression matrix with the same number of columns and $v_1, \ldots, v_r \in \mathcal{M}^N$ are the corresponding microarray games, then $\frac{\sum_{i=1}^r v_i}{r}$ coincides with the microarray game corresponding to the abnormal expression matrix obtained juxtaposing the matrices $\mathbf{B}_i^{E_i,m}, \ldots, \mathbf{B}_r^{E_r,m}$.

To prove these facts a cumbersome notation is needed. So, we prove it in detail only for $r = 2$.

Let $k, l, p \in \mathbb{N}$ and let $\oplus : I\!R^{k \times l} \times I\!R^{k \times p} \rightarrow I\!R^{k \times (l+p)}$ be a matrix operator such that if $A \in I\!R^{k \times l}$ and $B \in I\!R^{k \times p}$, then $A \oplus B = C$ is such that $C^i = A^i$ for each $i \in \{1, \ldots, l\}$ and $C^{j+l} = B^j$ for each $j \in \{1, \ldots, n\}$.

Let $\mathbf{A}^{E_A,m} \in \{0,1\}^{k \times l}$ and $\mathbf{B}^{E_B,m} \in \{0,1\}^{k \times p}$ be two abnormal expression matrix arising from the application of a given discriminant method m on two different microarray experimental situations on the same set of genes G and the same set of reference samples S_R, with $|G| = k$, and where l and p are the cardinality of the respective sets of tumor samples. Consider $v_A, v_B \in \mathcal{M}^N$ the two

[3] Assuming the continuity of F, it can be proved, using functional equation theory, that the ES property is equivalent to the simpler property of requiring that F satisfies $F(\frac{v+w}{2}) = \frac{F(v)+F(w)}{2}$ for each pair $v, w \in \mathcal{M}^N$.

corresponding microarray games, respectively obtained from $\mathbf{A}^{E_A,m}$ and $\mathbf{B}^{E_B,m}$ by Definition 3. It is easy to check that the game $\frac{v_A+v_B}{2}$ is the microarray game corresponding to the abnormal expression matrix $\bigoplus_{i=1}^{p} \mathbf{A}^{E_A,m} \oplus \bigoplus_{i=1}^{l} \mathbf{B}^{E_B,m}$. Therefore, if $l = p$, the microarray game $\frac{v_A+v_B}{2}$ corresponds to $\mathbf{A}^{E_A,m} \oplus \mathbf{B}^{E_B,m}$.

For $r > 2$ similar arguments hold too.

The ES property underlies a principle of equivalence of reliability levels for microarray games arising from equal splitting of the same MES. Let $< G, S_D, S_R, A^{S_D}, A^{S_R} >$ be an MES and let $S_{D_1}, \dots, S_{D_m}$ form a partition of the set of samples S_D such that $|S_{D_1}| = |S_{D_2}| = \dots = |S_{D_m}|$. If the ES property holds, then the relevance index computed on the microarray game corresponding to $< G, S_D, S_R, A, A^{S_D}, A^{S_R} >$ equals the average of the relevant indices computed on the microarray games arising from the microarray experimental situations $< G, S_{D_1}, S_R, A^{S_{D_1}}, A^{S_R} >, \dots, < G, S_{D_m}, S_R, A^{S_{D_m}}, A^{S_R} >$, respectively; differently stated, the relevance index is independent from the equal splitting partition $\{S_{D_1}, \dots, S_{D_m}\}$ chosen.

The last property involves the definition of *null player* of a game (N, v), that is a player $i \in N$ such that $v(S \cup i) = v(S)$ for each $S \subseteq N \setminus \{i\}$.

Property 5 Let $v, w \in \mathcal{M}^N$. The solution F has the *Null Player* (NP) property, if for each null player $i \in N$

$$F_i(v) = 0.$$

The interpretation of the NP property is straightforward: if a player does not contribute anything to each coalition $S \in 2^N$ then he gets null relevance.

Remark 4 It is well known in literature that the Shapley value satisfies the NP property on each class of TU-games $\mathcal{C}^N \subseteq \mathcal{G}^N$. The ES property directly follows from Remark 3 together with additivity and homogeneity of the Shapley value ϕ on $\mathcal{G}^N$, that is $\phi(\alpha v + \beta w) = \alpha\phi(v) + \beta\phi(w)$ for each $v, w \in \mathcal{G}^N$.

Lemma 1 *Let $v \in \mathcal{M}^N$ and let $S \in 2^N \setminus \{\emptyset\}$ be a maximal partnership in v. Then the Shapley value attributes the same relevance index to players in S.*

Proof Let $\phi(v)$ be the Shapley value on the game v. For each $U \subseteq N$ such that $i \in U$ the marginal contribution of player $i \in S$ is the following

$$\begin{aligned}
&v(U) - v(U \setminus \{i\}) \\
&= v([U \cap S] \cup [U \setminus S]) - v([(U \cap S) \setminus \{i\}] \cup [U \setminus S]) \\
&= \begin{cases} v(U \setminus S) - v(U \setminus S) & \text{if } U \cap S \neq S \\ \\ v(U) - v(U \setminus S) & \text{if } U \cap S = S \end{cases} \\
&= \begin{cases} 0 & \text{if } U \cap S \neq S \\ \\ v(U) - v(U \setminus S) & \text{if } U \cap S = S, \end{cases}
\end{aligned}$$

where the second equality follows by Definition 4 on partnership S.

Then, the marginal contribution of each player $i \in S$ to coalition U is different from zero only if S is a subset of U, which means that by equation (1.2) the Shapley value of player i is

$$\begin{aligned}
&\textstyle\sum_{U \subseteq N : i \in U} \frac{(u-1)!(n-u)!}{n!} (v(U) - v(U \setminus \{i\})) \\
&= \textstyle\sum_{U \subseteq N : S \subseteq U} \frac{(u-1)!(n-u)!}{n!} (v(U) - v(U \setminus S)),
\end{aligned}$$

for each $i \in S$, proving that the Shapley value is the same for each player $i \in S$. ■

Lemma 2 *Let $v \in \mathcal{M}^N$ and let $S \in 2^N \setminus \{\emptyset\}$ be a maximal partnership in v. Then*

$$v(U) = 0$$

for each $U \subsetneq S$.

Proof Suppose on the contrary $v(U) \neq 0$. Then, by Definition 3, $v(R \cup U) > v(R)$ for each $R \subseteq N \setminus U$, which yields a contradiction by Definition 4. ■

Proposition 2 *The Shapley value satisfies the properties PM, PR, PF.*

Proof Let $v \in \mathcal{M}^N$ and let $\phi(v)$ be the Shapley value on the game v.

i) Let S and T two disjoint partnerships such that $v(S) = v(T)$ and $v(S \cup T) = v(N)$.

   If S and T are subsets of the same maximal partnership, then their Shapley index is the same by Lemma 1, and PM is directly satisfied.

   If S and T are subsets of two different maximal partnerships U and V, respectively, then $S = U$ and $T = V$. In fact, suppose on the contrary that $S \subset U$ or $T \subset V$. By condition $v(S) = v(T)$ and Lemma 2 we have $v(S) = v(T) = 0$, and then, by definition 3, it follows $v(S \cup T) \neq v(N)$, which yields a contradiction.

   We still have to prove PM when S and T are two maximal partnerships. By condition $v(S \cup T) = v(N)$ and Definition 3, it turns out that $v(U) = v(U \cap (S \cup T))$ for each $U \subseteq N$. By Lemma 2 and Definition 4 $v(R) = 0$ for each $R \subseteq S \cup T$, with $S, T \nsubseteq R$. Hence, it is possible to write the game v in terms of unanimity games in the following way

$$v = \frac{1}{|S_D|}\big(v(S)(u_T + u_S) + v(N)u_{S \cup T}\big),$$

   where S_D is the number of samples in the corresponding MES. Finally, by equation 1.3, $\phi_i = \frac{v(S)}{|S|} + \frac{v(N)}{|S|+|T|}$ for each $i \in S$ and $\phi_j = \frac{v(S)}{|T|} + \frac{v(N)}{|S|+|T|}$ for each $j \in T$, which concludes the proof of the PM property of the Shapley value.

ii) The convexity of microarray games by Proposition 1 guarantees that the Shapley value $\phi(v)$ is in the core of the microarray game v. The PR property follows directly from intermediate rationality of core allocations.

iii) For each $S \in 2^N \setminus \{\emptyset\}$ such that S is a maximal partnership in v, by monotonicity of v and the fact that $\phi(v)$ is in the core of the microarray game v we have $\sum_{i \in S} \phi_i(v) \geq v(S) \geq 0$. On the other hand, by efficiency of the Shapley value, $\sum_{i \in N} \phi_i(v) = v(N)$ and then $\sum_{i \in S} \phi_i(v) \leq v(N)$, which proves that the Shapley value satisfies the PF property.

■

Theorem 1 *Let be given a finite set N. The Shapley value on the class $\mathcal{M}^N$ of microarray games is the unique relevance index which satisfies the properties PR, PF, PM, ES and NP.*

Proof We already know by Proposition 2 and Remark 4 that the Shapley value satisfies the five properties PR, PF, PM, ES and NP. To prove the uniqueness consider a map $\psi : \mathcal{M}^N \to I\!R^N$ satisfying PR, PF, PM, ES and NP.

Consider the unanimity game $(N, u_S) \in \mathcal{M}^N$, where $S \in 2^N \setminus \{\emptyset\}$. First note that players $j \in N \setminus S$ are null players. Then by NP property, $\psi_j(u_S) = 0$ for each $j \in N \setminus S$.

Moreover, it is easy to see that S is a maximal partnership in u_S. Then by Lemma 2, for each pair of nonempty sets $U, W \subseteq S$ such that $U \cap W = \emptyset$ and $U \cup W = S$, $u_S(U) = u_S(W) = 0$ and $u_S(U \cup W) = u_S(S) = u_S(N)$. Since PM property holds for ψ, then $\psi_i(u_S) = \psi_j(u_S)$ for each $i, j \in S$.

It follows that $\sum_{i \in S} \psi_i(u_S) = |S|\psi_k(u_S)$, with $k \in S$. By PR $|S|\psi_k(u_S) \geq 1$ and, by PF $|S|\psi_k(u_S) \leq 1$. Hence, $\psi_k(u_S) = \frac{1}{|S|}$ for each $k \in S$ and $\psi_k(u_S) = 0$ for each $k \in N \setminus S$.

Finally, we have

$$\begin{aligned} &\psi(v) = \psi\Big(\tfrac{\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_S u_S}{\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_S}\Big) \\ &= \tfrac{1}{\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_S} \sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_S \psi(u_S) \\ &= \tfrac{1}{|S_D|} \sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_S \psi(u_S) = \tfrac{1}{|S_D|} \sum_{S \subseteq N : S \neq \emptyset} \tfrac{\bar{\lambda}_S}{|S|}, \end{aligned} \tag{2.6}$$

where $\bar{\lambda}_S = |\{k \in \{1, \ldots, |S_D|\} \ : \ sp(\mathbf{B}^{E,m}(k)) = S\}|$, S_D is the set of samples of an MES corresponding to v (note that $\sum_{S \subseteq N : S \neq \emptyset} \bar{\lambda}_S = |S_D|$), the first equality follows by equation (2.5) and the second one by the ES property,.

According to equation (1.3), it has been proved that $\psi(v) = \phi(v)$, where $\phi(v)$ is precisely the Shapley value on the microarray game v. ■

2.4 Discussion

In this chapter we introduced an application of cooperative TU-games to gene expression analysis related with disease onset. An axiomatic characterization of the Shapley value aimed at identifying a relevance index for genes has been also presented.

As far as we know, cooperative game theory has been previously used in gene analysis in a recent work by Kaufman et al.(2004) as an application of the Multi-perturbation Shapley value Analysis (MSA) (Keinan et al.(2004)). The aim of that work was to identify the importance in terms of causal responsibility of some genes in performing a certain function in yeast cells. In their approach, Kaufman et al.(2004) evaluate the worth of each coalition as a measure of the biological system's performance for a certain function (e.g. the ability of the system to survive the UV irradiation). In order to obtain such a worth for each coalition, they carried out a series of experiments where genes of each different subset of n genes were perturbed concomitantly; on each experiment the performance score was also measured and the score assigned to the corresponding subset of perturbed genes, finally obtaining a TU-game. For 2^n experiments were needed to obtain a TU-game, implying the impossibility to deal with the complete structure of the game, both for practical and computational reasons, authors suggested two complementary approaches: a) the use of mathematical predictors on the available data set to predict the missing performance scores (Doudoit and Fridlyand (2003), Golub et al.(1999)); b) limiting the focus to one and two dimensional interactions (Grabish and Roubens (1999), Keinan et al.(2004)).

In our application setting, where samples of tumoral individuals are involved, of course we cannot imagine to perform such perturbation experiments. Moreover, from the computational point of view, the procedure to obtain the Shapley value of a microarray game is very simple to be implemented. On the other hand, the interpretation of the Shapley value as a measure of the functional causal contribution of genes in a biological system, as provided by Kaufman et al.(2004) seems to corroborate our interpretation of the Shapley value as indicator of the relevance of genes in tumor onset.

Finally, note that an axiomatic characterization of the Shapley value with the axioms PR, PF, PM, NP, together with the additivity property (see for instance Shapley (1953)) holds on the more general class of TU-games which are a positive linear combination of unanimity games.

Chapter 3

Statistical analysis of the Shapley value for microarray games

3.1 Introduction

In Chapter 2, a game theoretical approach, based on a cooperative game in coalitional form with the set of genes as set of players, has been used to describe the strength of each subgroup (coalition) of genes in provoking a condition of interests and, as a consequence, to incorporate in the successive analysis all possible genes interaction links related with the condition. On the class of *microarray games*, an operative definition of relevance index for genes has been provided in terms of the well known Shapley value (Shapley (1954)) and the biological justification of its use has been circumstantiated via a new axiomatic characterization. Since gene expression is a stochastic, or “noisy”, process (Elowitz (2002), Swain (2002)) and a microarray game is defined on a gene expression data-set, a microarray game itself follows a stochastic law. For this reason, given an expression data-set of genes under a condition of interest, in Chapter 2 the estimates of genes relevance in provoking the condition have been attained by the Shapley value of the corresponding microarray game, which is defined as the average game across all the observed single sample based games.

The first goal of this chapter is to answer the question on how accurate are the relevance estimates provided in Chapter 2. That question is the prelude for the second subject of this work, i.e. comparing the relevance of genes under different biological conditions or responses, for instance two different sub-types of tumors, or two different treatments etc. In practice, we present an algorithm to perform statistical inference based on the sampling distributions of the sample statistic of microarray games and the corresponding statistic of Shapley values.

Section 3.3 describes how to estimate, from the information provided by a microarray experiment, the average game in the population of cells/samples under the same biological condition. Section 3.4 introduce the Shapley value distribution on the population of cells under the same biological condition, and shows that a good estimate of the average Shapley value in the population of cells is the Shapley value of a microarray game.

Section 3.3 and 3.4 together introduce the statistical framework to set up the bootstrap based algorithm presented in Section 3.5. The basic idea of Bootstrap (Efron (1979); see also Efron and Gong (1983) Efron and Tibshirani (1993)) is to use re-sample techniques to collect information about the shape, center, and spread of the sampling distribution of the statistic of interest. This idea is particularly valuable when it is not possible to assume a given model describing the gene expression distributions in the population and, consequently, it is not possible to calculate the parameter of the corresponding sampling distribution. This is the case of many microarray experiments where gene expression distributions present high heterogeneity (see for example Grant *et al.* (2002) concerning the Golub *et al.* (1999) leukemia data set). The problem is even more complex dealing with transformations of the gene expression distributions, as in the present study, where the statistics of microarray games must be considered. The problem of simultaneous comparison of thousands of null hypothesis is also tackled in Section 3.5.

Section 3.6 is dedicated to the application of the bootstrap based method presented in Section 3.5 to the analysis of the well studied 38 leukemia samples data-set published by Golub *et al.* (1999). Section 3.7 concludes the work with some remarks on genes found significant from the application described in Section 3.6.

3.2 Preliminary notations

To help the reader in following the argumentations of this chapter, we collect here the required basic definitions introduced in Chapter 2. Let $v \in \{0,1\}^n$, $n \in \{1,2,\ldots\}$. We define the *support* of v denoted by $sp(v)$ the set

$$sp(v) = \{i \in \{1,\ldots,n\} \mid v_i = 1\}.$$

Let $\mathbf{B} \in \{0,1\}^{n\times k}$, $n,k \in \{1,2,\ldots\}$, be a boolean matrix. We define the *microarray game* corresponding to $\mathbf{B}$ as the TU-game $(N,\bar{v})$ such that $N = \{1,\ldots,n\}$ and $\bar{v} : 2^N \to I\!R_+$ is such that for each $T \in 2^N \setminus \{\emptyset\}$, $\bar{v}(T)$ is the number of occurrences of the coalition T as a superset of the supports in the abnormal expression matrix $\mathbf{B}$, in formula

$$\bar{v}(T) = \frac{|\Theta(T)|}{k} \tag{3.1}$$

where $|\Theta(T)|$ is the cardinality of the set

$$\Theta(T) = \{j \in \{1,\ldots,k\} \mid sp(\mathbf{B}_j) \subseteq T,\ sp(\mathbf{B}_j) = \emptyset\}$$

and $\bar{v}(\emptyset) = 0$. Equivalently, the game $(N,\bar{v})$ can be represented via the relation

$$\bar{v}(S) = \sum_{j=1,\ldots,k} \frac{u_{sp(\mathbf{B}_j)}(S)}{k} \tag{3.2}$$

for each $S \in 2^N \setminus \emptyset$, where $(N, u_{sp(\mathbf{B}_j)})$ is the unanimity game on the set $sp(\mathbf{B}_j)$.

The class of microarray games will be denoted with the symbol $\mathcal{M}$.

Example 9 Consider the boolean matrix $\mathbf{B} \in \{0,1\}^{4\times 3}$ such that

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then $sp(\mathbf{B}(1)) = \{1,3\}$, $sp(\mathbf{B}(2)) = \{2,3\}$ and $sp(\mathbf{B}(3)) = \{1,2,4\}$. By equation 3.2 the corresponding microarray game $(\{1,2,3,4\}, v)$ is such that

$$v = \frac{1}{3}\big(u_{\{1,3\}} + u_{\{2,3\}} + u_{\{1,2,4\}}\big).$$

It follows that $v(\emptyset) = v(\{1\}) = v(\{2\}) = v(\{3\}) = v(\{4\}) = v(\{1,4\}) = v(\{2,4\}) = v(\{1,2\}) = v(\{3,4\}) = 0$; $v(\{1,3\}) = v(\{2,3\}) = v(\{1,3,4\}) = v(\{2,3,4\}) = v(\{1,2,4\}) = \frac{1}{3}$; $v(\{1,2,3\}) = \frac{2}{3}$, $v(\{1,2,3,4\}) = 1$.

The Shapley value of the microarray game $(\{1,2,3,4\}, v)$ is $(\frac{5}{18}, \frac{5}{18}, \frac{1}{3}, \frac{1}{9})$.

Let k be the number of cells/arrays. After the application of specific procedure aimed to remove bias and to normalize gene expressions, the data set which results from a microarray experiment consists of an expression value matrix of n rows (genes) and k columns (cells / biological samples). Once that the application of the discriminant method used to establish whether a gene i shows the expression property a or not in sample j is performed, for each $j \in \{1, \ldots, k\}$ and $i \in \{1, \ldots, n\}$, the original data set is transformed in a data set that can be represented by means of an abnormal boolean matrix $\mathbf{B} \in \{0,1\}^{n \times k}$, where the $\mathbf{B}_{ij} = 1$ if gene i in the sample j shows the expression property a, and $\mathbf{B}_{ij} = 0$ otherwise, for each $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$.

3.3 Microarray game as estimation of gene associations

Let N be a set of genes. The goal of this section and of the next one is twofold: first we want to show how the game theoretical terminology is well suited to describe the variability of genes properties across different biological conditions (e.g. normal or tumoral tissues, under pathogenetically distinct tumor types, under different treatment etc.) in the population of original cells; second, we want to introduce the probabilistic background that we need to justify the application of inferential methods on the statistics provided as the results of the game theoretical analysis. We are interested in building a TU-game (N, v) where the characteristic function v assigns to each coalitions $S \subseteq N$ the frequency of *associations* of a given biological condition with a given expression property of genes realized in the coalition S. Different expression properties for genes might be considered like, e.g., over-expression, up or down regulation, strong variation etc.

A key issue for the definition of such game (N, v) is an operational definition of what we mean for *associations* between a gene expression property and a biological condition realized in a coalition $S \in 2^N \setminus \{\emptyset\}$. In Chapter 2 we introduced an operational definition to find associations, claiming that a sufficient conditions to realize in a coalition $S \subseteq N$ the association between an expression property and a biological condition of the original cell is that all the genes which present such expression property in the cell belongs to the coalition S (*sufficiency principle for groups of genes*). Said differently, a group of genes

$S \subseteq N$ which contains all the genes showing the expression property coded by a (e.g. abnormal expression) under the biological condition of the original cell coded by t (e.g. tumoral cell) is said to realize the association between a and t. We will call the coalitions which realize the association between the expression property and the biological condition of the original cell a *winning coalition*. Note that if $m \leq n$ is the number of genes showing the expression property a, the number of winning coalitions is 2^{n-m}.

Things would be much easier if the set of winning coalitions of two cells under the same biological condition would be always the same. On the contrary, a difference in terms of expression properties of genes across cells under the same biological condition is usual, mainly due to individual, environmental and temporal variability. Moreover, all the quantitative methods used to establish the expression property of genes in a cell introduce some bias which affects the decision process of gene expression attributions. Further, the high complex network of regulative relations among genes potentially involved in a biological situation could amplify each single source of error thousands of times. Last but not least, the gene expression amount is a continuous variable which hypothetically could assume whatever value across different individuals, then it is not at all easy to identify good criteria to discriminate between different expression property (in Appendix we propose a discretization technique for this purpose).

In order to tackle this problem, we assign to each coalition $S \in 2^N \setminus \{\emptyset\}$ the expected frequency of cells in the population in which such a coalition S is a winning coalition, in formula

$$v(S) = \bar{F}^S, \tag{3.3}$$

where $\bar{F}^S$ is the expectation of F^S, i.e. the probability distribution on the set $\{0, 1\}$, where 1 means that S is a winning coalition and 0 means that S is not. Note that according to the sufficiency principle and Remark 1, the probability distribution F^S can be calculated as follows

$$F^S(1) = I\!P(\{i \in S | G_i = 0\} \neq S) I\!P(\{i \in N \setminus S | G_i = 1\} = \emptyset) \tag{3.4}$$

where G_i, $i \in \{1, \ldots, n\}$, are n (possibly dependent) random variables on the set $\{0, 1\}$, where 1 means that gene i shows the expression property a and 0 means that the gene i does not show the expression property a; consequently, $I\!P(\{i \in S | G_i = 0\} \neq S)$ is the probability that at least one gene $i \in S$ shows the expression property under consideration and $I\!P(\{i \in N \setminus S | G_i = 1\} = \emptyset)$ is the probability that no genes in $N \setminus S$ show the same expression property.

Let k be the number of samples/arrays and n be the number of genes. Consider an abnormal boolean matrix $\mathbf{B} \in \{0,1\}^{n \times k}$. For each $S \in 2^N \setminus \{\emptyset\}$ it is possible to check whether S contains the support $sp(\mathbf{B}_j)$, for each $j \in \{1, \ldots, k\}$; so, via matrix $\mathbf{B}$ we actually face $2^N - 1$ *random samples* of size k (we define a random sample of size k as a family of k independent and equally distributed random variables) from the unknown probability distribution F^S on the set $\{0,1\}$, i.e.

$$X_1^S, X_2^S, \ldots, X_k^S \sim F^S, \tag{3.5}$$

for each $S \in 2^N \setminus \{\emptyset\}$. Having observed via the boolean matrix $\mathbf{B}$ the random sample $X_1^S = x_1^S, X_2^S = x_2^S, \ldots, X_k^S = x_k^S$ for each $S \in 2^N \setminus \{\emptyset\}$, with $x_j^S \in \{0,1\}$ for each $j = 1, \ldots, k$, we can compute the sample average $\bar{x}^S = \sum_{j=1,\ldots,k} \frac{x_j^S}{n}$ for use as an estimate of the expectation of F^S. Then, we can define the TU-game $(N, \bar{v})$ where for each $S \in 2^N \setminus \{\emptyset\}$

$$\bar{v}(S) = \bar{x}^S = \sum_{j=1,\ldots,k} \frac{x_j^S}{k} \tag{3.6}$$

and $\bar{v}(\emptyset) = 0$.

Comparing relations (3.2) and (3.6), it is easy to check that $(N, \bar{v})$ is the *microarray game* corresponding to the boolean matrix $\mathbf{B}$ as defined in (3.1).

The random sample $X_1^S = x_1^S, X_2^S = x_2^S, \ldots, X_k^S = x_k^S$ provides also an estimate of the accuracy of $\bar{v}(S)$, for each $S \in 2^N \setminus \{\emptyset\}$, namely

$$\hat{\sigma}^S = \left[\frac{1}{k(k-1)} \sum_{j=1}^{k} (x_j^S - \bar{x}^S)^2\right]^{\frac{1}{2}}; \tag{3.7}$$

$\hat{\sigma}^S$ is the estimated standard error of $\bar{X}^S = \bar{x}^S = \bar{v}(S)$, the mean squared root of estimation.

Example 10 *Consider the expression matrix presented in Table 3.1. Suppose we are interested in encoding each gene $i \in \{1, \ldots, 8\}$ in each cell/biological sample j, for each $j \in \{1, \ldots, 7\}$ according to a gene expression property. Different discretization operators that, given user defined parameters, can be used to transform each numerical value from continuous gene expression data into one boolean value per gene expression property, deciding whether the true or the false value must be assigned to gene i in cell/biological sample j with respect to the expression property under consideration (Pensa* et al. *(2004)). Table 3.2 shows a*

	1	2	3	4	5	6	7
gene 1	9	0	-1	-3	0	0	8
gene 2	0	-1	-1	7	0	14	-1
gene 3	-1	-1	-2	4	6	10	13
gene 4	-1	0	0	-1	8	14	0
gene 5	1	0	5	10	0	1	-1
gene 6	0	-1	0	-2	13	8	0
gene 7	-1	14	0	7	1	-1	0
gene 8	0	0	-1	-2	0	8	13

Table 3.1: A toy example of microarray expression matrix with $n = 8$ genes and $k = 7$ cells/biological samples collected under the same biological situation.

	1	2	3	4	5	6	7
gene 1	1	0	0	0	0	0	1
gene 2	0	0	0	1	0	1	0
gene 3	0	0	0	1	1	1	1
gene 4	0	0	0	0	1	1	0
gene 5	0	0	1	1	0	0	0
gene 6	0	0	0	0	1	1	0
gene 7	0	1	0	1	0	0	0
gene 8	0	0	0	0	0	1	1

Table 3.2: Boolean matrix obtained via Algorithm 2 on the expression matrix presented in Table 3.1.

possible boolean matrix derived from the expression data in Table 3.1 via the application of the dichotomization algorithm described in Appendix A, considering the gene 'over-expression' as gene expression property and setting the parameter d equal to 0. *The corresponding microarray game* $(\{1, 2, 3, 4, 5, 6, 7, 8\}, \bar{v})$ *is reported in Table 3.3. Note that already with* $n = 8$ *genes, the number of possible coalitions is* 256 *(in general* 2^n*), making already difficult the exhaustive evaluation of the frequency of each coalition of genes in realizing the association between the expression property and the biological condition considered.*

3.4 The Shapley value of a microarray game

In Chapter 2 it has been proposed the Shapley value of a microarray game as an index suitable to evaluate the role covered by each gene in realizing the association between the expression property and the biological condition of the original cell considered. In order to support this idea, in that chapter a biologically sound axiomatic characterization of the Shapley value on the class of microarray games has been proposed. This chapter is aimed to show that the Shapley value of a microarray game is an unbiased estimator of the game on the

S	$\bar{v}(S)$	$\hat{\sigma}^S$	S	$\bar{v}(S)$	$\hat{\sigma}^S$	S	$\bar{v}(S)$	$\hat{\sigma}^S$	S	$\bar{v}(S)$	$\hat{\sigma}^S$
1	0.1429	0.1429	2 4 7	0.1429	0.1429	2 3 4 6	0.1429	0.1429	1 4 5 6 7	0.4286	0.2020
2	0.0000	0.0000	2 4 8	0.0000	0.0000	2 3 4 7	0.1429	0.1429	1 4 5 6 8	0.2857	0.1844
3	0.0000	0.0000	2 5 6	0.1429	0.1429	2 3 4 8	0.0000	0.0000	1 4 5 7 8	0.4286	0.2020
4	0.0000	0.0000	2 5 7	0.2857	0.1844	2 3 5 6	0.1429	0.1429	1 4 6 7 8	0.2857	0.1844
5	0.1429	0.1429	2 5 8	0.1429	0.1429	2 3 5 7	0.4286	0.2020	1 5 6 7 8	0.4286	0.2020
6	0.0000	0.0000	2 6 7	0.1429	0.1429	2 3 5 8	0.1429	0.1429	2 3 4 5 6	0.2857	0.1844
7	0.1429	0.1429	2 6 8	0.0000	0.0000	2 3 6 7	0.1429	0.1429	2 3 4 5 7	0.4286	0.2020
8	0.0000	0.0000	2 7 8	0.1429	0.1429	2 3 6 8	0.0000	0.0000	2 3 4 5 8	0.1429	0.1429
1 2	0.1429	0.1429	3 4 5	0.1429	0.1429	2 3 7 8	0.1429	0.1429	2 3 4 6 7	0.2857	0.1844
1 3	0.1429	0.1429	3 4 6	0.1429	0.1429	2 4 5 6	0.1429	0.1429	2 3 4 6 8	0.2857	0.1844
1 4	0.1429	0.1429	3 4 7	0.1429	0.1429	2 4 5 7	0.2857	0.1844	2 3 4 7 8	0.1429	0.1429
1 5	0.2857	0.1844	3 4 8	0.0000	0.0000	2 4 5 8	0.1429	0.1429	2 3 5 6 7	0.4286	0.2020
1 6	0.1429	0.1429	3 5 6	0.1429	0.1429	2 4 6 7	0.1429	0.1429	2 3 5 6 8	0.1429	0.1429
1 7	0.2857	0.1844	3 5 7	0.2857	0.1844	2 4 6 8	0.0000	0.0000	2 3 5 7 8	0.4286	0.2020
1 8	0.1429	0.1429	3 5 8	0.1429	0.1429	2 4 7 8	0.1429	0.1429	2 3 6 7 8	0.1429	0.1429
2 3	0.0000	0.0000	3 6 7	0.1429	0.1429	2 5 6 7	0.2857	0.1844	2 4 5 6 7	0.2857	0.1844
2 4	0.0000	0.0000	3 6 8	0.0000	0.0000	2 5 6 8	0.1429	0.1429	2 4 5 6 8	0.1429	0.1429
2 5	0.1429	0.1429	3 7 8	0.1429	0.1429	2 5 7 8	0.2857	0.1844	2 4 5 7 8	0.2857	0.1844
2 6	0.0000	0.0000	4 5 6	0.1429	0.1429	2 6 7 8	0.1429	0.1429	2 4 6 7 8	0.1429	0.1429
2 7	0.1429	0.1429	4 5 7	0.2857	0.1844	3 4 5 6	0.2857	0.1844	2 5 6 7 8	0.2857	0.1844
2 8	0.0000	0.0000	4 5 8	0.1429	0.1429	3 4 5 7	0.2857	0.1844	3 4 5 6 7	0.4286	0.2020
3 4	0.0000	0.0000	4 6 7	0.1429	0.1429	3 4 5 8	0.1429	0.1429	3 4 5 6 8	0.2857	0.1844
3 5	0.1429	0.1429	4 6 8	0.0000	0.0000	3 4 6 7	0.2857	0.1844	3 4 5 7 8	0.2857	0.1844
3 6	0.0000	0.0000	4 7 8	0.1429	0.1429	3 4 6 8	0.1429	0.1429	3 4 6 7 8	0.2857	0.1844
3 7	0.1429	0.1429	5 6 7	0.2857	0.1844	3 4 7 8	0.1429	0.1429	3 5 6 7 8	0.2857	0.1844
3 8	0.0000	0.0000	5 6 8	0.1429	0.1429	3 5 6 7	0.2857	0.1844	4 5 6 7 8	0.2857	0.1844
4 5	0.1429	0.1429	5 7 8	0.2857	0.1844	3 5 6 8	0.1429	0.1429	1 2 3 4 5 6	0.4286	0.2020
4 6	0.0000	0.0000	6 7 8	0.1429	0.1429	3 5 7 8	0.2857	0.1844	1 2 3 4 5 7	0.5714	0.2020
4 7	0.1429	0.1429	1 2 3 4	0.1429	0.1429	3 6 7 8	0.1429	0.1429	1 2 3 4 5 8	0.4286	0.2020
4 8	0.0000	0.0000	1 2 3 5	0.2857	0.1844	4 5 6 7	0.2857	0.1844	1 2 3 4 6 7	0.4286	0.2020
5 6	0.1429	0.1429	1 2 3 6	0.1429	0.1429	4 5 6 8	0.1429	0.1429	1 2 3 4 6 8	0.5714	0.2020
5 7	0.2857	0.1844	1 2 3 7	0.2857	0.1844	4 5 7 8	0.2857	0.1844	1 2 3 4 7 8	0.4286	0.2020
5 8	0.1429	0.1429	1 2 3 8	0.2857	0.1844	4 6 7 8	0.1429	0.1429	1 2 3 5 6 7	0.5714	0.2020
6 7	0.1429	0.1429	1 2 4 5	0.2857	0.1844	5 6 7 8	0.2857	0.1844	1 2 3 5 6 8	0.4286	0.2020
6 8	0.0000	0.0000	1 2 4 6	0.1429	0.1429	1 2 3 4 5	0.2857	0.1844	1 2 3 5 7 8	0.7143	0.1844
7 8	0.1429	0.1429	1 2 4 7	0.2857	0.1844	1 2 3 4 6	0.2857	0.1844	1 2 3 6 7 8	0.4286	0.2020
1 2 3	0.1429	0.1429	1 2 4 8	0.1429	0.1429	1 2 3 4 7	0.2857	0.1844	1 2 4 5 6 7	0.4286	0.2020
1 2 4	0.1429	0.1429	1 2 5 6	0.2857	0.1844	1 2 3 4 8	0.2857	0.1844	1 2 4 5 6 8	0.2857	0.1844
1 2 5	0.2857	0.1844	1 2 5 7	0.4286	0.2020	1 2 3 5 6	0.2857	0.1844	1 2 4 5 7 8	0.4286	0.2020
1 2 6	0.1429	0.1429	1 2 5 8	0.2857	0.1844	1 2 3 5 7	0.5714	0.2020	1 2 4 6 7 8	0.2857	0.1844
1 2 7	0.2857	0.1844	1 2 6 7	0.2857	0.1844	1 2 3 5 8	0.4286	0.2020	1 2 5 6 7 8	0.4286	0.2020
1 2 8	0.1429	0.1429	1 2 6 8	0.1429	0.1429	1 2 3 6 7	0.2857	0.1844	1 3 4 5 6 7	0.5714	0.2020
1 3 4	0.1429	0.1429	1 2 7 8	0.2857	0.1844	1 2 3 6 8	0.2857	0.1844	1 3 4 5 6 8	0.5714	0.2020
1 3 5	0.2857	0.1844	1 3 4 5	0.2857	0.1844	1 2 3 7 8	0.4286	0.2020	1 3 4 5 7 8	0.5714	0.2020
1 3 6	0.1429	0.1429	1 3 4 6	0.2857	0.1844	1 2 4 5 6	0.2857	0.1844	1 3 4 6 7 8	0.5714	0.2020
1 3 7	0.2857	0.1844	1 3 4 7	0.2857	0.1844	1 2 4 5 7	0.4286	0.2020	1 3 5 6 7 8	0.5714	0.2020
1 3 8	0.2857	0.1844	1 3 4 8	0.2857	0.1844	1 2 4 5 8	0.2857	0.1844	1 4 5 6 7 8	0.4286	0.2020
1 4 5	0.2857	0.1844	1 3 5 6	0.2857	0.1844	1 2 4 6 7	0.2857	0.1844	2 3 4 5 6 7	0.5714	0.2020
1 4 6	0.1429	0.1429	1 3 5 7	0.4286	0.2020	1 2 4 6 8	0.1429	0.1429	2 3 4 5 6 8	0.4286	0.2020
1 4 7	0.2857	0.1844	1 3 5 8	0.4286	0.2020	1 2 4 7 8	0.2857	0.1844	2 3 4 5 7 8	0.4286	0.2020
1 4 8	0.1429	0.1429	1 3 6 7	0.2857	0.1844	1 2 5 6 7	0.4286	0.2020	2 3 4 6 7 8	0.4286	0.2020
1 5 6	0.2857	0.1844	1 3 6 8	0.2857	0.1844	1 2 5 6 8	0.2857	0.1844	2 3 5 6 7 8	0.4286	0.2020
1 5 7	0.4286	0.2020	1 3 7 8	0.4286	0.2020	1 2 5 7 8	0.4286	0.2020	2 4 5 6 7 8	0.2857	0.1844
1 5 8	0.2857	0.1844	1 4 5 6	0.2857	0.1844	1 2 6 7 8	0.2857	0.1844	3 4 5 6 7 8	0.4286	0.2020
1 6 7	0.2857	0.1844	1 4 5 7	0.4286	0.2020	1 3 4 5 6	0.4286	0.2020	1 2 3 4 5 6 7	0.7143	0.1844
1 6 8	0.1429	0.1429	1 4 5 8	0.2857	0.1844	1 3 4 5 7	0.4286	0.2020	1 2 3 4 5 6 8	0.7143	0.1844
1 7 8	0.2857	0.1844	1 4 6 7	0.2857	0.1844	1 3 4 5 8	0.4286	0.2020	1 2 3 4 5 7 8	0.7143	0.1844
2 3 4	0.0000	0.0000	1 4 6 8	0.1429	0.1429	1 3 4 6 7	0.4286	0.2020	1 2 3 4 6 7 8	0.7143	0.1844
2 3 5	0.1429	0.1429	1 4 7 8	0.2857	0.1844	1 3 4 6 8	0.4286	0.2020	1 2 3 5 6 7 8	0.7143	0.1844
2 3 6	0.0000	0.0000	1 5 6 7	0.4286	0.2020	1 3 4 7 8	0.4286	0.2020	1 2 4 5 6 7 8	0.4286	0.2020
2 3 7	0.1429	0.1429	1 5 6 8	0.2857	0.1844	1 3 5 6 7	0.4286	0.2020	1 3 4 5 6 7 8	0.7143	0.1844
2 3 8	0.0000	0.0000	1 5 7 8	0.4286	0.2020	1 3 5 6 8	0.4286	0.2020	2 3 4 5 6 7 8	0.7143	0.1844
2 4 5	0.1429	0.1429	1 6 7 8	0.2857	0.1844	1 3 5 7 8	0.5714	0.2020	N	1.0000	0.0000
2 4 6	0.0000	0.0000	2 3 4 5	0.1429	0.1429	1 3 6 7 8	0.4286	0.2020	$\emptyset$	0.0000	0.0000

Table 3.3: The microarray game corresponding to the boolean matrix in Table 3.2.

entire population of original cells defined in (3.3).

Given a TU-game (N, v) as defined by relation (3.3), from equation (1.2) it turns out that to calculate the Shapley value of player $i \in N$ in v corresponds to calculate the expected marginal contribution, over all orderings, of gene i in realizing the association between the expression property and the considered biological condition of the original cell in the coalition of players who precede i. On the other hand, note that formula (1.2) is computationally intractable due to the number coalitions (2^n with $n = |N|$) that must be considered for each $i \in N$. Luckily, formula (1.3) can be used to reduce the computational complexity to polynomial dimension in the number of genes n.

Let P^T be the probability distribution on the set $\{0, 1\}$ where 1 means that all the genes in T show the expression property considered and no genes outside of T show the same expression property and 0 means that not all the genes in T show the expression property or at least one gene in $N \setminus T$ show the expression property, for each $S \in 2^N \setminus \{\emptyset\}$. Then, for each $S \in 2^N \setminus \{\emptyset\}$, the probability $F^S(1)$ can be calculated as follows

$$\begin{aligned} &F^S(1) = I\!P(\{i \in S | G_i = 0\} \neq S) I\!P(\{i \in N \setminus S | G_i = 1\} = \emptyset) \\ &= \textstyle\sum_{T \in 2^S \setminus \{\emptyset\}} \Big(I\!P(\{i \in T | G_i = 1\} = T) I\!P(\{i \in N \setminus T | G_i = 1\} = \emptyset) \Big) \\ &= \textstyle\sum_{T \in 2^S \setminus \{\emptyset\}} P^T(1). \end{aligned} \tag{3.8}$$

In fact, $I\!P(\{i \in T | G_i = 1\} = T)$ is the probability the exactly all the genes in T show the expression property under consideration, for each $T \in 2^N \setminus \{\emptyset\}$. Moreover, it is easy to see that

$$\bar{F}^S = F^S(1) = \textstyle\sum_{T \in 2^S \setminus \{\emptyset\}} \bar{P}^T, \tag{3.9}$$

where $\bar{P}^S$ is the expectation of the probability distribution P^S for each $S \in 2^N \setminus \{\emptyset\}$.

Consequently, it is possible to decompose the characteristics function $v(S)$ for each $S \in 2^N \setminus \{\emptyset\}$ using the basis of unanimity games in such a way that

$$v = \sum_{S \in 2^N \setminus \{\emptyset\}} \big(\bar{P}^S u_S \big), \tag{3.10}$$

where u_S is the unanimity game on coalition S.

Then, via formula (1.3), one can calculate the Shapley value of the game v as follows

$$\phi_i(v) = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{\bar{P}^S}{|S|}, \tag{3.11}$$

where $|S|$ is the cardinality of the set S.

In an analogous way to what we presented in the last section, given a boolean matrix $\mathbf{B} \in \{0,1\}^{n\times k}$ corresponding to a microarray experiment with n genes and k samples, we can check whether $sp(\mathbf{B}_j) = S$, for each $S \in 2^N \setminus \{\emptyset\}$, finally deriving $2^N - 1$ random samples of size k from the unknown probability distribution P^S on the set $\{0,1\}$, in formula

$$Z_1^S, Z_2^S, \ldots, Z_k^S \sim P^S, \tag{3.12}$$

for each $S \in 2^N \setminus \{\emptyset\}$ and with $z_j^S \in \{0,1\}$ for each $j = 1,\ldots,k$ such that $z_j^S = 1$ if in the j-th sample the set of genes which show the expression property under consideration coincide with S (i.e. $sp(\mathbf{B}_j) = S$) and with $z_j^S = 0$ otherwise. Having observed via the boolean matrix $\mathbf{B}$ the random samples $Z_1^S = z_1^S, Z_2^S = z_2^S, \ldots, Z_k^S = z_k^S$ for each $S \in 2^N \setminus \{\emptyset\}$, we can compute the sample average $\bar{z}^S = \frac{\sum_{j=1}^k z_j^S}{k}$ for use as an estimate of the expectation of P^S for each $S \in 2^N \setminus \{\emptyset\}$.

Since any linear combination of unbiased estimators is unbiased for the same linear combination of the parameters (by the linearity of the expectations), an unbiased estimator of the Shapley value $\phi(v)$ in the TU-game (N, v) is by relation (3.11) the following one

$$\bar{\phi}_i(v) = \sum_{S\in 2^N\setminus\{\emptyset\}: i\in S} \frac{\bar{z}^S}{|S|}, \tag{3.13}$$

for each $i \in N$.

Now, we want to show that the unbiased estimator $\bar{\phi}_i(v)$ of the Shapley value in the game (N, v) equals the Shapley value $\phi_i(\bar{v})$ of the microarray game $(N, \bar{v})$. First consider, for each $j \in \{1,\ldots,k\}$, the random variables Z_j^S, for each $S \in 2^N \setminus \{\emptyset\}$ can be transformed to form other random variables Y_j^i, for each $i \in N$, giving a random sample of size k such that

$$Y_1^i, Y_2^i, \ldots, Y_k^i \sim Q^i, \tag{3.14}$$

with $Y_j^i = \sum_{S\in 2^N\setminus\{\emptyset\}: i\in S} \frac{Z_j^S}{|S|}$, for each $j \in \{1,\ldots,k\}$ and for each $i \in N$.

Remark 5 Note that for each $j \in \{1,\ldots,k\}$ and each pair of coalition $S, T \in 2^N \setminus \{\emptyset\}$ with $S \neq T$, the realizations of events $Z_j^S = 1$ and $Z_j^T = 1$ are incompatible, then realizations of Y_j^i, for each $j \in \{1,\ldots,k\}$ and each $i \in N$, take values on the set $\{1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, 0\}$.

As a consequence of Remark 5, Q^i is a probability distribution on the set $\{1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{n}, 0\}$ such that

$$Q^i(\frac{1}{t}) = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S, |S| = t} P^S(1), \tag{3.15}$$

for each $t \in \{1, \dots, n\}$, and $Q^i(0) = 1 - \sum_{t \in \{1,\dots,n\}} Q^i(\frac{1}{t})$.

Now, having observed the random samples $Y_1^i = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{z_1^S}{|S|} = y_1^i, Y_2^i = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{z_2^S}{|S|} = y_2^i, \dots, Y_k^i = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{z_k^S}{|S|} = y_k^i$ for each $i \in N$, we can compute the sample average $\bar{y}^i = \frac{\sum_{j=1}^k y_j^i}{k}$ for use as an estimate of the expectation of Q^i for each $i \in N$. Note that by relation (3.13),

$$\bar{y}^i = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{\bar{z}^S}{|S|} = \bar{\phi}_i(v) \tag{3.16}$$

for each $i \in N$.

Note also that each observation of the random variable Y_j^i, for each $j \in \{1, \dots, k\}$, corresponds to the Shapley value of player $i \in N$ in the corresponding microarray game observed as realizations of the random variables X_j^S, for each $S \in 2^N \setminus \{\emptyset\}$.

Moreover note that

$$\begin{aligned}
&\phi_i(\bar{v}) = \\
&= \phi_i\big(\textstyle\sum_{j=1}^k \frac{u_{sp(\mathbf{B}_j)}}{k})\big) \\
&= \textstyle\sum_{j \in \{1,\dots,k\} : i \in sp(\mathbf{B}_j)} \Big(\frac{1}{|sp(\mathbf{B}_j)|k}\Big) \\
&= \textstyle\sum_{j=1}^k \frac{1}{k}\Big(\sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{z_j^S}{|S|}\Big) \\
&= \frac{\sum_{j=1}^k y_j^i}{k} \\
&= \bar{\phi}_i(v),
\end{aligned} \tag{3.17}$$

for each $i \in N$, where the first equality follows from relation (3.2), the second one follows from relation (1.3), the third one from Remark 5 and the fourth one from relation (3.16).

Relation (3.17) means that the estimation of Shapley value $\bar{\phi}(v)$ on the TU-game v coincide with the Shapley value of the microarray game $\bar{v}$

We can also provide an estimate of the accuracy of $\bar{\phi}_i(v)$ for each $i \in N$, namely

$$\hat{\sigma}^{\bar{\phi}_i(v)} = \Big[\frac{1}{k(k-1)} \sum_{k}^{j=1} (y_j^i - \bar{y}^i)^2\Big]^{\frac{1}{2}}; \tag{3.18}$$

$\hat{\sigma}^{\bar{\phi}_i(v)}$ for each $i \in N$ is the estimated standard error of $\bar{Y}^i = \bar{y}^i = \bar{\phi}_i(v)$.

gene i	$\phi_i(\bar{v})$	$\hat{\sigma}^{\phi_i}$
gene 1	0.19047619	0.13363062
gene 2	0.06428571	0.03915020
gene 3	0.15952381	0.05526212
gene 4	0.07619048	0.04797486
gene 5	0.17857143	0.13223131
gene 6	0.07619048	0.04797486
gene 7	0.17857143	0.13223131
gene 8	0.07619048	0.04797486

Table 3.4: Shapley value of the microarray game presented in Table 3.3 and its estimate of the accuracy.

Example 11 *Consider the boolean matrix* $\mathbf{B}$ *of Table 3.2 and the corresponding microarray game* $(N, \bar{v})$ *in Table 3.3 of Example 10. The Shapley value of the microarray game* $(N, \bar{v})$ *and its estimated standard error is reported in Table 3.4. Note that the most relevant gene according to the Shapley value* $\phi(\bar{v})$ *is gene 1 directly followed by gene 5 and gene 7 with the same relevance and gene 3, with a lower Shapley value than genes 1,5 and 7. Note that gene 3 has a standard error much lower than gene 1,5 and 7 (about* 33% *of its Shapley value against the* 66% *of the respective Shapley values for genes 1,5,and 7) so its relevance index, although a bit smaller, could be more reliable than the higher relevance index observed on the other genes.*

Next section shows that in comparing Shapley values of single genes in microarray games corresponding to different biological conditions of the original cells, the observed variability of the Shapley value across the biological samples plays an important role.

3.5 Test statistics

Consider a boolean matrix $\mathbf{B} \in \{0,1\}^{n \times k}$ corresponding to a data set from an expression microarray experiment with n genes and k cells/biological samples which has been dicretized according to a discriminant method (for example the algorithm provided in the Appendix).

Suppose that samples can be partitioned in two groups, according to two different biological conditions of the original cells (let us say condition 1 and 2) where samples are collected. Without loss of generality, let $F^1 = \{1, \ldots, h\}$ be the group of samples under condition 1 and let $F^2 = \{h+1, \ldots, k\}$ be the group of samples under condition 2, for some $h \in \{1, \ldots, k-1\}$.

Let $\mathbf{B}^{F^1} \in \{0,1\}^{n\times h}$ and $\mathbf{B}^{F^2} \in \{0,1\}^{n\times(k-h)}$ be the two matrix obtained from $\mathbf{B}$ such that $\mathbf{B}_j^{F^1} = \mathbf{B}_j$ for each $j \in \{1,\ldots,h\}$ and $\mathbf{B}_j^{F^1} = \mathbf{B}_{j+h}$ for each $j \in \{1,\ldots,k-h\}$.

Let $\bar{v}^1, \bar{v}^2 \in \mathcal{M}$ be the microarray games corresponding to the abnormal expression matrix $\mathbf{B}^{F^1}$ and $\mathbf{B}^{F^2}$, respectively. Let $\phi(\bar{v}^1)$ be the Shapley value on the game $\bar{v}^1$ and let $\phi(\bar{v}^2)$ be the Shapley value on the game $\bar{v}^2$.

We want to answer the following question: is the Shapley value of the gene i in determining the association between the expression property under consideration and condition 1 significantly different from the Shapley value of the same gene i in determining the association between the expression property under consideration and condition 2, for each $i \in N$?

Consider the following observed difference of Shapley values

$$\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2)) := |\phi_i(\bar{v}^1) - \phi_i(\bar{v}^2)|, \tag{3.19}$$

for each $i \in N$, where $\phi_i(\bar{v}^1)$ is the Shapley value of gene i in the microarray game corresponding to the boolean matrix $\mathbf{B}^{F^1}$ and $\phi_i(\bar{v}^2)$ is the Shapley value of gene i in the microarray game corresponding to the boolean matrix $\mathbf{B}^{F^2}$.

Our goal in this section is to propose a method that can test the null hypothesis that a gene has no differences of Shapley values between the two conditions 1 and 2. In fact we want to test the *null hypothesis* that $\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2)) = 0$ against the *alternative hypothesis* that $\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2)) \neq 0$.

Let P_1^S, for each $S \in 2^N \setminus \{\emptyset\}$, be the probability distribution on the set $\{0,1\}$ in the population of original cells in condition 1. As we already said in Section 3, the event 1, which can happen with probability $P_1^S(1) = I\!P(\{i \in S | G_i = 1\} = S) I\!P(\{i \in N \setminus S | G_i = 1\} = \emptyset)$, means that all the genes in S show the expression property considered and no genes outside of S show the same expression property; the event 0 means that not all the genes in S show the expression property or at least one gene in $N \setminus S$ show the expression property. Let P_2^S, for each $S \in 2^N \setminus \{\emptyset\}$, be the probability distribution on the set $\{0,1\}$ in the population of original cells in condition 2, with an analogous meaning.

Consider the random samples of size h from the unknown probability distribution P_1^S on the set $\{0,1\}$

$$Z_1^{1,S}, Z_2^{1,S}, \ldots, Z_h^{1,S} \sim P_1^S, \tag{3.20}$$

for each $S \in 2^N \setminus \{\emptyset\}$ and with $z_j^{1,S} \in \{0,1\}$ for each $j = 1,\ldots,k$ such that $z_j^{1,S} = 1$ if in the j-th sample the set of genes which show the expression property under consideration coincide with S (i.e. if $sp(\mathbf{B}_j^{F^1}) = S$) and $z_j^{1,S} = 0$ if $sp(\mathbf{B}_j^{F^1}) \neq S$.

Similarly, consider the random samples of size $k - h$ from the unknown probability distribution P_2^S on the set $\{0, 1\}$

$$Z_1^{2,S}, Z_2^{2,S}, \ldots, Z_{k-h}^{2,S} \sim P_2^S. \tag{3.21}$$

for each $S \in 2^N \setminus \{\emptyset\}$ and with $z_j^{2,S} \in \{0, 1\}$ for each $j \in \{1, \ldots, k\}$ such that $z_j^{2,S} = 1$ if in the j-th sample the set of genes which show the expression property under consideration coincide with S ($sp(\mathbf{B}_j^{F^2}) = S$) and $z_j^{2,S} = 0$ if $sp(\mathbf{B}_j^{F^2}) \neq S$.

Then the transformed random samples

$$Y_1^{f,i}, Y_2^{f,i}, \ldots, Y_{k-h}^{f,i} \sim Q_f^i. \tag{3.22}$$

introduced in (3.23) are such that $Y_j^{g,i} = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S} \frac{Z_j^{g,S}}{|S|}$ for each $j \in \{1, \ldots, k - h\}$, $i \in N$ and $f \in \{1, 2\}$, and where

$$Q_f^i(\frac{1}{t}) = \sum_{S \in 2^N \setminus \{\emptyset\} : i \in S, |S| = t} P_f^S, \tag{3.23}$$

for each $t \in \{1, \ldots, n\}$ and $f \in \{1, 2\}$.

Suppose that there are no evidences in favor of a priori assumptions concerning neither the parametric nature of probability Q_1^S and Q_2^S, nor the equality between the two probability distributions Q_1^S and Q_2^S under the null hypothesis.

In such a situation we found appropriate to use a test procedure based on a non parametric bootstrap methods of re-sampling with replacement (see Efron and Tibshirani (1993), Efron and Gong (1983) as general introduction to bootstrap methods; see Bickel (2002) as a bootstrap application to microarray analysis), which is able to test the null hypothesis of no difference between two means of two random samples without assuming under the null hypothesis that the probability distributions in the populations are the same. In this respect, remember that via relation (3.17), the Shapley value of gene i of the microarray game corresponding to $\mathbf{B}^{F^1}$ is the mean of the random sample $Y_j^{1,i}$, $j \in \{1, \ldots, h\}$, and the Shapley value of the microarray game corresponding to $\mathbf{B}^{F^2}$ is the mean of the random sample $Y_j^{2,i}$, $j \in \{1, \ldots, k - h\}$, for each $i \in N$.

We describe the nonparametric approach to estimate the (un-adjusted for multiple comparisons) p-values in the next algorithm:

Algorithm 1 (Multiple hypotheses test for Shapley differences)
INPUT: a boolean matrix $\mathbf{B} \in \{0,1\}^{n\times k}$, $n, k \in \{1, 2, \dots\}$, *with* n *rows (genes) and* k *columns (samples); a partition* $\{F^1, F^2\}$ *of the set of* k *samples; an integer number* m *of Monte Carlo bootstrap re-samples (with replacement).*
OUTPUT: a bootstrap statistics of Shapley value differences for each one of the n *genes; a vector of* n *(un-adjusted for multiple comparisons) estimated p-values.*

```
step 1 : Compute the observed Shapley value difference $\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2))$ for each $i \in N$;

step 2 : Fix $m$ as the number of Monte Carlo bootstrap re-samples (with replacement).

step 3 : for $r : 1$ to $m$ $\{$

    step 4.r Let $\boldsymbol{s}^{r,1} = (s_j^{r,1})_{j\in\{1,\dots,h\}} \in \{1,\dots,h\}^h$ and $\boldsymbol{s}^{r,2} = (s_j^{r,2})_{j\in\{1,\dots,h\}} \in \{1,\dots,k-h\}^{k-h}$ be the vectors representing the $r$-th bootstrap re-sample (with replacement) on the cells/biological samples in condition $1$ and $2$, respectively.

    step 5.r Consider the new boolean matrix $\mathbf{B}^{s^{r,1}} \in \{0,1\}^{n\times h}$ such that $\mathbf{B}_j^{s^{r,1}} = \mathbf{B}_{s_j^{r,1}}$ for each $j \in \{1,\dots,h\}$ and the boolean matrix $\mathbf{B}^{s^{r,2}} \in \{0,1\}^{n\times(k-h)}$ such that $\mathbf{B}_j^{s^{r,2}} = \mathbf{B}_{s_j^{r,2}}$ for each $j \in \{1,\dots,h-j\}$.

    step 6.r Compute the bootstrap Shapley value difference

        $$\delta_i^r(\phi(\bar{v}_r^1), \phi(\bar{v}_r^2)) := \big|\big(\phi_i(\bar{v}_r^1) - \phi(\bar{v}^1)\big) - \big(\phi_i(\bar{v}_r^2) - \phi(\bar{v}^2)\big)\big|, \qquad (3.24)$$

        for each $i \in N$, where $\bar{v}_r^1, \bar{v}_r^2 \in \mathcal{MG}$ are the microarray games corresponding to the boolean matrix $\mathbf{B}^{s^{r,1}}$ and $\mathbf{B}^{s^{r,2}}$, respectively;

$\}$

step 7 : for each $i \in N$, compute the (un-adjusted for multiple comparisons) estimate Achieved Significance Level (ASL) or p-value $p_i$ of each gene $i \in N$ in the following way =

    $$p_i = \frac{1}{m}\Big|\{r \in \{1,\dots,m\} : \delta_i^r(\phi(\bar{v}_r^1), \phi(\bar{v}_r^2)) \geq \delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2))\}\Big|. \qquad (3.25)$$
```

Remark 6 In order to preserve the ties among genes in each sample, on step 5.r, the entire columns of the boolean matrix $\mathbf{B}$ are re-sampled according to the vectors $\mathbf{s}^{r,1}$ and $\mathbf{s}^{r,2}$ defined on step 4.r, for each $r \in \{1,\dots,m\}$.

Remark 7 Subtracting $\phi(\bar{v}_r^1)$ and $\phi(\bar{v}_r^2)$ in (3.24 from the Shapley values in the game defined on matrix $\mathbf{B}^{\mathbf{s}^{r,1}}$ and $\mathbf{B}^{\mathbf{s}^{r,2}}$, respectively, makes the bootstrap Shapley values correspond to the null hypotheses that $\delta_i(\phi(\bar{v}^1) = 0$ (Efron and Tibshirani (1993), Bickel (2002)).

Note that the estimated p-values provided by bootstrap methods (with replacement) are less exact than p-values obtained from permutation tests (without replacement) (see e.g. Dudoit et al.(2002)) but, as we already mentioned, can be used to test the null hypothesis of no differences between the means of two statistics (Efron and Tibshirani (1993)) without assuming that the distributions are otherwise equal (see also Bickel (2002)).

Applying the previous algorithm to a microarray game, thousands of null hypothesis can be tested separately; so we need to consider the problem of multiple comparison. In fact, if n is the number of statistical tests, each performed at level α, if the tests are independent, the expected number of false positive is αn, which is very large for large n. It is possible to alleviate this problem by adjusting the individual p-value of the tests for multiplicity. Several methods have been proposed in literature to tackle this problem (see for a summary Amaratunga and Cabrera (2004)), mainly assuming independence of the test statistics. In Algorithm 1, test statistics are likely not independent; in fact they are statistics on the Shapley value distribution in the population of genes, which should be representative of the relevance of each gene (interacting with many others) in determining the association between the genes expression property of groups of genes and the biological condition of the original cell under consideration. On the other hand the problem of multiplicity is still there, but to establish its entity is even harder with respect to the case of test statistics independency.

Moreover, given the very high number of null hypothesis tested in a typical microarray game, aggressively adjusting the p-values for multiplicity could seriously impede the ability of the test to find genes with respective relevance index which are truly different under the two biological conditions at hand.

Traditional statistical procedures often control the family-wise error rate (FWER), i.e. the probability that at least one of the true null hypothesis is rejected. Classical p-value adjustment methods for multiple comparisons which control FWER have been found to be too conservative in analyzing differential expression in large-screening microarray data, and the False Discovery Rate

(FDR), i.e. the expected proportion of false positives among all positives, has been recently suggested as an alternative for controlling false positives (Benjamini and Hochberg (1995), Dudoit *et al.* (2002)). It is not possible at this moment to express similar considerations suitable for the game theoretical context in which we are moving.

For all these reasons, in the sequel we separately present the results provided by our method controlling for the FDR and for the FWER, respectively, facing the problem of possible dependent statistical tests. One possible approach is to make estimation for both FDR and FWER using again re-sampling methods (Bickel (2002), Jain *et al.* (2005)).

Let $V(c)$ be the average number of bootstrap Shapley value differences equal to or greater than c, in formula

$$V(c) = \frac{1}{m}\sum_{r=1}^{m}\Big|\{i \in N : \delta_i^r(\phi(\bar{v}_r^1), \phi(\bar{v}_r^2)) \geq c\}\Big|, \tag{3.26}$$

with the convention that the cardinality of the empty set is zero, i.e. $|\emptyset| = 0$. Let $R(c)$ be the average number of observed Shapley value differences equal to or greater than c, in formula

$$R(c) = \Big|\{i \in N : \delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2)) \geq c\}\Big|. \tag{3.27}$$

The simplest way to estimate FDR at the a threshold value c is obtained via the following relation (Bickel (2002), Jain *et al.* (2005))

$$\widehat{FDR}(c) = \frac{V(c)}{R(c)}, \tag{3.28}$$

to control the estimated FDR at a level ϵ, let γ be the minimum value of $\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2))$ for which $\widehat{FDR}(\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2))) \leq \epsilon$ and reject the j-th null hypothesis if $\delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2)) \geq \gamma$.

For what concerns controlling the FWER, as we already said different approach have been proposed. Here we present a method to adjust the p-values obtained in step 7 of Algorithm 1 according to a procedure introduced in Bickel (2002). For each $i \in N$, consider the adjusted p-value $\tilde{p}_i$ defined as follows

$$\tilde{p}_i = \frac{1}{m}\Big|\{r \in \{1, \ldots, m\} : max_{j \in N}\big(\delta_j^r(\phi(\bar{v}_r^1), \phi(\bar{v}_r^2))\big) \geq \delta_i(\phi(\bar{v}^1), \phi(\bar{v}^2))\}\Big|; \tag{3.29}$$

given the FWER α', reject the i-th null hypothesis if $\tilde{p}_i \leq \alpha'$.

	1	2	3	4	5	6	7
gene 1	9	-1	8	0	-3	0	1
gene 2	7	14	-1	-2	0	0	-1
gene 3	-1	-2	-1	4	6	10	13
gene 4	0	8	14	-1	-1	0	0
gene 5	5	1	10	0	0	1	-1
gene 6	13	0	8	0	0	-1	-2
gene 7	1	14	7	0	0	-1	-1
gene 8	8	0	13	0	-2	-1	0

Table 3.5: Another toy example of microarray expression matrix with $n = 8$ genes and $k = 7$ samples and the same expression values as in Table 3.1.

	1	2	3	4	5	6	7
gene 1	1	0	1	0	0	0	0
gene 2	1	1	0	0	0	0	0
gene 3	0	0	0	1	1	1	1
gene 4	0	1	1	0	0	0	0
gene 5	1	0	1	0	0	0	0
gene 6	1	0	1	0	0	0	0
gene 7	0	1	1	0	0	0	0
gene 8	1	0	1	0	0	0	0

Table 3.6: Boolean matrix obtained by the application of Algorithm 2 to the expression matrix in Table 3.5.

Example 12 *Consider another toy example of microarray expression matrix with* $n = 8$ *genes and* $k = 7$ *samples Note that the expression vector of each gene* $i \in \{1, \ldots, 8\}$ *takes the same values of the corresponding* i*-th expression vector in Table 3.1 (in other terms, the values in the* i*-th row, for each* $i \in \{1, \ldots, 8\}$*, in Table 3.1 are obtained as a permutations of the values in the* i*-th row in Table 3.1. So, no genes con be considered abnormally expressed with respect their expression profiles in Table 3.1 and 3.5, respectively. Following the terminology used in this section, we will refer to this new situation as the condition* 2*, whereas the situation introduced in Example 10 will be referred as condition* 1*.*

If we apply the same discriminant method used in Example 10 on Table 3.1, i.e. Algorithm 2 described in Appendix A with the same input parameters, we find the same cutoffs used to obtain the Boolean matrix in Table 3.2. Given the new form of the expression matrix in Table 3.5, the corresponding boolean matrix will be the one presented in Table 3.6. We escape the presentation of the table representing the microarray game corresponding to the new boolean matrix in Table 3.6. Instead we directly present, the table of the Shapley values and their estimated standard errors in Table 3.7. Comparing the respective Shapley value for each gene $i \in \{1, \ldots, 8\}$ *in Tables 3.4 and 3.7, it is interesting to see that in*

gene i	$\phi_i(v)$	$\hat{\sigma}^{\phi_i}$
gene 1	0.05238095	0.03181045
gene 2	0.07619048	0.04797486
gene 3	0.57142857	0.18898224
gene 4	0.07142857	0.04636239
gene 5	0.05238095	0.03181045
gene 6	0.05238095	0.03181045
gene 7	0.07142857	0.17857143
gene 8	0.05238095	0.07619048

Table 3.7: Shapley value of the microarray game corresponding to the boolean matrix presented in Table 3.6 and its accuracy.

gene i	p-value
gene 1	0.364
gene 2	0.838
gene 3	0.032
gene 4	0.939
gene 5	0.420
gene 6	0.746
gene 7	0.444
gene 8	0.715

Table 3.8: Un-adjusted for multiple comparisons p-values obtained by Algorithm 1 applied to the microarray expression matrix presented in Table 3.1 and 3.5.

this new situations only genes 2 *and* 3 *increase their respective Shapley values. Gene* 3 *increases its Shapley value of about four times even if its expression vector is exactly the same of Table 3.1 (expression values for row* 3 *in Table 3.1 have not been permuted in Table 3.5).*

Applying the test method described in Algorithm 1 to the boolean matrix in Table 3.6 with 1000 re-samples, we estimated the table of (un-adjusted for multiple comparisons) p-values presented in Table 3.8

Performing the control of the estimated FDR as introduced in relation (3.28) at a level $\epsilon = 0.05$*, the unique null hypothesis of no Shapley value differences between condition* 1 *and* 2 *that can be rejected is the one for gene* 3*. The same conclusion is inferred performing the control for the adjusted p-values as introduced in relation (3.29) at a FWER* $\alpha' = 0.05$.

3.6 Analysis of real data

This section is devoted to the presentation of the results from the application of Algorithm 1 to the gene expression data-set (3051 genes and 38 tumor mRNA samples, 27 acute lymphoblastic leukemia (ALL) cases and 11 acute myeloid leukemia (AML)) from the leukemia microarray study of Golub *et al.* (1999). Pre-processing was done as described in Dudoit *et al.* (2002), implemented by the R Development Core Team (2004) code in the Bioconductor package *multtest* (Gentleman *et al.* (2204)).

The resulting real-valued expression matrix (3051 rows, 38 columns) has been dichotomized according to Algorithm 2 described in Appendix A. Sorting the real-valued expression data of each gene (see step 2 in Algorithm 2), it has been observed that many genes presented at most three very low values with respect the average expression. One consequence was that the thresholds for binarization selected by Algorithm 2 took the position very close to the big jump corresponding to such low-bound outliers. For this reason, it has been decided to exclude the three lowest value of each row in the application of Algorithm 2 and hence to set the low-bound outliers parameter $d = 3$. Thresholds selected by Algorithm 2 for nine genes have been presented in Figure 3.1. The output from the application of Algorithm 2 was stored in a boolean matrix $\mathbf{B} \in \{0,1\}^{3051\times 38}$.

Algorithm 1 has been applied to the boolean matrix $\mathbf{B}$, with the partition $\{F^1, F^2\}$ of the 38 samples such that all the 11 AML samples belong to the set F^1 and the remaining 27 ALL samples belong to the set F^2. The number of bootstrap re-samples with replacement was $m = 1000$. Figure 3.2 shows, for five genes, the histograms of the Shapley values observed in the two classes of samples and the corresponding bootstrap statistic of Shapley differences.

Figure 3.3 shows the QQ plot of the observed Shapley value differences and the expected Shapley value differences produced by Algorithm 1. The graph shows 70 genes whose difference in terms of Shapley values is greater than 0.0004614116 and 77 genes whose difference in terms of Shapley values is lower than -0.0004614116, for a total number of 147 genes which corresponds to the number of rejected null hypothesis when the estimated FDR is controlled at a level 0.05, according with relation (3.28).

Figure 3.4 shows the plotting of the corrected p-values controlling the FWER at a level of 0.05 using the procedure introduced via relation (3.29). The set of 40 null hypothesis rejected in this case is a subset of the set of 147 rejected null hypothesis controlling the estimated FDR at level 0.05. In Table 3.9 are reported details for the identification of the 40 genes corresponding to the rejected null hypothesis controlling the FWER at the level of 0.05, together with the Shapley value observed in the game built on the 11 AML samples (fourth column), the Shapley value observed in the game built on the 27 ALL samples (fifth column) and the corresponding adjusted p-value provided by the FWER control at the level 0.05. Similar information is provided in Table 3.10, concerning the remaining 107 rejected null hypothesis when the control is performed on the estimated FDR at a level of 0.05.

The Shapley values distributions observed in the two sample groups have been plotted in Figure 3.5.

3.7 Discussion

In this chapter, a new method to analyze the relevance of genes in the mechanisms which provoke a biological condition or a response of interest has been described, based on the game theoretical model introduced in Chapter 2. The main novelty of the approach, with respect the model of Chapter 2, is that the present method considers the stochastic process governing the gene expression observations and its level of influence in determining differences of the observed relevance index of genes under two distinct biological conditions or responses of interest.

An application of the method to the gene expression data set from the leukemia microarray study by Golub *et al.* (1999) is presented.

Preliminary examination of the literature concerning the role of the significant genes presented in Table 3.8 provides evidence of the effective capacity of Algorithm 2 in selecting genes with an effective role in pathogenesis of subtypes of leukemia. For example, over-expression of the TCL1 oncogene has been shown to play a causative role in T cell leukemias of humans and mice (Narducci *et al.* (2002). IFI 16 gene product is a nucleoprotein expressed in association with the differentiation of myeloid precursor cell lines (Dawson and Trapani (1995)).

Other genes were already known as leukemia markers. As already observed by Golub *et al.* (1999), CD33 and MB-1 encode cell surface proteins for which monoclonal antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells.

Other markers of hematopoietic lineage provided in Table 3.8, observed again by Golub *et al.* (1999) and related to cancer pathogenesis, are Cyclin D3, which encodes proteins critical for S-phase cell cycle progression, and zyxin, which encodes proteins for adhesion.

SPTAN1 is involved in secretion and it interacts with calmodulin in a calcium-dependent manner and has been indicated as good marker for ALL in two studies using gene expression data to distinguish subtypes of leukemia (Armstrong *et al.* (2002), Tan *et al.* (2005)).

LYN is an oncogene and Hasegawa *et al.* (2001) proved that the expression of Cd19, which is also in Table 3.8, is required for the development of autoimmunity in Lyn deficient mice. Note that the Shapley value of LYN is 0.000729 in the AML microarray game and it is 0 in the ALL microarray game, whereas the Shapley value of Cd19 is more or less the opposite with respect to LYN.

3.8 Figures and Tables

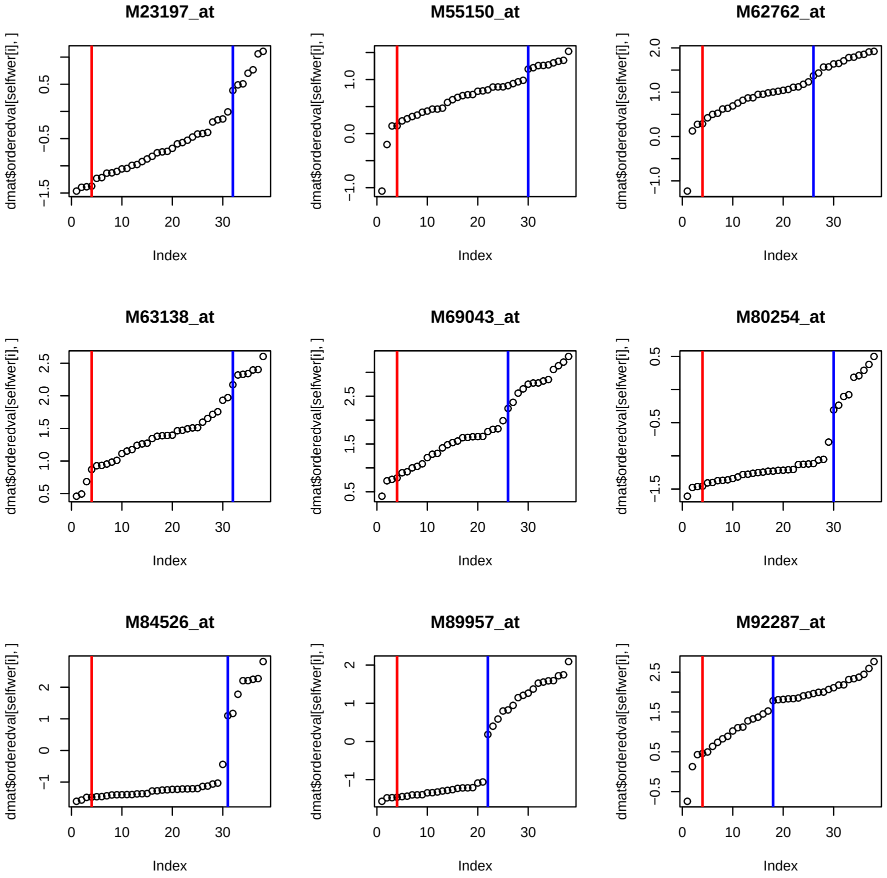

Figure 3.1: Sorted expression values (on the y-axis) corresponding to 38 samples (labels on the x-axis) for nine genes of the Golub *et al.* (1999) data-set. Dichotomization thresholds (y-coordinate of circles placed on the right vertical straight line) have been selected by Algorithm 2. Real-valued expressions which are strictly lower than the threshold have been labelled by 0, whereas expression values higher than or equal to the threshold have been labelled by 1. Values on the left side of the index labelled by the left vertical straight line have not been taken into account in the algorithmic computation of the thresholds.

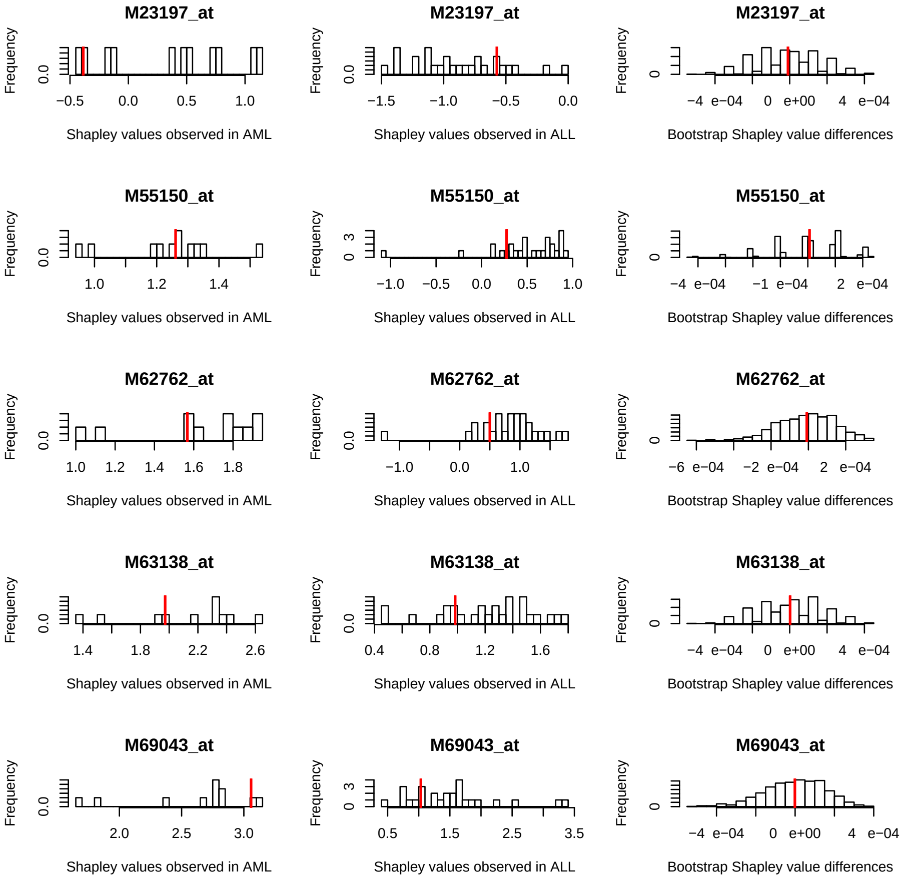

Figure 3.2: In the first two columns of the figure, the histograms of the Shapley values observed for genes in AML and ALL samples, respectively, have been given, for five genes. The last column shows the corresponding histograms of Shapley value differences obtained via the bootstrap procedure described in Algorithm 1 under the null hypothesis of no difference between the Shapley values computed under the two conditions AML and ALL. The vertical straight line indicates the mean of the respective distributions.

ID	Gene name	Gene code	Shapley value in AML samples	Shapley value in ALL samples	Adjusted p-value
1	FAH Fumarylacetoacetate	M55150_at	0.000939	0	0
2	Cytoplasmic dynein light chain 1 (hdlc1) mRNA	U32944_at	0	0.000925	0
3	Leukotriene C4 synthase (LTC4S) gene	U50136_rna1_at	0.000937	0	0
4	Zyxin	X95735_at	0.001142	0	0
5	CCND3 Cyclin D3	M92287_at	0	0.000896	0.001
6	CD22 CD22 antigen	X59350_at	0.000105	0.000981	0.001
7	Interleukin 8 (IL8) gene	M28130_rna1_s_at	0.001039	0.000132	0.001
8	MYL1 Myosin light chain (alkali)	M31211_s_at	0	0.000898	0.001
9	CYSTATIN A	D88422_at	0.000935	8.86E-05	0.002
10	GLUTATHIONE S-TRANSFERASE, MICROSOMAL	U46499_at	0.000935	8.08E-05	0.002
11	PROTEASOME IOTA CHAIN	X59417_at	0	0.000852	0.002
12	INTERLEUKIN-8 PRECURSOR	Y00787_s_at	0.001039	0.000174	0.002
13	DF D component of complement (adipsin)	M84526_at	0.000832	0	0.003
14	Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA	U46751_at	0.000829	0	0.003
15	Small Nuclear Ribonucleoprotein, Polypeptide C, Alt. Splice 2	HG1322-HT5143_s_at	0.000204	0.001021	0.004
16	Inducible protein mRNA	L47738_at	0.000207	0.001013	0.005
17	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR	M80254_at	0.000829	4.56E-05	0.007
18	GB DEF = (lambda) DNA for immunoglobin light chain	D88270_at	0	0.000764	0.01
19	ATP6C Vacuolar H+ ATPase proton channel subunit	M62762_at	0.000935	0.000168	0.01
20	MB-1 gene	U05259_rna1_at	0	0.000764	0.01
21	CD19 gene	M84371_rna1_s_at	0	0.000764	0.01
22	DHPS Deoxyhypusine synthase	U26266_s_at	0.00021	0.000975	0.01
23	ANPEP Alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13)	M22324_at	0.000837	7.87E-05	0.011
24	MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3	M69043_at	0.000928	0.000171	0.011
25	Dihydropyrimidinase related protein-2	U97105_at	0.000305	0.001061	0.013
26	Interferon-gamma induced protein (IFI 16) gene	M63838_s_at	0.00031	0.001054	0.017
27	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	M16038_at	0.000729	0	0.021
28	CD33 CD33 antigen (differentiation antigen)	M23197_at	0.000729	0	0.021
29	IGB Immunoglobulin-associated beta (B29)	M89957_at	0	0.000726	0.021
30	CTSD Cathepsin D (lysosomal aspartyl protease)	M63138_at	0.000723	0	0.023
31	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	J05243_at	0.000423	0.001144	0.025
32	GB DEF = Neurotensin receptor	X70070_at	0.000928	0.000207	0.025
33	TCL1 gene (T cell leukemia) extracted from H.sapiens mRNA for Tcell leukemia/lymphoma 1	X82240_rna1_at	0	0.000722	0.025
34	KIAA0097 gene	D43948_at	0	0.000718	0.026
35	NF-IL6-beta protein mRNA	M83667_rna1_s_at	0.000829	0.000129	0.035
36	ALDR1 Aldehyde reductase 1 (low Km aldose reductase)	X15414_at	0.000407	0.001106	0.036
37	Gal-beta(1-3/1-4)GlcNAc alpha-2.3-sialyltransferase	X74570_at	0.00104	0.000339	0.036
38	Spermidine/spermine N1-acetyltransferase (SSAT) gene	U40369_rna1_at	0.000829	0.000132	0.038
39	KIAA0200 gene	D83785_at	0.000104	0.000798	0.039
40	Nuclear factor NF45 mRNA	U10323_at	0.000205	0.000895	0.045

Table 3.9: Genes with adjusted p-values for FWER control lower than 0.05.

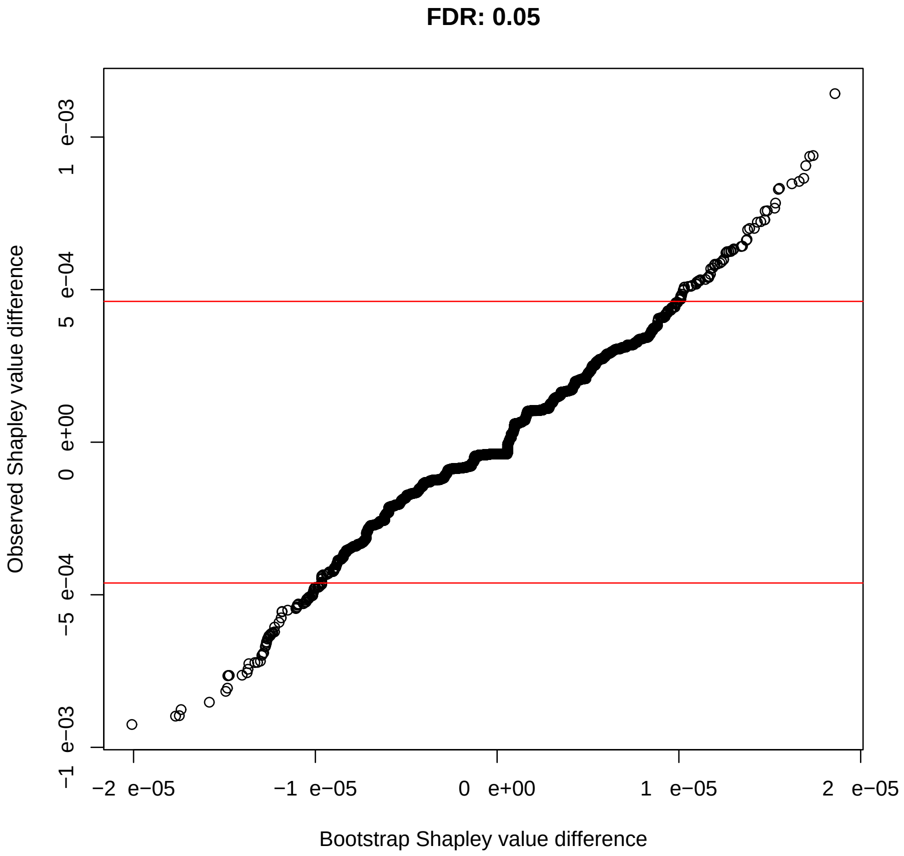

Figure 3.3: QQ plot of the observed Shapley value differences and the expected Shapley value differences produced by Algorithm 1. Null hypothesis of genes with observed Shapley values difference outside the interval between the two horizontal straight lines have been rejected controlling the FDR at the level 0.05.

ID	Gene name	Gene code	Shapley value in AML samples	Shapley value in ALL samples	Adjusted p-value
41	TCF12 Transcription factor 12 (HTF4, helix-loop-helix transcription factors 4)	M83233_at	0	0.000671	0.075
42	Transcription factor (CBFB) mRNA, 3' end	L20298_at	0.000304	0.00097	0.08
43	Fc-epsilon-receptor gamma-chain mRNA	M33195_at	0.000834	0.000169	0.083
44	ANX1 Annexin I (lipocortin I)	X05908_at	0.001041	0.000379	0.088
45	SERYL-TRNA SYNTHETASE	X91257_at	0.000312	0.000971	0.092
46	GTF2E2 General transcription factor TFIIE beta subunit, 34 kD	X63469_at	0.000313	0.000967	0.101
47	KIAA0235 gene, partial cds	D87078_at	0.00011	0.000755	0.113
48	PIM1 Pim-1 oncogene	M16750_s_at	0.000724	8.16E-05	0.116
49	Adenosine triphosphatase, calcium	Z69881_at	0.000424	0.001064	0.12
50	V-ERBA RELATED PROTEIN EAR-1	M24900_at	0.000928	0.000287	0.12
51	KIAA0067 gene	D31891_at	0.000208	0.000843	0.131
52	CD24 signal transducer mRNA and 3' region	L33930_s_at	0	0.000634	0.131
53	FTH1 Ferritin heavy chain	L20941_at	0.000724	9.05E-05	0.133
54	Orphan receptor mRNA, partial cds	U07132_at	0.000929	0.000299	0.139
55	IRF2 Interferon regulatory factor 2	X15949_at	0	0.000629	0.141
56	CA2 Carbonic anhydrase II	Y00339_s_at	0.000627	0	0.148
57	MEF2A gene (myocyte-specific enhancer factor 2A, C9 form) extracted from Human myocyte-specific enhancer factor 2A (MEF2A) gene, first coding	U49020_cds2_s_at	0	0.000626	0.15
58	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)	M27891_at	0.000625	0	0.153
59	TRANSFORMATION-SENSITIVE PROTEIN IEF SSP 3521	M86752_at	0.000311	0.000936	0.153
60	Azurocidin gene	M96326_rna1_at	0.000624	0	0.154
61	MANB Mannosidase alpha-B (lysosomal)	U05572_s_at	0.000624	0	0.154
62	ZNF91 Zinc finger protein 91 (HPF7, HTF10)	L11672_at	0.000521	0.001144	0.156
63	LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)	M57710_at	0.000622	0	0.158
64	PFC Properdin P factor, complement	M83652_s_at	0.000622	0	0.158
65	Pre-Mrna Splicing Factor Sf2p33, Alt. Splice Form 1	HG3546-HT3744_s_at	9.76E-05	0.000719	0.158
66	Nuclear Factor Nf-Il6	HG3494-HT3688_at	0.00062	0	0.159
67	PRKAR1A CAMP-dependent protein kinase regulatory subunit type I	M33336_at	0.000321	0.000926	0.195
68	BZIP protein NF-IL3A (IL3BP1) mRNA	U26173_s_at	0.000726	0.000127	0.217
69	ORF, Xq terminal portion	D16469_at	0.000937	0.000339	0.219
70	SMT3A protein	X99584_at	0.000935	0.000343	0.237
71	SNRPN Small nuclear ribonucleoprotein polypeptide N	J04615_at	0.000512	0.001102	0.243
72	ARHG Ras homolog gene family, member G (rho G)	X61587_at	0.001142	0.000555	0.255
73	AFFX-HUMTFRR/M11507_M_at (endogenous control)	AFFX-HUMTFRR/M11507_M_at	0.000627	4.39E-05	0.282
74	SELL Leukocyte adhesion protein beta subunit	M15395_at	0.000625	4.21E-05	0.284
75	SPI1 Spleen focus forming virus (SFFV) proviral integration oncogene spi1	X52056_at	0.000834	0.000254	0.288
76	MCM3 Minichromosome maintenance deficient (S. cerevisiae) 3	D38073_at	0.000524	0.0011	0.304
77	NFKB2 Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)	S76638_at	0.000614	4.21E-05	0.321
78	Epican, Alt. Splice 1	HG2981-HT3125_s_at	0.001039	0.00047	0.337
79	BLK Protein-tyrosine kinase blk	S76617_at	0	0.000556	0.391
80	Protein phosphatase 2A 74 kDa regulatory subunit (delta or B	subunit)"	L76702_at	0.000208	0.017788
81	DGUOK Deoxyguanosine kinase	U41668_at	0.000425	0.000975	0.421
82	HMOX1 Heme oxygenase (decycling) 1	X06985_at	0.001142	0.000592	0.422
83	Novel T-cell activation protein	X94232_at	0.000518	0.001064	0.438
84	Clone CIITA-8 MHC class II transactivator CIITA mRNA	U18259_at	0	0.000545	0.442
85	KIAA0063 gene	D31884_at	0.001142	0.000599	0.444
86	ORF mRNA	M68864_at	0.00052	0.001063	0.445
87	K+ channel beta 2 subunit mRNA	U33429_at	0.000621	8.08E-05	0.461
88	HS1 binding protein HAX-1 mRNA, nuclear gene encoding mitochondrial protein	U68566_at	0.00031	0.00085	0.461
89	PIM1 Pim-1 oncogene	M54915_s_at	0.000621	8.16E-05	0.461
90	ACADM Acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	M91432_at	0.00052	0.001054	0.482
91	Serine palmitoyltransferase (LCB2) mRNA, partial cds	U15555_at	0.00062	8.68E-05	0.483
92	MEF2C MADS box transcription enhancer factor 2, polypeptide C (myocyte enhancer factor 2C)	L08895_at	0.000105	0.000639	0.483

Table 3.10: Genes corresponding to rejected null hypothesis (together with genes in Table 3.9) when controlling the estimated FDR at the level 0.05 (follows).

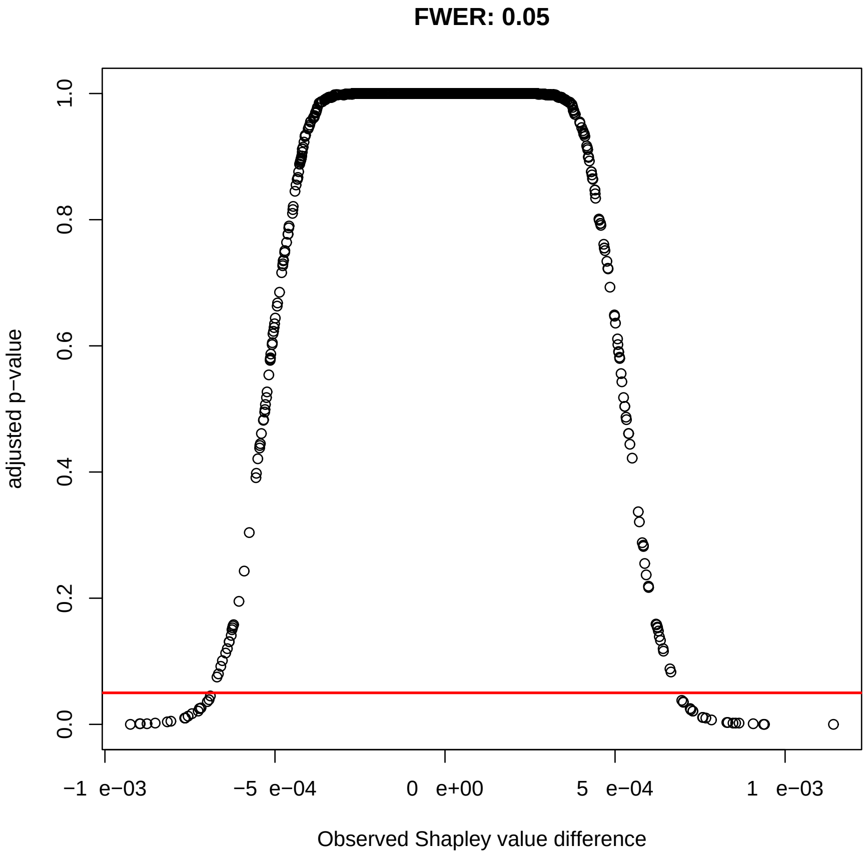

Figure 3.4: Plot of the corrected p-values controlling the FWER at a level of 0.05. The null hypothesis of genes with $\tilde{p}$ below the horizontal straight line have been rejected.

93	OAT Ornithine aminotransferase (gyrate atrophy)	M29927_at	0.001037	0.000505	0.487
94	ATM Ataxia telangiectasia mutated (includes complementation groups A, C and D)	U33841_at	0.000104	0.000634	0.495
95	Ionotropic ATP receptor P2X5a mRNA	U49395_at	0.000104	0.000633	0.499
96	Major Histocompatibility Complex, Dg	HG1872-HT1907_at	0.000827	0.000298	0.504
97	SELL Leukocyte adhesion protein beta subunit	X64072_s_at	0.001039	0.00051	0.504
98	Chromosome segregation gene homolog CAS mRNA	U33286_at	0.000104	0.000632	0.507
99	Mac25	HG987-HT987_at	0.001039	0.000514	0.518
100	Estrogen sulfotransferase mRNA	U20499_at	0.000619	0.001144	0.518
101	Macmarcks	HG1612-HT1612_at	0.000621	0.001144	0.527
102	CATHEPSIN G PRECURSOR	J04990_at	0.00052	0	0.543
103	IL2RG Interleukin 2 receptor gamma chain	D11086_at	0.000626	0.001144	0.554
104	GB DEF = Beta-2 integrin alphaD subunit (ITGAD) gene, exons 25-30, and partial cds	U40279_at	0.000718	0.000201	0.556
105	Oncoprotein 18 (Op18) gene	M31303_rna1_at	0.00063	0.001144	0.577
106	Folylpolyglutamate synthetase mRNA	M98045_at	0.00021	0.000724	0.579
107	Epican, Alt. Splice 11	HG2981-HT3127_s_at	0.000514	0	0.58
108	Putative enterocyte differentiation promoting factor mRNA, partial cds	U62136_at	0.000631	0.001144	0.581
109	PPGB Protective protein for beta-galactosidase (galactosialidosis)	M22960_at	0.000513	0	0.582
110	HLA CLASS I HISTOCOMPATIBILITY ANTIGEN, F ALPHA CHAIN PRECURSOR	X17093_at	0	0.000512	0.587
111	Adult heart mRNA for neutral calponin	D83735_at	0.000935	0.000425	0.59
112	CaM kinase II isoform mRNA	U81554_at	0.001142	0.000632	0.591
113	MST1R Protein-tyrosine kinase RON	X70040_at	0.001142	0.000634	0.602
114	OBF-1 mRNA for octamer binding factor 1	Z49194_at	0	0.000508	0.602
115	POLYPOSIS LOCUS PROTEIN 1	M73547_at	0	0.000508	0.605
116	DLX7 Distal-less homeobox 7	U73328_at	0.001142	0.000635	0.611
117	HKR-T1	S50223_at	0	0.000506	0.619
118	DNA-dependent protein kinase catalytic subunit (DNA-PKcs) mRNA	U47077_at	0	0.000504	0.623
119	Butyrophilin (BTF5) mRNA	U90552_s_at	0	0.000503	0.629
120	Clone 23721 mRNA sequence	U79291_at	0	0.000501	0.635
121	GB DEF = Cystic fibrosis antigen mRNA	M26311_s_at	0.000935	0.000434	0.636
122	Transcription factor SIM2 long form mRNA	U80457_at	0.000214	0.000713	0.644
123	PRG1 Proteoglycan 1, secretory granule	X17042_at	0.001142	0.000644	0.647
124	SNCA Synuclein, alpha (non A4 component of amyloid precursor)	U46901_at	0.000623	0.000125	0.649
125	Carboxyl Methyltransferase, Aspartate, Alt. Splice 1	HG1400-HT1400_s_at	0.000518	0.001012	0.663
126	ATP6E ATPase, H+ transporting, lysosomal (vacuolar proton pump) 31kD	X71490_at	9.76E-05	0.00059	0.668
127	Autoantigen DFS70 mRNA, partial cds	U94319_at	0.000105	0.000592	0.685
128	Tryptase-III mRNA, 3' end	M33493_s_at	0.000524	3.88E-05	0.693
129	PRKCB1 Protein kinase C, beta 1	X06318_at	0.000105	0.000586	0.716
130	Lysophosphatidic acid acyltransferase-beta mRNA	U56418_at	0.001038	0.000558	0.722
131	CSNK1D Casein kinase 1, delta	U29171_at	0.000524	4.49E-05	0.723
132	KIAA0030 gene, partial cds	D21063_at	0.000629	0.001106	0.727
133	LPAP gene	X97267_rna1_s_at	0.000629	0.001106	0.73
134	RSU-1/RSP-1 mRNA	L12535_at	0.001142	0.000666	0.734
135	TCRB T-cell receptor, beta cluster	M12886_at	0.000415	0.000891	0.735
136	PHB Prohibitin	S85655_at	0.000207	0.000683	0.735
137	DAGK1 Diacylglycerol kinase, alpha (80kD)	X62535_at	0.00063	0.001106	0.735
138	TFIID subunit TAFII55 (TAFII55) mRNA	U18062_at	0.000627	0.001102	0.736
139	Transcriptional activator hSNF2b	D26156_s_at	0	0.000471	0.748
140	LGALS1 Ubiquinol-cytochrome c reductase core protein II	J04456_at	0.000937	0.000467	0.751
141	Non-histone chromosomal protein (NHC) mRNA	U90549_at	0.00063	0.001101	0.751
142	PRKCD Protein kinase C, delta	D10495_at	0.000514	4.56E-05	0.755
143	GLUL Glutamate-ammonia ligase (glutamine synthase)	M63438_s_at	0.001142	0.000674	0.755
144	Low-Mr GTP-binding protein (RAB32) mRNA, partial cds	U59878_at	0.000927	0.000461	0.761
145	GOT2 Glutamic-oxaloacetic transaminase 2, mitochondrial (aspartate aminotransferase 2)	M22632_at	0	0.000466	0.764
146	DNA-binding protein ABP/ZF mRNA	U82613_at	0.000208	0.00067	0.777
147	MHC-encoded proteasome subunit gene LAMP7-E1 gene (proteasome subunit LMP7) extracted from H.sapiens gene for major histocompatibility complex encoded proteasome subunit LMP7	Z14982_rna1_at	0.00052	0.000981	0.777

Follows Table 3.10.

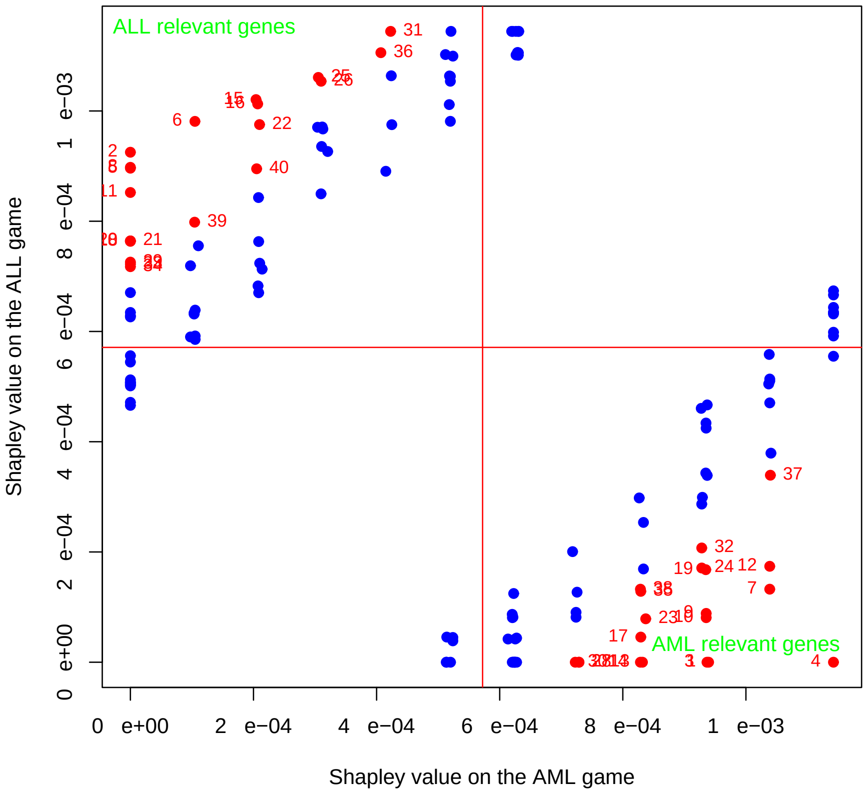

Figure 3.5: Shapley values on AML and ALL corresponding to 147 rejected null hypothesis controlling the FDR at the level of 0.05. Dots labelled with numbers correspond to genes with adjusted p-values $\tilde{p}$ lower than 0.05. Numerical labels refer to ID numbers in Table 3.9 and 3.10.

Chapter 4

Other games on gene expression data

4.1 Minimum cost spanning tree and gene expression

Microarray games are based on a dichotomization process applied to a real-valued expression matrix. In this chapter we present an alternative class of cooperative games where the dichotomization process is not required. On the other hand, a different way to asses the value of each coalition must be used to avoid the arbitrariness in choosing the cutoffs to dichotomize the expression data. First, a method based on the framework of Minimum Cost Spanning Trees (MCSTs) has been introduced to represent the interaction structure of the involved genes. In gene expression analysis Xu *et al.* (2001) and Speer *et al.* (2003) have already used MCSTs representation as starting point for clustering algorithms. Here we exploit the MCSTs representation of a gene expression data-set to construct a corresponding MCST game.

Similarly to the case of microarray games, the objective of this different class of games is again to evaluate the genes relevance in provoking a biological condition or response of interest. With this goal, the approach here is to describe the overall level of *similarity* (or *dissimilarity*) of each sub-group of genes with a reference vector of expressions representing a pre-selected gene (or a group of pre-selected genes) which is assumed not to be involved in the processes related to the biological condition or response of interest under studying. Differently stated, observed fluctuations in the expression values of a reference gene across the samples, should uniquely be ascribed to experimental random noise.

Note that the concept of association between expression properties and biological conditions realized in coalitions of genes, in this context must be reformulated to catch the idea of association between real valued gene expression vectors and biological conditions. The smaller the overall level of similarity of a coalition with a reference gene expression vector is, the higher the level of association of real valued gene expression vectors with the biological condition or response of interest should be.

To avoid the arbitrariness related to the dichotomization cutoffs is not the only advantage of this new approach. In comparison with the microarray game model, this new model based on MCSTs also improves the resolution of the analysis. For example, it is well known that genes that are co-regulated by common transcription factors have similar expression patterns. A characteristic function based on levels of similarity among genes is highly representative of the influence of different transcription factors in all possible clusters of co-regulated genes. With respect to microarray games, the application of a relevance index to games based on MCST representation would be more efficient in selecting those genes who are able to regulate other genes.

4.1.1 Preliminary notations

Here we introduce some basic notions on graph theory. An (undirected) *graph* is a pair $< V, E >$, where V is a set of vertices or nodes and E is a set of edges e of the form $\{i, j\}$ with $i, j \in V$, $i \neq j$. The *complete graph* on a set V of vertices is the graph $< V, E_V >$, where $E_V = \{\{i, j\} | i, j \in V \text{ and } i \neq j\}$. A *path* between i and j in a graph $< V, E >$ is a sequence of nodes $(i_0, i_1, \ldots, i_k)$, where $i = i_0$ and $j = i_k$, $k \geq 1$, and such that $\{i_s, i_{s+1}\} \in E$ for each $s \in \{0, \ldots, k-1\}$. A path $(i_0, i_1, \ldots, i_k)$ is *without cycles* if there do not exist $a, b \in \{0, 1, \ldots, k\}$, $a \neq b$, such that $i_a = i_b$.

Now, we consider *MCST situations.* In an MCST situation a set $N = \{1, \ldots, n\}$ of agents is involved willing to be connected as cheap as possible to a source denoted by 0. In the sequel we use the notation S' for $S \cup \{0\}$, for each $S \subseteq N$. An MCST situation can be represented by a tuple $< N', E_{N'}, w >$, where $< N', E_{N'} >$ is the complete graph on the set N' of nodes or vertices, and $w : E_{N'} \to I\!R_+$ is a map which assigns to each edge $e \in E_{N'}$ a nonnegative number $w(e)$ representing the weight or cost of edge e. We call w a *weight function.*

The cost of a network $\Gamma \subseteq E_{N'}$ is $w(\Gamma) = \sum_{e \in \Gamma} w(e)$. A network Γ is a *spanning network* on $S' \subseteq N'$ if for every $e \in \Gamma$ we have $e \in E_{S'}$ and for every $i \in S$ there is a path in Γ from i to the source. Given a spanning network Γ on N' we define the set of edges of Γ with nodes in $S' \subseteq N'$ as the set $E^{\Gamma}_{S'} = \{\{i,j\} | \{i,j\} \in \Gamma \text{ and } i,j \in S'\}$.

For any MCST situation $w \in \mathcal{W}^{N'}$ it is possible to determine at least one *spanning tree* on N', i.e. a spanning network without cycles on N', of minimum cost; each spanning tree of minimum cost is called an MCST for N' in w or, shorter, an MCST for w. Two famous algorithms for the determination of minimum cost spanning trees are the algorithm of Prim (1957) and the algorithm of Kruskal (1956).

The characteristic function of the *minimum cost spanning tree game* (G, c_w) (or simply c_w) (Bird (1976); see also Granot and Huberman (1981), Feltkamp (1995)), corresponding to a MCST situation $< G', w >$ based on a gene expression data-set X, is defined by

$$c_w(S) = \min\{w(\Gamma) | \Gamma \text{ is a spanning network on } S'\} \tag{4.1}$$

for every $S \in 2^G \backslash \{\emptyset\}$, with the convention that $c_w(\emptyset) = 0$.

4.1.2 MCST situations based on a gene expression data-set

Consider a microarray experiment on a finite set $G' = \{1, \ldots, g\} \cup \{0\}$ of genes studied in k different samples where gene 0 is the reference gene, that is the vector of expression values of a gene that should be constantly not expressed across the different samples. Let $X = (\mathbf{x}_i)_{i \in G'}$ be a set of expression vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik}) \in I\!R^k$, representing the expression value of gene $i \in G'$ across k samples. Let $\mathbf{x}_0 = (x_{01}, \ldots, x_{0k})$ be the expression vector of the reference gene. Such a vector could be obtained averaging the expression vectors of a set of invariant genes (see Chapter 5.5 in Amaratunga and Cabrera (2004) for a definition of *invariant genes*).

We define a *MCST situation based on a gene expression data-set* X as the tuple $< G', w >$, where each edge $\{u, v\} \in E_{G'}$ has a weight that is equal to the *dissimilarity measure* $d(\mathbf{x}_u, \mathbf{x}_v)$, where $d : X \times X \to I\!\!R_+$ states quantitatively how dissimilar $\mathbf{x}_u$ and $\mathbf{x}_v$ are to each other. Note that we do not require that a dissimilarity function would be a *metric* on X. [1]

We simply require that function d satisfies at least the property of symmetry and non-negativity on X.

Consider the *microarray MCST game* (G, c_d) (or simply c_d), corresponding to a MCST situation $< G', w >$ based on a gene expression data-set X (and with weight function w corresponding to the dissimilarity measure d on X), is defined by

$$c_d(S) = \frac{\min\{w(\Gamma)|\Gamma \text{ is a spanning network on } S'\}}{\min\{w(\Gamma)|\Gamma \text{ is a spanning network on } G'\}} \tag{4.2}$$

for every $S \in 2^G \backslash \{\emptyset\}$, with the convention that $c_d(\emptyset) = 0$. We chose to divide the cost of the MCST on each coalition $S \in 2^G \backslash \{\emptyset\}$ by the cost of the MCST on the great coalition G' in order to make possible the comparison of MCST games based on gene expression data-sets obtained from experiments performed on different sets of samples. Alternatively to definition (4.2), one can obtain the MCST game c_d directly by definition (4.1) as the MCST game corresponding to the MCST situation $(G', \hat{w})$, where

$$\hat{w}(\{u, v\}) = \frac{w(\{u, v\})}{\min\{w(\Gamma)|\Gamma \text{ is a spanning network on } G'\}}.$$

In the next example we present two different approaches to the analysis of microarray MCST games, whose essential differences follows from the kind of information known on data-sets.

Example 13 *Consider the real-valued expression matrix* X *in Table 4.1, on genes* $G' = \{0, 1, 2, 3\}$*, where* 0 *is the reference gene, and all genes are collected from* 7 *samples, such that samples* $1, 2, 3, 4$ *come from the biological condition* 1 *and samples* $5, 6, 7$ *come from the biological condition* 2*. Therefore, we can split*

[1] A metric m on a set N is a function $m : N \times N \to I\!\!R_+$, such that for each $i, j \in N$ we have

$m(i, j) \geq 0$ and $m(i, i) = 0$ (non-negativity);
$m(i, j) = m(j, i)$ (symmetry);
$m(i, k) \leq m(i, j) + m(j, k)$ (triangle inequality);
if $i \neq j$ then $m(i, j) > 0$ (positivity for distinguished points);

samples: genes	1	2	3	4	5	6	7
0	0.2	0.3	-0.1	0	0.1	-0.2	0.05
1	9	-3	2	-1	-2	8	-1
2	2.9	3	2.8	3.2	3.1	2.6	3.4
3	-1	-1	-2	-4	6	4	3

Table 4.1: A toy gene expression matrix on 4 genes (0 is the invariant gene) and 7 samples; $\{1, 2, 3, 4\}$ are samples under the biological condition 1 and $\{5, 6, 7\}$ are samples under the biological condition 2

the matrix in Table 4.1 in two sub-matrix X^1 *and* X^2 *such that*

$$X^1 = \begin{pmatrix} 0.2 & 0.3 & -0.1 & 0 \\ 9 & -3 & 2 & -1 \\ 2.9 & 3 & 2.8 & 3.2 \\ -1 & -1 & -2 & -4 \end{pmatrix}, \quad X^2 = \begin{pmatrix} 0.1 & -0.2 & 0.05 \\ -2 & 8 & -1 \\ 3.1 & 2.6 & 3.4 \\ 6 & 4 & 3 \end{pmatrix}$$

In this example we use the Euclidean distance as dissimilarity measure, i.e. for each $u, v \in G'$ *we consider the metrics* e^1 *and* e^2 *on* X^1 *and* X^2*, respectively, such that for each* $t \in \{1, 2\}$

$$e^t(\{\mathbf{x}_u^t, \mathbf{x}_v^t\}) = \Big(\sum_{j \in \{1,\dots,k^t\}} (\mathbf{x}_{uj}^t - \mathbf{x}_{vj}^t)^2 \Big)^{\frac{1}{2}}, \tag{4.3}$$

where $k^1 = 4$ *and* $k^2 = 3$.

The (approximated) values of $e^1(\mathbf{x}_u^1, \mathbf{x}_v^1)$ *between each pair of genes* $u, v \in \{0, 1, 2, 3\}$ *is represented by the weights of the edges in the following completed graph.*

The microarray MCST game (G, c_{e^1})*, corresponding to the MCST situation under the biological condition 1, is such that* $c_{e^1}(\{1\}) = 0.6$*,* $c_{e^1}(\{2\}) = 0.36$*,* $c_{e^1}(\{3\}) = 0.3$*,* $c_{e^1}(\{1, 2\}) = 0.95$*,* $c_{e^1}(\{1, 3\}) = 0.9$*,* $c_{e^1}(\{2, 3\}) = 0.65$*,* $c_{e^1}(\{1, 2, 3\}) = 1$.

The (approximated) values of $e^2(\mathbf{x}_u^2, \mathbf{x}_v^2)$ *between each pair of genes* $u, v \in \{0, 1, 2, 3\}$ *is represented by the weights of the edges in the following completed graph.*

The microarray MCST game (G, c_{e^2})*, corresponding to the MCST situation under the biological condition 2, is such that* (G, c_{e^2}) *is* $c_{e^2}(\{1\}) = 0.5$*,* $c_{e^2}(\{2\}) =$

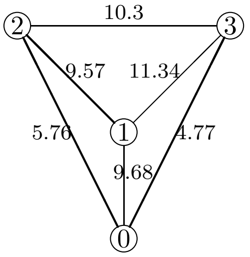

Figure 4.1: The MCST situation under the biological condition 1. The thick lines show the MCST on G'.

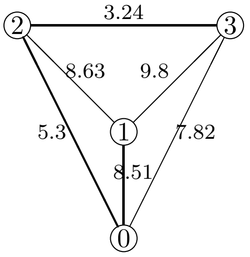

Figure 4.2: The MCST situation under the biological condition 2. The thick lines show the MCST on G'.

0.31, $c_{e^2}(\{3\}) = 0.46$, $c_{e^2}(\{1,2\}) = 0.81$, $c_{e^2}(\{1,3\}) = 0.96$, $c_{e^2}(\{2,3\}) = 0.5$, $c_{e^2}(\{1,2,3\}) = 1$.

Example 14 *Consider again the toy data-set in Table 1, but now suppose that no information about sample labels with respect to biological conditions* 1 *and* 2.

Then, using again the Euclidean distance, The (approximated) values of $e(\mathbf{x}_u, \mathbf{x}_v)$ *between each pair of genes* $u, v \in \{0,1,2,3\}$ *is represented by the weights of the edges in the following completed graph.*
The microarray MCST game (G, c_e) *is such that* (G, c_e) *is* $c_e(\{1\}) = 0.43$, $c_e(\{2\}) = 0.26$, $c_e(\{3\}) = 0.31$, $c_e(\{1,2\}) = 0.69$, $c_e(\{1,3\}) = 0.74$, $c_e(\{2,3\}) = 0.57$, $c_e(\{1,2,3\}) = 1$.

Another useful dissimilarity measure to use with microarray data (see for instance chapter 9.2.1 in Amaratunga and Cabrera (2004)) is the function

$$\rho(x_u, x_v) = 1 - |r(x_u, x_v)|, \tag{4.4}$$

where $|r(\boldsymbol{x}_u, \boldsymbol{x}_v)|$ *is absolute value of the Pearson correlation coefficient defined*

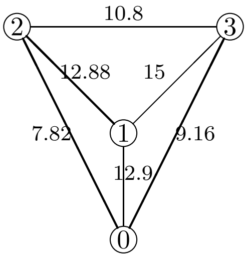

Figure 4.3: An MCST situation with three genes. The thick lines show the MCST on G'.

as

$$r(\boldsymbol{x}_u, \boldsymbol{x}_v) = \frac{\sum_{k=1,\ldots,k}(x_{uk} - \bar{x}_u)(x_{vk} - \bar{x}_v)}{\left(\sum_{j=1,\ldots,k}(x_{uj} - x_u)^2 \sum_{j=1,\ldots,k}(x_{vj} - \bar{x}_v)^2\right)^{\frac{1}{2}}}, \tag{4.5}$$

for each pair $u, v \in G'$, where $\bar{x}_u = \sum_{k=1,\ldots,k} x_{uk}$ and $\bar{x}_v = \sum_{k=1,\ldots,k} x_{vk}$. The dissimilarity measure ρ takes values in the interval $[0,1]$ and increases toward its maximum value of 1 the less linearly correlated $\boldsymbol{x}_u$ and $\boldsymbol{x}_u$ are. On the other hand, ρ is not a metric. [2]

The (approximated) values of $\rho(\mathbf{x}_u, \mathbf{x}_v)$ between each pair of genes $u, v \in \{0,1,2,3\}$ is represented by the weights of the edges in the following completed graph.

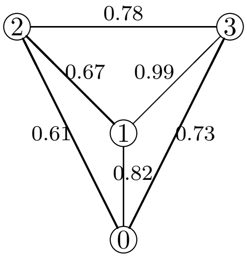

Figure 4.4: An MCST situation with three genes. The thick lines show the MCST on G'.

The corresponding microarray MCST game (G, c_ρ) is $c_\rho(\{1\}) = 0.41$, $c_\rho(\{2\}) =$

[2] Take for example vectors $a = (1,0,0), b = (0,1,0), c = (-1,1,0) \in \boldsymbol{R}^3$ and note that $\rho(a,b) = 0.5 > \rho(a,c) + \rho(c,b) = 0.2679492$, i.e. the triangle inequality is not satisfied. Of course function ρ does not satisfy neither the positivity for distinguished points condition, since for example $\rho((1,2),(2,4)) = 0$.

0.30, $c_\rho(\{3\}) = 0.36$, $c_\rho(\{1,2\}) = 0.64$, $c_\rho(\{1,3\}) = 0.77$, $c_\rho(\{2,3\}) = 0.69$, $c_\rho(\{1,2,3\}) = 1$.

Given a microarray MCST game (G, c_ρ) based on a microarray gene expression data-set, it is possible to apply a one-point solution for TU-games. For example, we could calculate the Shapley value of c_{e^1}, c_{e^2} and c_ρ.

Example 15 *The following table shows the Shapley values of games c_{e^1}, c_{e^2} and c_ρ in Example 13 and Example 17. Note that in the case where it is known the partition of samples with respect to the biological conditions, gene* 1 *has an high Shapley value in both microarray MCST games c_{e^1} and c_{e^2}, whereas gene* 2 *and* 3 *have an opposite behavior under the two conditions (when one of them has high Shapley value, the other one has a low Shapley value). If we look at the differences of Shapley values between the two conditions* 1 *and* 2*, we observe for gene* 1 *a difference close to zero, and for gene* 2 *and* 3 *differences closed to* 0.1*. In conclusion, gene* 1 *seems heavily involved in the mechanisms regulating both conditions; on the contrary, the Shapley values of genes* 2 *and* 3 *seem to differentiate more the two conditions.*

player:	1	2	3
$\phi(c_{e^1})$	0.48	0.28	0.24
$\phi(c_{e^2})$	0.5	0.18	0.32
$\phi(c_e)$	0.43	0.26	0.31
$\phi(c_\rho)$	0.36	0.27	0.37

Table 4.2: Shapley value $\phi()$ of four different microarray MCST games.

Looking at the game c_e, corresponding to the MCST situation on the whole data-set, it appears that genes 2 *and* 3 *increases their relevance with respect to gene* 1*, obtaining more or less the respective maximum between the two separated games c_{e^1} and c_{e^2}. This effect is even more evident when the dissimilarity measure ρ, which is aimed to measure linear correlation among expression vectors, is used and the corresponding microarray MCST game c_ρ is analyzed. In fact, note that gene* 3 *obtains a slightly greater Shapley value of gene* 1*. In our opinion, the correct interpretation of this fact is that on the whole data-set the Shapley value is able to provide hints about those genes which are able to discriminate well among two or more unknown different biological conditions under* *which samples are collected. In other terms, our conjecture is that the Shapley value on microarray MCST games is able to select genes which change much across the samples both in terms of expression values and in terms of similarity with the other genes.*

Of course, it is not possible to justify the use of the Shapley value as relevance index for genes on the basis of few examples. Even less correct would be to claim that the Shapley value is able to quantify the gene relevance on the class of microarray MCST games merely on the basis of the properties introduced in Chapter 2, which were defined just for another class of TU-games, the class of microarray games.

One possible way to proceed is again to propose sound properties, possibly with a biological meaning, that a relevance index applied on the class of microarray MCST games should satisfy and then to axiomatically characterize the Shapley value on the class of microarray MCST games, analogously to what we did in Chapter 2 for microarray games.

In any case, recall that to calculate the Shapley value for microarray MCST games is not easy for computational reasons. In fact, with the exception of few cases where the weight function of the MCST situation satisfies certain properties, to calculate the Shapley value of microarray MCST games with n players requires the computation of 2^{n-1} marginal contribution, which is a very big value when n takes the size of the number of genes analyzed in a conventional microarray experiments. In these cases a possible alternative which is polynomially computable is the solution called P-value by Branzei *et al.* (2004), already introduced in Feltkamp *et al.* (1994) as Equal Remaining Obligations rule and studied in Branzei *et al.* (2004), Tijs *et al.* (2004) also in connection with the Shapley value and other solutions in Tijs *et al.* (2005) and Moretti *et al.* (2005). Of course, it remains the open question on the justification of the use of the P-value as relevance index for microarray MCST games, whose answer could be provided once again by the property driven approach. In next example we show the normalized version of P-value on the four games introduced in Examples 13 and 17, without giving a formal definition of the P-value. For a formal definition of the P-value solution see for example Branzei *et al.* (2004).

Example 16 *The following table shows the P-values of games c_{e^1}, c_{e^2} and c_ρ in Example 13 and Example 17. Relevance index provided by the P-value so-*

player:	1	2	3
$P(c_{e^1})$	0.47	0.29	0.24
$P(c_{e^2})$	0.5	0.25	0.25
$P(c_e)$	0.43	0.26	0.31
$P(c_\rho)$	0.33	0.30	0.36

Table 4.3: P-value $P()$ of four different microarray MCST games.

lution seems to flatten the observed differences of Shapley values among genes calculated in Example 15.

4.1.3 Future work

Concerning the class of MCST games, there are many possible directions that can be addressed. First of all, it would be important to improve the understanding of the model potentiality, in particular with respect the "good" dissimilarity measure to be used and the meaning of the reference gene. For example, the problem of finding a good estimate of the reference gene is still an open question and, apparently, it is strongly related to the problem of data normalization in the pre-processing analysis of microarray games.

As we already stressed in the previous section, another point that should deserve more attention is the axiomatic characterization of different relevance index on the class of MCST games. The property driven approach would have a very important role in practice, for the selection of a proper relevance index. This approach would be useful also to make better interpretations of the results. For example, it is not clear at this moment why the P-value, at least on few small examples, shows such a flattening behavior as shown in Example 16.

To study the connections with the statistical method introduced in Chapter 2, it would be useful to understand whether a procedure similar to Algorithm 1 for finding significant Shapley value differences between two biological conditions can be maintained also for MCST games or, alternatively, which new assumptions should be introduced in order to achieve the same purposes.

With respect to practice, finally, the applications of the model to real data should be necessary to make a comparison with the results provided by microarray games.

4.2 Microarray games and the classification problem

As we already noted in the previous chapters, the goal of many analytical methods applied to gene expression data-sets is to develop a *classification rule*, that is a criterium to predict, as accurately as possible, the true class of samples. For example, suppose that some samples are collected from a class of tumors and some other from normal tissues, and their labels with respect to these two classes is known. The idea is to use in some way the information available either on single genes or on combination of genes as classifiers for classifying the samples into the right classes. Since most of genes contribute to add noise and to obfuscate the separation between classes, only few genes are able to perform almost correct discriminations. In order to select the set of genes with the best performance in classifying samples, the analytical method must solve a very hard problem: maximize the proportion of correct classification and minimize the number of misclassification in the data-set under consideration.

In this section we try to answer to the following interesting question: is the Shapley value of a microarray game of any help in studying the ability of genes in well classifying tumor samples according to a certain classification rule?

First note that the Shapley value of a microarray game seems meaningless as classification power index, since a microarray game does not consider any classification information in its characteristics function. On the other hand, intuitions based on the results provided by the application of the Shapley value on microarray games seem to go in the direction of a quite positive answer.

In the attempt to yield an analytical explanation of this fact, first of all we need to introduce a very simple classification rule based on boolean data.

For simplicity, we will consider only two classes, let us say the biological conditions 1 and 2. Moreover, we will refer to a dichotomized gene expression data-set. Let $\mathbf{B}^1 \in \{0,1\}^{n \times k_1}$ and $\mathbf{B}^2 \in \{0,1\}^{n \times k_2}$ be two boolean matrix, where n is the number of genes, k_1 is the number of samples under the biological condition 1 (for example, samples from normal tissues), k_2 is the number of samples under the biological condition 2 (for example, samples from tumoral tissues) and if $\mathbf{B}^t_{ij} = 1$ for some $i \in \{1, \dots, n\} = N$, $j \in \{1, \dots, k_t\}$ and $t \in \{1, 2\}$, then it means that gene i in sample j and condition t shows a certain expression property (for example, it is over-expressed) and if $\mathbf{B}^t_{ij} = 0$ then it means that gene i in sample j and condition t does not show such expression property.

Consider the following classification rule based on the subset of genes in $S \in 2^N \setminus \{\emptyset\}$:

$<$ **if** there exists $i \in S$ such that $\mathbf{B}^t_{ij} = 1$, **then** classify sample $j \in \{1, \ldots, k_t\}$ under the biological condition $t \in \{1, 2\}$ **in class 1** $>$. ($\bigstar$)

Let $r_c(S)$ be the rate of correct classifications provided by the classification rule ($\bigstar$) applied on the data-set $(\mathbf{B}^t)_{t\in\{1,2\}}$ using the set of genes $S \in 2^N \setminus \{\emptyset\}$ and let $r_m(S)$ be the rate of misclassifications made via the classification rule ($\bigstar$) applied on the data-set $(\mathbf{B}^t)_{t\in\{1,2\}}$ using the set of genes S. Then we define the *classification game* (N, d_1) as the TU-game on N with the characteristic function $d : 2^N \to [-1, 1]$ such that

$$d_1(S) = r_c(S) - r_m(S), \tag{4.6}$$

for each $S \in 2^N \setminus \{\emptyset\}$.

Remark 8 We can provide an alternative definition of classification game using dual unanimity games, which will be useful later. Recall that a dual unanimity game (N, u^*_T), $T \in 2^N$, is a TU-game described by $u_T(S) = 1$ if $R \cap T \neq \emptyset$ and $u_R(T) = 0$, otherwise.

Relation (4.6) can straightforwardly be rewritten as follows

$$d_1(S) = \frac{|\Omega(S)|}{k_1} - \frac{|\Delta(S)|}{k_2} \tag{4.7}$$

where $|\Omega(T)|$ is the cardinality of the set

$$\Omega(S) = \{j \in \{1, \ldots, k_1\} | sp(\mathbf{B}^1_j) \cap S \neq \emptyset\}$$

and $\bar{v}(\emptyset) = 0$ and $|\Delta(S)|$ is the cardinality of the set

$$\Delta(S) = \{j \in \{1, \ldots, k_2\} | sp(\mathbf{B}^2_j) \cap S \neq \emptyset\}$$

and $d_1(\emptyset) = 0$. Equivalently, the game (N, d_1) can be represented via the relation

$$d_1(S) = \sum_{j=1,\ldots,k_1} \frac{u^*_{sp(\mathbf{B}^1_j)}(S)}{k_1} - \sum_{j=1,\ldots,k_2} \frac{u^*_{sp(\mathbf{B}^2_j)}(S)}{k_2} \tag{4.8}$$

for each $S \in 2^N \setminus \emptyset$, where $(N, u^*_{sp(\mathbf{B}_j)})$ is the *dual unanimity game* on the set $sp(\mathbf{B}^t_j)$, $t \in \{1, 2\}$.

Remark 9 One could use the classification rule:

< **if** there exists $i \in S$ such that $\mathbf{B}^t_{ij} = 1$, **then** classify sample $j \in \{1, \ldots, k_t\}$ under the biological condition $t \in \{1, 2\}$ **in class 2** >, $(\bullet)$

for each $S \in 2^N \setminus \emptyset$, and define the classification game (N, d_2) according to relation (4.6) or (4.7). Note that $d_2(S) = -d_1(S)$ for each coalition $S \in 2^N \setminus \emptyset$.

A well known results for TU-games is that the Shapley value of a unanimity game on $T \in 2^N \setminus \{\emptyset\}$ is equal to the Shapley value of the dual unanimity game on T (see for example Tijs *et al.* (2003)). In formula,

$$\phi_i(u_T) = \phi_i(u^*_T) = \frac{1}{|T|}, \tag{4.9}$$

for each $i \in N$, where $\phi(u_T)$ is the Shapley value on the unanimity game u_T and $\phi(u^*_T)$ is the Shapley value on the dual unanimity game u^*_T, for each $T \in 2^N \setminus \{\emptyset\}$. Relation (4.9) makes Proposition 3 straightforward.

Proposition 3 *Let* $\mathbf{B}^1 \in \{0,1\}^{n \times k_1}$ *and* $\mathbf{B}^2 \in \{0,1\}^{n \times k_2}$ *be two boolean matrix. Let* (N, v_1) *be the microarray game corresponding to* $\mathbf{B}^1$ *and Let* (N, v_2) *be the microarray game corresponding to* $\mathbf{B}^2$*. Moreover, Let* (N, d_1) *be the classification game corresponding to* $\mathbf{B}^t$*,* $t \in \{1, 2\}$ *and to classification rule* ($\bigstar$)*. Then*

$$\phi(d_1) = \phi(v_1) - \phi(v_2). \tag{4.10}$$

Proof It follows directly by relations (1.3), (4.9), (3.2) and (4.8). ■

By Remark 9 follows that the absolute value $|\phi_i(d_1)| = |\phi_i(v_1) - \phi_i(v_2)|$ provides an indication of the ability of gene i in discriminating the two classes 1 and 2, for each $i \in N$. Note that, in order to test the null hypothesis that $|\phi_i(v_1) - \phi_i(v_2)| = 0$, a bootstrap procedure was introduced in Chapter 2. Note also that, if it exists $i \in N$ such that the coalition S is the unique coalition which classifies correctly all the samples under condition 1 and does not misclassifies any samples under condition 2, then $\phi_i(d_1) = max_{i \in N} \phi_i(d_1)$ for each $i \in S$.

Example 17 Consider the following boolean matrix

$$\mathbf{B}^1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{B}^2 = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

Then $sp(\mathbf{B}^1_1) = \{1,3\}$, $sp(\mathbf{B}^1_2) = \{2,3\}$, $sp(\mathbf{B}^1_3) = \{1,2,4\}$ and $sp(\mathbf{B}^2_1) = \{1,2\}$, $sp(\mathbf{B}^2_2) = \{1,4\}$, $sp(\mathbf{B}^2_3) = \{3\}$ and $sp(\mathbf{B}^2_4) = \{1,4\}$.

The classification game $(\{1,2,3,4\}, d_1)$ corresponding to $\mathbf{B}^t$, $t \in \{1,2\}$ and to classification rule ($\bigstar$) is such that $d_1(\{2\}) = d_1(\{3\}) = \frac{2}{3} - \frac{1}{4} = \frac{5}{12}$; $d_1(\{4\}) = \frac{1}{3} - \frac{2}{4} = -\frac{1}{6}$; $d_1(\{1\}) = d_1(\{1,4\}) = d_1(\{2,4\}) = \frac{2}{3} - \frac{3}{4} = -\frac{1}{12}$; $d_1(\{1,2\}) = d_1(\{3,4\}) = d_1(\{1,3\}) = 1 - \frac{3}{4} = 1 - \frac{1}{4}$; $d_1(\{2,3\}) = 1 - \frac{2}{4} = \frac{1}{2}$; $d_1(\{1,2,4\}) = 1 - \frac{3}{4} = \frac{1}{4}$; $d_1(\{1,3,4\}) = d_1(\{2,3,4\}) = d_1(\{1,2,3\}) = d_1(\{1,2,3,4\}) = 0$.
The Shapley value of the classification game $(\{1,2,3,4\}, d_1)$ is $(-\frac{7}{72}, \frac{11}{72}, \frac{6}{72}, -\frac{10}{72})$. The maximum classification power is allocated to gene 3 and 2.

4.2.1 Future work

Classification games provide a different interpretation of microarray games, in terms of the classification information contained in the characteristics function. Of course all these argumentations are coherent with the classification rule ($\bigstar$) (or, alternatively, with the classification rule ($\bullet$)). These classification rules are simple methods to classify samples (even if to evaluate the performance of the classification rule ($\bigstar$) on each coalition is not at all simple when the number of genes is high; once again this problem can be partially avoided during the Shapley value computations thanks to Proposition 3 and the considerations made in Chapter 2 on microarray games). It would be very interesting to study classification games corresponding to more complex classification rules, for example based on Support Vector Machines (Cortes and Vapnik (1995)) and other supervised classification technique applied to each coalition of genes. In this context, the Shapley value of classification games based on different classification rules could be also informative in comparing the validity of different classifiers.

4.3 Analysis of gene expression data from Real Time PCR

In this section we present a brief description of very preliminary results provided by a gene expression analysis concerning 19 samples collected from neuroblastic tumors, in particular 10 neuroblastoma (Schwannian stroma-poor) (NB-SP) samples and 9 Ganglioneuroblastoma intermixed (Schwannian stroma-rich) (GNBi-SR). The goal of this analysis was to screen and to give a priority level to genes according to their connections in differentiating and provoking NB-SP and GNBi-SR tumors.

The analysis is divided in two parts. In the first one, the expression values of 22283 genomic sequences from the 19 samples have been assessed using the Human Genome U133A GeneChip microarray technology (Affymetrix, Inc, CA, USA).

The second part of the analysis is based on the same samples, but the data collection has not yet concluded at this moment. In this case, the data collection of only 126 genes is done by means of quantitative *Real Time PCR* (Polymerase Chain Reactions), a very accurate (and much more expensive with respect to GeneChip Affymetrix microarray technology) method to evaluate gene expressions on a gene-by-gene basis. Real Time PCR is commonly used in biomedical literature as a confirmation of microarray results (see for example Amaratunga *et al.* (2004)).

In both the analysis, all data have been collected by the Unit of Translational Paediatric Oncology of the National Institute for Cancer Research (IST) and the Laboratory of Italian Neuroblastoma Foundation of the Advanced Biotechnology Center (ABC), both located in Genoa (Italy). For more details on the connections among the neuroblastic tumors under studying and for a statistical pre-analysis on a smaller data-set see our previous work in Coco *et al.* (2005).

From the first analysis, concerning microarray gene expression data, the raw data have been pre-processed using the function *expresso* in the package *affy* in Bioconductor libraries (Gentleman *et al.* (2004)), normalizing using the Variance Stabilization Normalization (vsn) method (Huber *et al.* (2002)).

On the microarray normalized data-set we performed two different procedures. The first procedure has been based on the application of the Statistical Analysis of Microarray (SAM) (Storey and Tibshirani (2003)), in order to find genes which are significantly differential expressed between the two conditions NB-SP and GNBi-SR. Controlling the FDR at a level approximately equal to zero (0.000005), the SAM method called 88 genes as significant.

The second procedure has been based on the game theoretical approach described in Chapter 2. For each one of the two conditions NB-SP and GNBi-SR, a microarray game has been constructed and the corresponding Shapley value has been computed on it. Then, for each gene, the absolute value of the observed difference of Shapley values of the two microarray games has been calculated. We selected the first 49 genes with the highest absolute value of Shapley difference. The overlap of genes selected by the two methods (SAM and microarray game) was of 11 genes. Note that at the time when this analysis has been performed, Algorithm 1 described in Chapter 2 was not yet completed. Here the number 49 reflects the constraint, due to practical reasons, implying that only 126 genes could be validated in the following analysis based on Real Time PCR.

4.3.1 Future work

As we already said, a Real Time PCR analysis has been planned to be used to confirm the results obtained by the two statistical procedure in the first part of the analysis, concerning microarray data.

Here we briefly anticipate the results of the validation procedure, at this moment performed on only 15 of the 19 samples (the remaining 4 samples are currently under experiment in the laboratories).

For what concerns the set of genes called significant by the SAM method, up to now we can only report that the agreement rate between the Real Time PCR technology and the GeneChip Affymetrix technology in indicating the condition (GNBi-SR or NB-SP) where the average expression value of the genes is greater is about 98%.

For what concerns the set of genes selected applying microarray games, a similarly high rate of agreement has been observed. On the other hand, we want to point out that in this case genes have been selected according to their Shapley values on microarray games, and that the Shapley value of a microarray game keeps into account both the expression values of genes and their cooperative interactions. For these reasons in our opinion the more accurate information about the agreement in average expression values provided by Real Time PCR is not enough to confirm the results obtained by means of the game theoretical method. In this direction, we find reasonable to apply Algorithm 1 to the dataset of 49 genes analyzed by Real Time PCR and to observe how many genes effectively show their Shapley values significantly different between the two conditions. A first attempt of application of Algorithm 1 to the Real Time PCR expression data-set concerning those 49 genes, suggests the rejection of 46 null hypothesis of no differences in terms of Shapley values when controlling the FDR at a level of 0.05. In our opinion, this is a quite positive response in the direction of the validation of the microarray game analysis.

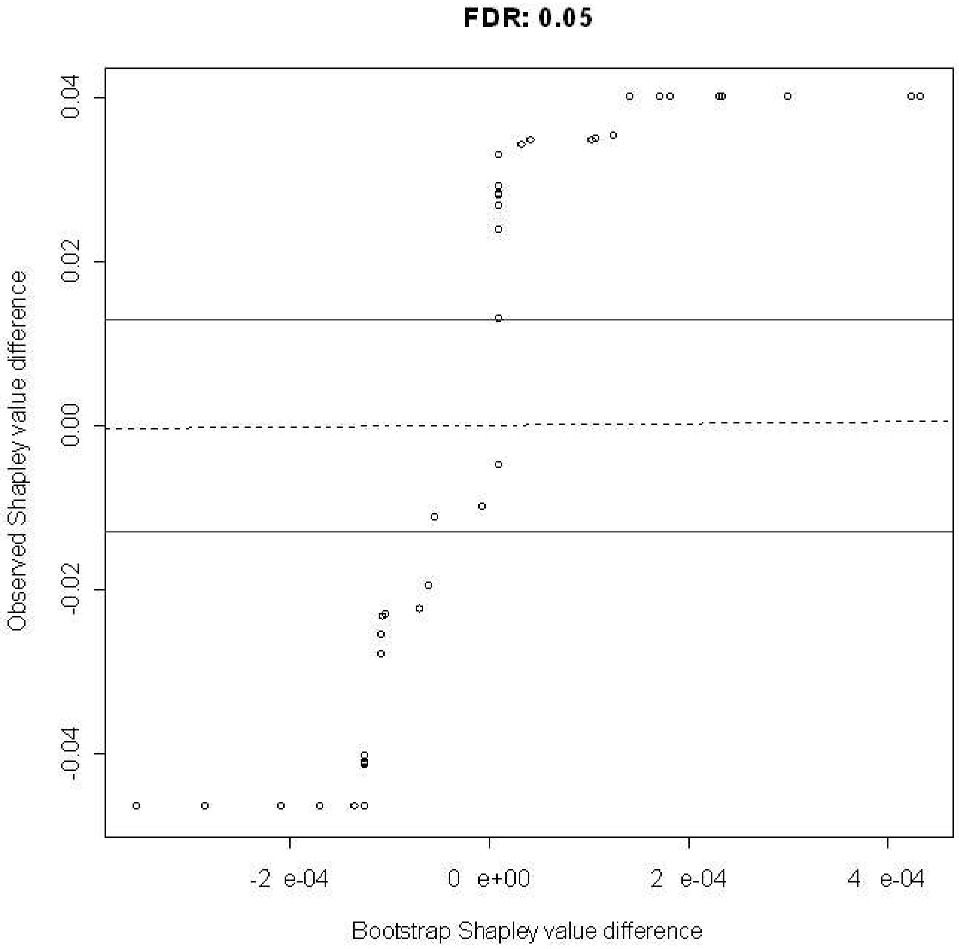

Figure 4.5: QQ plot of the observed Shapley value differences and the expected Shapley value differences produced by Algorithm 1. Application of Algorithm 1 to the Real Time PCR expression data-set concerning 49 selected genes suggests the rejection of 46 null hypothesis. Null hypothesis of genes with Shapley values difference observed on Real Time PCR data which are outside the interval between the two continuous horizontal straight lines have been rejected controlling the FDR at the level 0.05.

Appendix A

Dichotomization algorithm

In this appendix we face the problem of transforming real-valued gene expressions into binary values. This problem can be solved using a procedure which selects thresholds. In Shmulevich and Zhang (2002) an algorithm to find an individually selected threshold for each normalized gene expression vector is presented. For each gene expression vector, the basic idea of such an algorithm is to sort all the real-valued expression values and to locate the threshold in correspondence of the smallest separation between two successive sorted values which is greater than a predefined value, called the 'big jump'. In their algorithm implementation, the author used as predefined value in given a gene expression vector the length of the interval between to successive sorted values in the 'the worst case', that is when all the values in the gene expression vector are equally spaced between the maximum and the minimum.

Since the case in which the sorted true values are equally spaced between the maximum and the minimum is in fact the most critical one to be binarized, we think that the 'big jump' used by Shmulevich and Zhang (2002) is very sensitive to small fluctuations on the observed values due to random noise. Therefore, we present a different version of the algorithm, where the 'big jump' is calculated via an iterative procedure. As formally explained in the following pseudo-code, at each iteration, the algorithm provide a candidate value directly proportional to difference between the maximum and the minimum and inversely proportional to the number of iterations already done. Our algorithm fixes the 'big jump' as the biggest 'candidate value' smaller than some separations between two successive sorted values.

Algorithm 2 (Dichotomize) *INPUT: a real-valued expression matrix* $\mathbf{G} \in I\!R^{n\times k}$*, with* n *rows (i.e. genes) and* k *columns (i.e. samples); a low-bound outlier parameter* d*.*
OUTPUT: a boolean matrix $\mathbf{B} \in \{0,1\}^{n\times k}$*,* $n, k \in \{1, 2, \ldots\}$*.*

```
step 1 : $\mathbf{S}_i \leftarrow sort(G_{i,1}, \ldots, G_{i,k}),\ \forall i \in N;$

step 2 : for $j : 1$ to $k - 1$ do
        $D_{i,j} \leftarrow sort(S_{i,j+1} - S_{i,j}),\ \forall i \in N;$
        end do

step 3 : for $l : 1 + d$ to $k$ do
        $t_{i,l} = \frac{S_{i,k} - S_{i,1}}{l-1},\ \forall i \in N;$
        end do

step 4 $t_i^* = \max_{l\in\{1+d,\ldots,k\}}\{t_{i,l} : \exists j \in \{1,\ldots,k\} s.t. \in D_{i,j} > t_{i,l}\}\ \forall i \in N;$

step 5 $m_i = \min\{j \in \{1,\ldots,k\} : D_{i,j} > t_i^*\},\ \forall i \in N;$

step 3 : for $j : 1$ to $k$ do
        if $G_{i,j} \geq S_{i,m_i+1}$ then
        $\mathbf{B}_{i,j} \leftarrow 1;$
        else
        $\mathbf{B}_{i,j} \leftarrow 0;$
        end if
        end do
```

In Figure 3.6 the threshold selected by Algorithm 2 on the expressions of gene 1 in Example 10 is shown, together with the threshold selected via the application of the algorithm in Shmulevich and Zhang (2002). Note that the 'big jump' selected by Algorithm 2 is four times the 'big jump' selected by the algorithm in Shmulevich and Zhang (2002). The parameter d has been settled to 1 (no low-bound outliers have been detected).

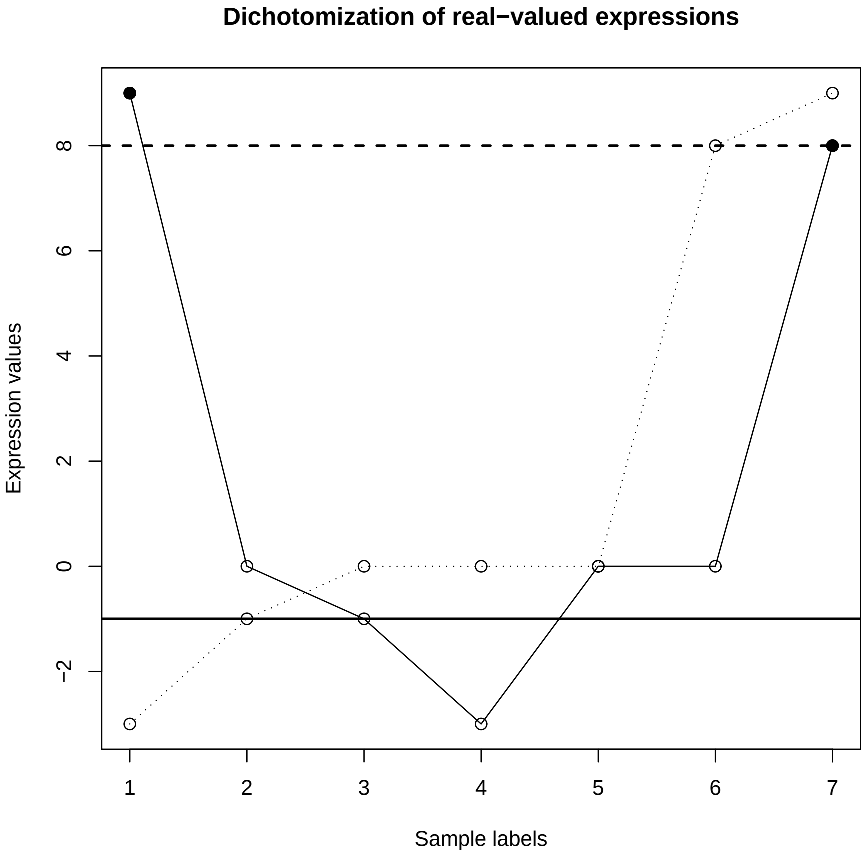

Figure A.1: Real-valued expressions of gene 1 in Example 10 corresponding to column index $1, \ldots, 7$ of Table 3.1 (continuous line) and sorted by value (dotted line). The dashed horizontal straight line indicates the dichotomization threshold selected by Algorithm 2 and the filled (unfilled) points correspond to the samples where gene 1 is labelled with value 1 (0), according to such a threshold. The continuous horizontal straight line indicates the dichotomization threshold selected by the algorithm introduced in Shmulevich and Zhang (2002).

References

Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745-6750.

Amaratunga D., Cabrera J. (2004). Exploration and Analysis of DNA Microarray and Protein Array Data, Wiley-Interscience, New Jersey.

Banzhaf J. F. III (1965). Weighted voting doesn't work: A game theoretic approach, Rutgers Law Review, 19, 317- 343.

Becquet C., Blachon S., Jeudy B., Boulicaut J.G. and O. Gandrillon (2002). Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human SAGE data, Genome Biology, 3(12), research 67.10067.16.

Benjamini Y., Hochberg Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57:289-300.

Bickel, D. R. (2002). Microarray gene expression analysis:Data transformation and multiple comparison bootstrapping, Computing Science and Statistics 34, 383-400, Interface Foundation of North America (Proceedings of the 34th Symposium on the Interface, Montreal, Quebec, Canada, April 17-20, 2002)

Bird C.G. (1976). On cost allocation for a spanning tree: a game theoretic approach. Networks 6:335-350.

Branzei, R., Moretti, S., Norde, H., Tijs, S., (2004). The *P*-value for cost sharing in minimum cost spanning tree situations, Theory and Decision 56, 47-61.

Coco S., Defferrari R., Scaruffi P., Cavazzana A., Longo L., Mazzocco K., Perri P., Gambini C., Moretti S., Bonassi S., Tonini G.P., (2005). Genome analysis and gene expression profiling of neuroblastoma (Schwannian stroma-poor) and ganglioneuroblastoma intermixed (Schwannian stroma-rich) cell subpopulations reveal a different degree of DNA imbalances and signature between Neuroblasts and Schwannian stromal cells, Journal of Pathology, 207: 346-357.

Cortes C., Vapnik V. (1995). Support vector networks, Machine Learning, 20, 1-25.

Dawson M.J., Trapani J.A. (1995). IFI 16 gene encodes a nuclear protein whose expression is induced by interferons in human myeloid leukaemia cell lines. Journal of Cellular Biochemistry, 57(1), 39-51.

Dudoit S., J. Fridlyand (2003). Classification in microarray experiments. In T. P. Speed (ed), Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall/CRC, Chapter 3, p. 93-158.

Dudoit S., Fridlyand J., Speed T.P. (2002a). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, Vol. 97, No. 457, p. 7787.

Dudoit S., Shaffer J.P., J.C. Boldrick (2003). Multiple hypothesis testing in microarray experiments, Statistical Science, 18(1), 71-103.

Dudoit S., Yang Y.H., Luu P., Speed T.P. (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), Microarrays: Optical Technologies and Informatics, Vol. 4266 of Proceedings of SPIE, p. 141-152.

Dudoit S., Yang Y., Speed T., Callow M. (2002b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, 12:111-139.

Efron B. (1979). Computers and the Theory of Statistics: Thinking the Unthinkable. SIAM Review, 21(4), 460-480.

Efron B., Gong G.(1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. American Statistician, 37, 36-48.

Efron B., Tibshirani R. J. (1993). An Introduction to the Bootstrap, Chapman & Hall/CRC: New York.

Elowitz M.B., Levine A.J., Siggia E.D., Swain P.S. (2002). Stochastic Gene Expression in a Single Cell, Science, 297, 11831186.

Felsenthal D., Machover M. (1998). The Measurement of Voting Power. Theory and Practice, Problems and Paradoxes. Cheltenham: Edward Elgar.

Feltkamp V. (1995). Cooperation in controlled network structures, PhD Dissertation, Tilburg University, The Netherlands.

Feltkamp V., Tijs S., Muto S. (1994). On the irreducible core and the equal remaining obligations rule of minimum cost spanning extension problems, CentER DP 1994 nr.106, Tilburg University, The Netherlands.

Fujarewicz K., Wiench M. (2003). Selecting differentially expressed genes for colon tumor classification, Int. J. Appl. Math. Comput. Sci., 13(3), 327335

Gentleman R.C., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J, Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F, Li C., Maechler M., Rossini A.J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J.Y.H., Zhang J. (2004). Bioconductor: Open software development for computational biology and bioinformatics, Genome Biology, 5, 2004, pp.80, http://genomebiology.com/2004/5/10/R80

Golub T., Slonim D., Tamayo P., Huard C., Gaasenbeek M., Mesirov J., Coller H., Loh M., Downing J., Caligiuri M., Bloomfield C., Lander E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286, 531-537.

Granot, D., Huberman, G. (1981). On minimum cost spanning tree games, Mathematical Programming, 21, 1-18.

Grabisch M., Roubens M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. International Journal of Game Theory, 28, 547 565.

Huber W., Heydebreck A., Sueltmann H., Poustka A., Vingron M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics, 18(Suppl.1), S96:S104.

Jain N., Cho HJ., O'Connell M., Lee J.K. (2005) Rank-Invariant Resampling Based Estimation of False Discovery Rate for Analysis of Small Sample Microarray Data. BMC Bioinformatics, 6, 187:195.

Kalai E., Samet D. (1988). Weighted Shapley Values. In The Shapley Value, Essays in Honor of Lloyd S. Shapley, A. Roth (ed.), Cambridge University Press, 83-100.

Kaufman A., Kupiec M., Ruppin E. (2004). Multi-Knockout Genetic Network Analysis: The Rad6 Example, Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04), August 16 - 19, 2004, Standford, California

Keinan A, Sandbank B., Hilgetag C.C., Meilijson I., Ruppin E. (2004). Fair attribution of functional contribution in artificial and biological networks. Neural Computation, 16(9).

Kreps D.M. (1990) Game Theory and Economic Modelling, Oxford University

Press.

Kruskal J.B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. Proceedings of the American Mathematical Society 7:48-50.

Moler E.J., Chow M.L, Mian I.S. (2000). Analysis of molecular profile data using generative and discriminative methods, 4, 109-126.

Moretti S. (2006). Statistical analysis of the Shapley value for microarray games, working paper presented at the 36th Annual Conference of the Italian Operations Research Society, Camerino, September 6-9, 2005.

Moretti S., Patrone F., Bonassi S. (2004). The class of Microarray games and the relevance index for genes, presented at the VI Spanish Meeting on Game Theory and Practice, July 12-14, 2004, Elche, Spain (Preprint of the Department of Mathematics n.531).

Moretti S., Tijs S., Branzei R., Norde H. (2005). Cost monotonic 'construct and charge' rules for connection situations. CentER for Economic Research Discussion Papers, n. 104-2005, Tilburg University, The Netherlands

Nadon R., Shoemaker J. (2002). Statistical issues with microarrays: processing and analysis. TRENDS in Genetics, 8(5):265-271.

Narducci M.G., Fiorenza M.T., Kang S.M., Bevilacqua A., Di Giacomo M., Remotti D., Picchio M.C., Fidanza V., Cooper M.D., Croce C.M., Mangia F., Russo G. (2002). TCL1 participates in early embryonic development and is overexpressed in human seminomas, PNAS, 99(18): 11712-11717.

Osborne M.J., Rubinstein A.(1994). A Course in Game Theory, MIT Press, Cambridge (Massachusetts).

Owen G. (1995). Game Theory, Academic Press - Third edition.

Parmigiani G., Garret E.S., Irizarry R.A., Scott S.L: (2003). The analysis of gene expression data: an overview of Methods and Software. In: The analysis of gene expression data: methods and software. Edited by G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger. Springer, New York.

Parmigiani G., Garret E.S., Irizarry R.A., Zeger S.L. (2003). The analysis of gene expression data: an overview of Methods and Software. Springer, New York.

Pensa R.G., Leschi C., Besson J., Boulicaut J.F. (2004). Assessment of discretization techniques for relevant pattern discovery from gene expression data, Proceedings of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics BIOKDD'04, Seattle, USA, August 2004, pp. 24-30.

Polkowski L., Araszkiewicz B. (2003). A Rough Set Approach to Estimating the Game Value and the Shapley Value from Data. Electr. Notes Theor. Comput. Sci. 82(4)

Prim R.C. (1957). Shortest connection networks and some generalizations. Bell Systems Technical Journal 36:1389-1401.

R Development Core Team (2004). R: A language and environment for statistical, R Foundation for Statistical Computing, Vienna, Austria, 2004, ISBN 3-900051-00-3, http://www.R-project.org

Roth A. E. (editor) (1988). The Shapley Value: Essays in Honor of Lloyd S. Shapley, Cambridge University Press.

Schena M. (2003). Microarray Analysis, J.Wiley & Sons Publishing, 630 pp.

Shapley L. S. (1953). A Value for n-Person Games, in Contributions to the Theory of Games II (Annals of Mathematics Studies 28), H. W. Kuhn and A. W. Tucker (eds.), Princeton University Press, 307-317.

Shapley L. S., Shubik M. (1954).A Method for Evaluating the Distribution of

Power in a Committee System, American Political Science Review 48, 787-792.

Shmulevich I., Zhang W. (2002). Binary analysis and optimization-based normalization of gene expression data, Bioinformatics, 18(4):555-565. Smith K. and T. Speed (2003). Normalization of cDNA microarray data, Methods, 31, 265-273.

Speer N., Spieth C., Merz P., Zell A. (2003). Clustering Gene Expression Data with Memetic Algorithms based on Minimum Spanning Trees", In Proceedings of the 2003 Congress on Evolutionary Computation, IEEE Press, 3,1848-1855.

Stöltzner M. (2004). On Optimism and Opportunism in Applied Mathematics, (Mark Wilson Meets John von Neumann on Mathematical Ontology), Erkenntnis 60, pp. 121-145. Available at
http://philsci-archive.pitt.edu/archive/00001225

Storey J.D., Tibshirani R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: The analysis of gene expression data: methods and software. Edited by G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger. Springer, New York.

Su Y., Murali T.M., Pavlovic V., Schaffer M., Kasif S.(2003). RankGene: identification of diagnostic genes based on expression data, Bioinformatics, 19(12), 1578-1579

Swain P.S., Elowitz M.B., Siggia E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. PNAS, 99(20):1279512800.

Tan A.C., Naiman D.Q., Xu Lei, Raimond L., (2005). Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics, 21(20), 38963904.

Tijs S. (2003). Introduction to Game Theory. Hindustan Book Agency ed.

Tijs S., Branzei R., Moretti S., Norde H. (2004). Obligation rules for minimum cost spanning tree situations and their monotonicity properties, CentER DP 2004-53, Tilburg University, The Netherlands (to appear in European Journal of Operational Research).

Tijs S., Moretti S., Branzei R., Norde H. (2005). The Bird core for minimum cost spanning tree problems revisited: monotonicity and additivity aspects. Recent Advances in Optimization, Lectures Notes in Economics and Mathematical Systems, Springer-Verlag ed., 563: 305-322

von Neumann J., Morgenstern O. (1944). Theory of Games and Economic Behavior, Princeton University Press, Princeton. Xu Y, Olman V, Xu D. (2001). Minimum spanning trees for gene expression data clustering. Genome Inform Ser Workshop Genome Inform, 12, 24-33.

Young H.P. (1995). Equity: In Theory and Practice. Princeton University Press, 253 pp.