Game Theory and Microarray Data Analysis

Paola Radrizzani

Advisor: **Prof. Roberto Lucchetti**

Tutor: Prof. Roberto Lucchetti

Acknowledgements

When three years ago I started this PhD adventure I could not realize that I would be involved in a so exciting and fascinating experience. I got my degree in mathematics several years ago and after so long time spent as a teacher of mathematics and physics in high school, I certainly did not imagine I would be able to carry out a research work in an area completely new for me as game theory. Here I would like to to thank all people who supported me in this research adventure. First of all I want to thank my thesis advisor Roberto Lucchetti for his acute, precise and precious supervision and especially for his enthusiastic and driving approach to the main issues I had to face in my thesis work. I am also grateful to my coauthors Stefano Moretti and Fioravante Patrone for their precious collaboration.

My thanks to my friend Carlo Angaroni for his support in the management of data base, to Emanuela Scacheri who introduced me to the microarray analysis explaining to me the whole process from the extraction of the DNA fragments to the final gene expression and to Luca Candiani and Stefano Bosia for their contribution in building the software to process the data. I am grateful to my parents for their warm encouragement and to my husband Paolo for his scientific support for molecular biology and for his constant encouragement especially in the most difficult moments. And thanks also to my sons, Tommaso and Andrea, who sometimes had to deal with a tired and a little nervous mom; thanks to them for their patience.

Chapter 1

Introduction

This thesis deals with an application of cooperative game theory to molecular biology. Nowadays, the discovery of the structure of the DNA allows studying those diseases which are of genetic origin, in order to try to find the genes potentially responsible of the disease itself. One of most powerful techniques is based on microarray data analysis. This technique provides expression of the genes of the individuals. Then, the medical literature usually compares the data obtained by two different groups of people. One group could be of individuals which are sane, and the other one of individuals which have a specific disease. But this is not the only case: it is also possible to compare two group of people having a similar, but not identical disease (for instance two similar types of tumour). By comparing the two sets of data, it can be seen when in the people affected by a specific disease the genes are normally expressed or else abnormally expressed (either under or over expressed).
But at this point the problem becomes how to handle the data. One of the main points is that usually the genes observed in these processes are very many. An estimation of the number of genes in the human genome is around 30,000. It thus becomes necessary to develop tools in order to give a meaning to the collected data. Of course, many statistical tools have been developed to tackle the problem.

Recently, in the literature a different approach was proposed, based on cooperative game theory (see [Moretti et al. (2007)]). The idea is to build, from the data obtained by microarray technique, a suitable TU cooperative game, where the players are the genes, and to use a power index to evaluate the strength of each player (gene). If the model is correct, then it is reasonable to expect that the genes ranked at the first places are more responsible than others in the rise of the disease. It is worth mentioning that a very recent paper based on this type of techniques (and other ones) has been published on the journal *Cancer*, showing that the interest in this model goes beyond the purely mathematical aspects.

I will now briefly describe the contents of the various chapters.
The first chapter contains the background needed to enter in the subject. I provide a short review of the main issues from molecular biology, trying to give a simple idea of the microarray technique. Then, I briefly remind all concepts from game theory needed for dealing with the model subsequently developed.
In the second chapter some axiomatic characterization of two relevant indices in the class of the microarray games is derived. The so called axiomatic approach is quite used in cooperative game theory. It means the following. As it is well known, there are some general ideas around the idea of solution for a cooperative game, but there is no a single, unifying concept of solution, like for instance the idea of Nash-Cournot equilibrium in noncooperative game theory. Of course, having several solutions (providing quite often rather different answers to the same problem) can be a little confusing. A way to better understand the underlying deep meaning of a solution concept, is to characterize the solution as the *unique* one, on a specific class of games, fulfilling a (short and reasonable) list of properties. This approach goes back at least to the pioneering papers by Nash and Shapley, one characterizing a solution for the bargaining problem, the other one to characterize one of the stars of this thesis, i.e. the Shapley value. This approach is interesting since it allows comparisons among different solutions, by looking at the different properties they fulfill. Thus in this chapter I consider the class of the microarray games, in order to provide characterizations of the Shapley and Banzhaf values. More precisely, starting from an older characterization of the Shapley value (see [Moretti et al. (2007)]), a similar characterization for the Banzhaf index is offered, and alternative characterizations for both are also derived. We finally compare the two indices on a set of data taken from the medical literature.
The third chapter provides the definition of a new family of indices, and studies them from various points of view. The starting idea is the remark that the Banzhaf and Shapley indices give a different weight to players in the so called winning coalitions. Essentially, Shapley assigns to players of the winning coalition S the power $\frac{1}{|S|}$, while Banzhaf assigns to players of the winning coalition S the power $\frac{1}{2^{|S|-1}}$[1]. This causes different results in ranking the genes. To have a better insight into the problem, it then makes sense to consider new indices, in a sense intermediate between the two main ones. In few words, the idea is to assign to the players belonging to the winning coalition S the quantity $\frac{1}{|S|^a}$, for some natural number a. For $a = 1$ we have Shapley's index and as a grows we can think to approach the Banzhaf value. Thus, I study this family of new indices. I give a formula for them for general games, which at least in the case $a = 2$ is manageable. I also provide a list of properties of the indices (leaving the attempt of finding an

[1]Given a finite set S I denote by $|S|$ the number of its elements

appropriate axiomatization to subsequent work) and I provide, for the case $a = 2$ a fast way to calculate the index for weighted majority games, in the spirit of algorithms available in the literature for the two main ones. A classical application to the study of the power of the nations in the EU council shows the fact that the new 2-index provides an intermediate result with respect the Shapley and Banzhaf in the following sense: the ratio between the power of the weakest nation and the strongest one is approximatively of $\frac{1}{100}$ for the Shapley value, of $\frac{1}{10}$ for the Banzhaf's value. Our index gives an intermediate result between them, and in this sense we believe it is more "reasonable" than both the other ones. Applications of the new index to microarray games are mostly performed in the last chapter.
The final chapter of the thesis deals with the introduction of variants of the microarray game, in order to better differentiate the ranking between the genes. In some experimental data it turns out that hundred of genes are given the same power. This can be annoying, especially for the first ranked genes: it is clear that having the first 100 genes, so to say, well separated as far as their power index is concerned, can be of great interest. Thus, by considering the proposed variant, we are led to consider weighted power indices, already introduced in the literature, that are shown to do the job. Having a group of one hundred genes well identified allows performing a deeper analysis. I propose to consider a new model of game, derived from the results of the (modified) microarray game, by considering a weighted majority game, with a much restricted set of genes, selected by means of the ranking of the indices.

It is clear that all of this must be considered, at the current state of the art, purely experimental. Several facts do not have, at the present, strong theoretical motivations. For instance, which index should be used to select a group of genes to analyze further with the weighted majority game. Then, how many genes should be used in the subsequent game. Of course, we must take into account the complexity of the calculations. Fortunately for this type of games the evaluation of the indices is much easier. In any case, I apply the whole machinery to data sets taken from the available literature. In particular, I consider data relative to *colon rectal tumour*, to *neuroblastic tumour*, to *lobular and ductal breast carcinomas*, to *colon tumour*. Very interestingly, a check made in the medical literature shows that some of the selected genes by our methods in particular experiments are considered to be of great importance from a medical point of view, in the onset of the disease.

To handle all calculations needed to evaluate the indices in the various experimantal data I have considered, I have developed a (simple) MATLAB program. I have also developed a similar program performed in C^{++}.

To conclude, I want to stress the following fact. On one side, the med-

ical literature and recent studies seem to suggest that clusters of genes could/should be considered when analyzing genetic responsibility in the onset of a disease. On the other side, however, it seems to be very important as well to try to single out small groups of genes individaully responsible of certain diseases. I think that the work done in this thesis suggests that the game theoretic approach, with the use of classical indices, such as Shapley's and Banzhaf, or the other ones here introduced, and the weighted indices, or also the use of the weighted majority games, serves well in the two approaches. On one side, the symmetric indices tend to group genes in small families, and this can help when considering clusters of them, on the other side the approach suggested in the last chapter seems to be promising in trying to better differentiate them.

This in my opinion motivates the idea to develop further research in this subject by using game heory, and to enhance interaction with scholars in molecular biology and medicine to suggest new developments of this approach.

Chapter 2

Preliminaries on Molecular Biology and Game Theory

2.1 Brief review on the molecular biology of cancer and on the microarray technology

The first version of human genome sequence was published at the beginning of this decade ([Lander et al. (2001)], [Venter et al. (2001)]). After the initial draft sequence, the information has been updated (International Human Genome Sequencing Consortium, 2004). The availability of the sequence information has promoted development of a number of high-throughput technologies, including microarrays. The microarrays have played an important role in changing the concept in biological research from investigation of single genes to an omics approach ([Ge et al. (2003)], [Liu et al. (2006)]). Omics studies are characterized by the use of high-throughput methods that produce large quantities of data. Microarrays can measure RNA, DNA, or protein levels from cells or tissues on a genome-wide scale. For example, DNA and RNA level alterations measured from the same sample provide information about genes in which expression is corrupted due to increased or decreased copy number. Copy number alterations represent an important mechanism for cancer cells to promote or suppress the expression of genes involved in cancer progression. Furthermore, genes deregulated in association with high level amplifications have been linked to poor outcome of cancer, representing potential drug targets ([Chin et al. (2006)]). Thus the integrated array data can identify therapeutic targets which might then provide alternative options to surgery and radiation therapy cancer.

2.1.1 Molecular biology of cancer

The ever increasing rate at which the different genomes are going to be decoded has opened a new area of biological research, named functional ge-

nomics, which is concerned with assigning biological function to the DNA sequences. With the complete DNA sequences of many genomes already available and the recent release of the first draft of the human genome, an essential and formidable task is to define the role of each gene and understand how the genome functions as a whole. Innovative approaches have been developed to exploit DNA sequence data and yield information about gene expression levels for entire genomes.
I now briefly review the basic genetic notions useful to understand the microarray experimental area. A gene consists of a segment of DNA which codes for a particular protein, the ultimate expression of the genetic information. A deoxyribonucleic acid or DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides. Each nucleotide comprises a phosphate group, a deoxyribose sugar, and one of four nitrogen bases. The four different bases found in DNA are: adenine (A), guanine (G), cytosine (C), and thymine (T). The two chains are held together by hydrogen bonds between nitrogen bases, with base-paring occurring according to the following rule: G pairs with C, and A pairs with T. While a DNA molecule is built from a four letter alphabet elements, proteins are sequences of twenty different types of amino acids. The expression of the genetic information stored in the DNA molecule occurs in two stages:

1. transcription during which DNA is transcribed into messenger ribonucleic acid or mRNA, a single stranded complementary copy of the base sequence in the DNA molecule, with the base uracil (U) replacing thymine;

2. translation during which mRNA is translated to produce a protein. The correspondence between DNA's four-letter alphabet elements and protein's twenty-letter basic units is specified by the genetic code which relates nucleotide triplets to amino acids.

Proteins and nucleic acids are two of the main biochemical components of the biological systems. As their full names imply, both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are chemically classified as nucleic acids and their main function is to store and encode the information used to synthesize proteins. Chromosomes, the molecular units of the genetic heredity, are composed of DNA organized into genes, while RNA, a less stable nucleic acid, is used to direct the process of protein synthesis. Under regulated conditions, specific regions of DNA corresponding to particular genes are transcribed into RNA that is then translated into proteins. Proteins are often mainly known for their enzymatic role in biological catalysis, but they are also needed for structure and support, movement and cellular communication. Following the discovery of the double helix structure of DNA by Watson and Crick in 1953, molecular biologists and biochemists have been interested in exploring the means by which nucleic acids encode information.

The gradual accumulation of knowledge has revealed this process to be a marvel of intricate complexity. DNA is chemically quite simple, composed of variations of only four nucleotides. This repeating polymer is organized into functional units (genes), and the collection of genes that make up an organism is referred to as its genome. With few exceptions, each cell of an organism contains a complete copy of its genome. The differences between individual cells in a multicellular organism are due to the regulated interactions and differential expression of particular genes. The protein products of gene expression interact with each other, with existing proteins in the cell, and often with the DNA itself to carefully control cellular conditions in a complicated pathway of feedback loops. Nowadays, a revolutionary technique, the microarray technology, allows for the collection of huge amount of information concerning the function of human genes. Cancer is regarded as a genetic disease that occurs due to sequential accumulation of genetic alterations in oncogenes, tumour suppressor genes and stability genes. These alterations cause abnormal activation or inactivation of a number of pathways resulting in uncontrolled cellular grow ([Volgelstein and Kinzler (2004)]). Environmental,viral, and chemical agents as well as physical substance can promote carcinogenesis ([Peto (2001)]). The risk of cancer can be associated with lifestyle and environmental factor even though hereditary factors also play a role. The majority of tumours derive from a single progenitor cell. Within a tumour, different subclones can have distinct alterations caused by simultaneous clonal expansion of different clones as a result of instability in a tumour genome ([Weinberg (2006)]). Instability can be acquired during tumour development or by inherited mutations occurring, for example, in genes that are responsible for genome integrity. Therefore a person with inherited mutations in critical genes becomes predisposed to cancer ([Fearon (1997)]). Moreover, the accelerated cell proliferation in cancer allows mutations to occur an increased rate. Cancer cells are characterized by acquired functional capabilities such as limitless replicative potential and acquisition of invasiveness and metastatic ability ([Hanahan and Weinberg (2000)]). Although recent studies have illuminated genetic changes needed to transform human cells ([Sjöblom et al. (2006)]), the exact number of changes required is still under debate. To date, 367 human genes have been causally implicated in cancer development either through mutation, copy number alteration or rearrangement (www.sanger,uk/genetics/CGP/Census). Recently cancer genes were mapped by a large-scale sequencing effort but the list of cancer genes is still not complete.

2.1.2 Microarray technology

DNA microarray is based upon the mutual and specific affinity of complementary strands of DNA. This approach provides a quantitative measurement of the gene expression (the amount of mRNA in a cell sample) for

thousands of genes in the same experiment. Array size can range from a small subset of 500 genes to a large pool of 30,000 genes. Once the purified samples have been prepared, they are individually spotted, usually in duplicate, onto glass slides in a predetermined array. While these are generally modified to promote the chemistry used in printing, these slides appear identical to the microscope slides used in any basic biology lab. A printed slide will contain two spots, each corresponding to a particular gene present in the array. Microarrays are used to probe differences in gene expression. In order to highlight these differences, the use of proper controls is vital. mRNA must be extracted from a normal control as well as the experimental samples and purified for use in the array experiment. This RNA can be obtained from a variety of sources including cell culture, tissue samples from animal models or clinical patients, and histologically-archived samples. Following mRNA extraction, reverse transcription PCR is used to convert the RNA transcripts into DNA. The complete pool of DNA obtained is representative of transcriptional events in the tissue source of the RNA. The genes that were being actively transcribed in the sample will have mRNA copies that should have been first purified and then copied into DNA during the PCR step. The reverse transcription event for the control and experimental mRNA are identical in every step except one, and it is this step that enables differential gene expression to be determined. Detection of the nucleic acid amount in the samples is performed using nucleotides typically labelled with fluorescent probes. In particular, nucleotides labelled with Cy3, a green fluorescent dye, are incorporated into the control DNA while nucleotides labelled with Cy5, a red fluorescent dye, are incorporated into the DNA coming form the biological samples. After extraction and labelling, both probes are mixed and allowed to *hybridize* onto the glass slide. The term hybridization refers to the annealing of nucleic acid strands from different sources according to the base-pairing rules described above. Excess hybridization buffer is removed after washing following an overnight incubation, and the slides are then ready to be subjected to quantification using a specific scanner. Hybridization is the crucial step of this procedure: many DNA regions immobilized on a small glass, plastic or nylon (probes), bind to a complementary sequence from the sample under study labelled with fluorescent dyes that flag their presence when exposed to a specific wavelength of light. If one of the single-stranded DNA probes corresponds to a single-stranded DNA gene printed on the slide, complementary interactions between the two will affix the probe to the slide. Then a laser ray activates the fluorescent dyes incorporated into the probe, and areas on the slide with hybridized probes will be visible on the scanned image as red or green spots. Gene spots with no affixed probe appear black. The red spots correspond to genes expressed in the experimental sample while green spots correspond to genes expressed in the control sample. If a gene is expressed under both conditions, both probes will hybridize and the spot will appear

yellow. Sophisticated laser scanning equipment is used to import data into image analysis software that can be quantify the gene expression on the basis of light intensity of the corresponding probes. Ratios comparing Cy5 and Cy3 intensities can be used to quantitatively evaluate gene expression. Under differing biological conditions, individual genes may be up-regulated or down-regulated, and the fluorescent signal of the marker dyes reflects these changes. Indeed, presently, the evaluation of the data generated from this analysis is one of the most complicated tasks of this technology. The array format definitely simplify the technical issues related to the investigation of the genome interactions, but the complexity of the data management remains still high. Since a 10,000-gene array generates 10,000 data points results must be validated through replication. A typical microarray experiment may utilize also thirty slides and produce vast quantities of data, whose analysis must generate a coherent picture of the system under investigation.

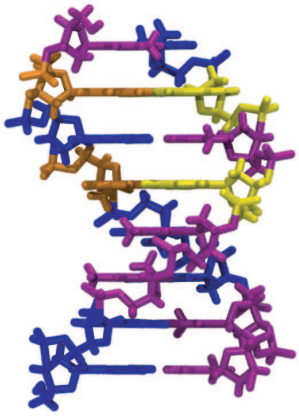


Figure 2.1: Double-stranded DNA.

Figure 2.2: Hybridization.

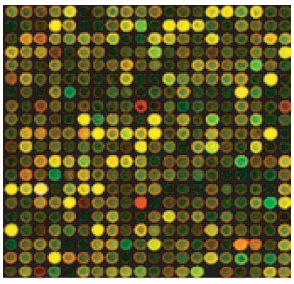


Figure 2.3: DNA microarray.

2.2 Brief review of game theory applied to gene expression analysis

In literature we find many models for data analysis aimed at understanding, from a matrix of gene expression data, the role of genes and their interactions when some changes in the biological system occur. By using the microarray technique to extract a gene expression data-set from samples, it is possible to produce a map of all genes expressed in the samples. Since many diseases, specifically tumours, are known to be of different classes, we can think to differentiate tumours classes according to the expression profile of the genes classified by a rule discriminating them. The classification rule could be exploited both to predict the class of a new tumour sample of unknown class by analyzing its gene expression profile, and to have meaningful information to apply in the field of cancer research.
From the mathematical point of view, the most puzzling problem in applying any method to analyze gene expression data-sets, is to find a strategy to reduce the number of genes under analysis: even if the definition of gene itself is not precisely given, the average number of genes present in the human genome is estimated around 30,000. The choice of a particular method to analyze microarray data about genes, is based on the possibility to select genes (or clusters of genes) having the most relevant role in mechanisms that cause biological changes (e.g. a tumour). As mentioned in the introduction, in this thesis I will apply game theory (in particular *coalitional games*) to the study of the interactions among genes, which can be considered, according to a very recent model developed in the literature, the players in a particular game, called *the microarray game*. The characteristic function of the microarray game picks up the information that can be successively exploited to define the role of each gene in each possible coalition by applying suitable solution concepts for cooperative games.

It is time to quickly introduce the basic concepts of game theory needed to develop the subsequent ideas.

2.2.1 Preliminaries

I start by introducing notations and some basic game theoretical notions. Let T be a (finite) set. To denote a subset S of T we use the notation $S \subseteq T$; $S \subsetneq T$ means $S \subseteq T$ and $S \neq T$; $S \nsubseteq T$ means that $S \subseteq T$ is not true. Let $|T|$ denote the cardinality of a finite set T: we shall often use the convention that $|T| = t$.
A *coalitional game* or *characteristic-form game* is a pair (N, v), where N denotes the finite set of *players* and $v : 2^N \to \mathbb{R}$ is its *characteristic function*, with $v(\emptyset) = 0$. If the set N of players is fixed, we identify a coalitional game (N, v) with the corresponding characteristic function v. We shall implicitly

assume from now on that $N = \{1, \ldots, n\}$. A group of players $T \subseteq N$ is called a *coalition* and $v(T)$ is called the *value* of this coalition. A coalitional game (N, w) such that $w : 2^N \to \{0, 1\}$ is called a $\{0, 1\}$-*game*. We shall denote by $\mathcal{W}$ the class of all $\{0, 1\}$-games, where $\mathcal{W} \subsetneq \mathcal{G}$, being $\mathcal{G}$ the class of all coalitional games.

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

The set $\mathcal{G}^N$ is a vector space, of dimension $2^n - 1$ (since all characteristic functions are valued zero at the empty set). Two collections of games provide interesting bases for the vector space. Let us introduce them. For each $R \subseteq N$, let the *unanimity game* (N, u_R) be defined as

$$u_R(T) = \begin{cases} 1 & \text{if } R \subseteq T \\ 0 & \text{otherwise} \end{cases}.$$

Another collection of games providing a base, surely less meaningful from the point of view of the interpretation as a game, but useful for purposes we shall see later, is obviously given by the canonical base of the corresponding Euclidean space. In terms of games, it is the collection of games v_R such that

$$v_R(T) = \begin{cases} 1 & \text{if } T = R \\ 0 & \text{otherwise} \end{cases}.$$

A *payoff vector* or *allocation* $x = (x_1, \ldots, x_n)$ of a coalitional game (N, v) is an n-dimensional vector describing the payoffs [1] of the players, such that each player $i \in N$ receives x_i. An allocation x is called *imputation* if it verifies the conditions:

1. $x_i \geq v(\{i\})$ for all $i = 1, \ldots, n$;
2. $\sum_i x_i = v(N)$.

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

The most famous solution in the theory of coalitional games is the *Shapley value*, introduced by Shapley (1953). Such a solution can be described

[1] More precisely, we can speak about payoffs when the characteristic function v is given the meaning of utility assigned to the calitions. More generally, the meaning of v induces the corresponding meaning to the allocation. In this work almost always v will represent the strenght of the coalitions, and thus an imputation represents the power (strenght) of the players in the game.

in several ways. I just give the Shapley value σ applied to game $(N, v) \in \mathcal{G}^N$ by means of its formula:

$$\sigma_i(v) = \sum_{S \in 2^{N \setminus \{i\}}} \frac{s!(n-1-s)!}{n!} m_i(v, S), \tag{2.1}$$

where the quantity $m_i(v, S) = v(S \cup \{i\}) - v(S)$ is the *marginal contribution* of player i to the coalition S. When $v \in \mathcal{W}$, and is monotonic (i.e. $S \supseteq T$ implies $v(S) \geq v(T)$) $m_i(v, S)$ can assume only the values zero and one: when $m_i(v, S) = 1$ we say that i is a *swing* for the coalition S.

Another one-point solution for coalitional games is the *Banzhaf value*, introduced by Banzhaf (1965). The Banzhaf value $\beta(v)$ of the game $v \in \mathcal{G}^N$, is defined as follows:

$$\beta_i(v) = \sum_{S \in 2^{N \setminus \{i\}}} \frac{1}{2^{n-1}} m_i(v, S), \tag{2.2}$$

for each $i \in N$.

However, it must be noticed that the Banzhaf value is not an imputation. It is quite possible to introduce different indices. An interesting analysis of some of them is carried out in the paper [Monderer and Samet (2001)]. In particular, a solution ψ is called a *probabilistic* value, if for each player i there exists a probability measure p^i on $2^{N \setminus \{i\}}$ such that

$$\psi_i(v) = \sum_{S \in 2^{N \setminus \{i\}}} p_i(S) m_i(v, S). \tag{2.3}$$

Thus for instance, in the case of Shapley, for all i

$$p_i(S) = \frac{1}{n\binom{n-1}{s}}.$$

It should be noticed that $p_i(S)$ does not depend, from the Shapley and Banzhaf indices, from the single player i. This property is clearly a symmetry property.

Probabilistic values were characterized by Weber in [Weber (1988)], where a formula too is offered in order to explicitly provide the coefficients in the formula. I will show in a subsequent chapter that the indices I introduce fulfill the conditions given by Weber, and that it is possible to provide a formula for the probabilistic coefficient.

A probabilistic value which is symmetric, is called a *semivalue*. If the probabilistic coefficient is positive for every i and S, it is called a *regular semivalue*.

I will mainly consider two types of (cooperative) games. The first is very well known: it is the class of the *weighted majority games*, a subclass of $\mathcal{W}$,

the class of the $\{0,1\}$-games. Suppose there are n players and that $n+1$ positive integers $q, w_1, \ldots, w_n$ are given. The associated weighted majority game v is defined as:

$$v(S) = \begin{cases} 1 & \text{if } w(S) \geq q \\ 0 & \text{if } w(S) < q \end{cases},$$

where $w(S) = \sum_{i \in S} w_i$. Such a game will be denoted by $[q; w_1, \ldots w_n]$. The meaning of the coefficients is clear: each player i has assigned a (positive) weight w_i, and a quota q is necessary to get the majority. Thus a coalition S is winning provided the sum of the weights of its players joins the quota. It is clear that such a model well serves to analyze, for instance, what can happen in a parliament where different parties are present. Here I will introduce a weighted majority game related to the analysis of microarray data. The key point will be how to assign weights to the genes. A less important point, that however must be taken into account, is the fact that I will consider also the case when a player in a game has zero weight. This is needed, since I will consider averaged sums of (weighted majority) games, and in a single game a player could be with no weight assigned, even if in the resulting game its weight is positive.

It is time now to introduce the main class of games object of this thesis. It is the class $\mathcal{M}^N$ of Microarray games, where the set N of the players is a given family of genes. Here I recall only the relevant facts for this work, for more, especially for the motivations to consider such a model, see [Moretti et al. (2007)].

Consider an $(n \times m)$ matrix $M = (m_{ij})$, such that m_{ij} is either zero or one, and such that for every j there is i with $m_{ij} \neq 0$. Given the column $m_{\cdot j}$, $j = 1, \ldots, m$, define its *support* as the set $\operatorname{supp} m_{\cdot j} = \{i : m_{ij} = 1\}$, and define the associated unanimity game v^j generated by $\operatorname{supp} m_{\cdot j}$, i.e.

$$v^j(T) = \begin{cases} 1 & T \supseteq \operatorname{supp} m_{\cdot j} \\ 0 & \text{otherwise} \end{cases}.$$

Then the microarray game associated to $M = (m_{ij})$ is defined as

$$v = \frac{1}{m} \sum_{j=1}^{m} v^j.$$

It makes sense that, in studying a particular disease, the set N of the genes is kept fixed, while the set of patients can vary. Thus microarray games $\mathcal{M}^N$ can be described by means of $(n \times m)$ matrices like above, with m ranging over the natural numbers. Sometimes, for $v \in \mathcal{M}^N$, we shall use the notation $v = (v^1, \ldots, v^j, \ldots, v^m)$ to stress the role of the generic patient j in the game v.

Example 2.2.1 Consider the matrix $M \in \{0,1\}^{3\times 3}$ such that

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

Then $\operatorname{supp} m_{\cdot 1} = \{2\}$, $\operatorname{supp} m_{\cdot 2} = \{1,2\}$ and $\operatorname{supp} m_{\cdot 3} = \{1,2,3\}$. The corresponding microarray game $(\{1,2,3\}, v)$ is such that

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

It follows that $v(\emptyset) = v(\{1\}) = v(\{3\}) = v(\{1,3\}) = 0$; $v(\{2\}) = v(\{2,3\}) = \frac{1}{3}$; $v(\{1,2\}) = \frac{2}{3}$; $v(\{1,2,3\}) = 1$. The Shapley value of the microarray game $(\{1,2,3\}, v)$ is $\sigma(v) = (\frac{5}{18}, \frac{11}{18}, \frac{2}{18})$, whereas the Banzhaf value is $\beta(v) = (\frac{3}{12}, \frac{7}{12}, \frac{1}{12})$. ■

For more on cooperative games, see for instance [Owen (1995)]

Chapter 3

Axiomatic Characterization for Microarray Games

As already mentioned in the introduction, it is particularly meaningful, in cooperative game theory, to characterize solutions by means of a short list of reasonable properties: it is the so called axiomatic approach. One of its merits is to highlight the features of a solution with respect to other solutions. This characterization becomes even more meamningful when considering *specific* classes of games, and not only the class of all games. Thus, it meakes sense to do this for the class of microarray games.
In [Moretti et al. (2007)] it was proved that the Shapley value is the *only* one point solution, on the class of microarray games, fulfilling a pool of reasonable properties that we shall describe later. In this chapter I will give another pool of properties characterizing the Banzhaf value. In doing this, I also produce another alternative characterization of the Shapley value.

To start with, I introduce two classical properties, often used in this context.

Property 1 *Let $v \in \mathcal{G}^N$. The solution ψ has the* dummy player *(DP) property, if for each player $i \in N$ such that $v(A \cup \{i\}) = v(A) + v(\{i\})$, then*

$$\psi_i(v) = v(\{i\}). \tag{3.1}$$

In other words, the player i is useless in joining any coalition, so that the solution does not assign to him more than what he is able to get by himself, without making coalitions with other players.

Property 2 *Let be given a finite set N of genes, and let $\pi : N \to N$ a permutation on N. Given the game v, denote by $\pi^* v$ the following game: $\pi^*(v(A)) = v(\pi(A))$, and by $\pi^*(x) = (x_{\pi^*(1)}, \ldots, x_{\pi^*(n)})$. The solution ψ has the* symmetry *(S) property on $\mathcal{M}^N$, if $\psi(\pi^*(v)) = \pi^*(\psi(v))$.*

This is a clear condition of symmetry between the players. It essentially implies that if two players bring the same marginal utility when joining coalitions, then the solution will assign the same to both.

Now, some properties, more specific for our context.
To start with, I introduce a new definition, motivated by an analogous one given in [Kalai and Samet (1987)], for general cooperative games.

Definition 3.0.1 *Let $v \in \mathcal{M}^N$. A coalition $S \in 2^N \setminus \{\emptyset\}$ such that for each $T \subsetneq S$ and each $R \subseteq N \setminus S$ it holds that*

$$v(R \cup T) = v(R), \tag{3.2}$$

is said to be a partnership of genes *in v.*

We call *relevance index* for genes a one point solution solution $F : \mathcal{M}^N \to \mathbb{R}^N$ with the property that $F(v) \geq 0$ and $F(v) \neq 0$ for all v.[1] Some interesting properties for relevance indices, related to the concept of partnership of genes, are the following.

Property 3 *Let be given a finite set N of genes. The solution F has the* Partnership Rationality *(PR) property on $\mathcal{M}^N$, if for every $v \in \mathcal{M}^N$*

$$\sum_{i \in S} F_i(v) \geq v(S) \tag{3.3}$$

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

Property 4 *Let be given a finite set N of genes. The solution F has the* Partnership Feasibility *(PF) property on $\mathcal{M}^N$, if for every $v \in \mathcal{M}^N$,*

$$\sum_{i \in S} F_i(v) \leq v(N) \tag{3.4}$$

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

Property 5 *Let be given a finite set N of genes. The solution F has the* Partnership Monotonicity *(PM) property on $\mathcal{M}^N$, if for every $v \in \mathcal{M}^N$:*

$$F_i(v) \geq F_k(v)$$

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

[1] Inequalities among vectors are intended coordinatewise.

The (PR) property states that the total relevance of a partnership of genes in determining the onset of the tumour in the individuals should not be lower than the average number of cases of tumour enforced by the partnership itself.
The (PF) property states that the total relevance of a partnership of genes in determining the tumour onset in the individuals should not be greater than the average number of cases of tumour enforced by the grand coalition.
Finally, the (PM) property means the following: consider two disjoint partnerships of genes enforcing the same average number of cases of tumour 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 tumours - then genes in the smaller partnership should receive a higher relevance index than genes in the bigger one, where the likelihood that some genes are redundant is higher.

Property 6 *Let be given a finite set N of genes. The solution F has the* Equal Splitting *(ES) property on $\mathcal{M}^N$, if for every $v_1, \ldots, v_k \in \mathcal{M}^N$*

$$F(\frac{\sum_{i=1}^{k} v_i}{k}) = \frac{\sum_{i=1}^{k} F(v_i)}{k}. \tag{3.5}$$

The equal splitting property (ES) clearly reminds the classical linearity assumption introduced by Shapley in order to characterize its index. However it should be observed that in this context it looks much more intuitive, due to the fact that it is required (only) for the class of games which are averages of unanimity games.

Property 7 *Let be given a finite set N of genes. The solution F has the* null gene *(NG) property on $\mathcal{M}^N$, if for every $v \in \mathcal{M}^N$ and for each null gene*[2] *$i \in N$*

$$F_i(v) = 0. \tag{3.6}$$

Clearly, this property is a simple adaptation of the dummy player property to this context.

We address the interested reader to the paper [Moretti et al. (2007)] for a deeper discussion of the meaning of the above properties, as well as for the proof of the following theorem.

Theorem 3.0.1 *Let be given a finite set N of genes. 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 (NG).*

[2]The gene i is said to be a *null gene* if $m_{ij} = 0$ for all j.

We now introduce some new properties, called *symmetry*, *individual consistency*, *average loss*, and *total loss*, respectively.

Property 8 *Let be given a finite set N of genes. The solution F has the* symmetry *(S) property on $\mathcal{M}^N$, if for every game $v \in \mathcal{M}^N$, for every partnership S of v and every $i, k \in S$, $F_i(v) = F_k(v)$.*

Property 9 *Let be given a finite set N of genes. The solution F has the* individual consistency *(IC) property, if*

$$F_i(u_{\{i\}}) = 1 \tag{3.7}$$

for each $i \in N$.

Property 10 *Let be given a finite set N of genes. Let $v = (v^1, \ldots, v^m) \in \mathcal{M}^N$, let S be a partnership of genes in v, let $l \in \{1, \ldots, n\}$. Define a new microarray game v_{Sl} in the following way:*

1. *for j such that $v^j(S) = 1$*

$$v^j_{Sl}(T) = \begin{cases} 1 & T \supseteq S \cup \{l\} \\ 0 & \text{otherwise} \end{cases};$$

2. *otherwise, $v^j_{Sl} = v^j$.*

Then the solution F has the average loss *(AL) property on $\mathcal{M}^N$, if for every v, v_{Sl} as above*

$$\frac{1}{s} \sum_{i \in S} [F_i(v) - F_i(v_{Sl})] = F_l(v_{Sl}) - F_l(v). \tag{3.8}$$

On the other hand, F has the total loss *(TL) property, if*

$$\sum_{i \in S} [F_i(v) - F_i(v_{Sl})] = F_l(v_{Sl}) - F_l(v). \tag{3.9}$$

Note that both axioms (TL) and (AL) concern the effect of adding a gene to a partnership in a microarray game v. Following the interpretation of similar axioms introduced in [Laruelle, Valenciano (2001)], constant total (respectively, average) loss here postulates that the total (respectively, average) loss of the genes in the partnership S equals the total (respectively, average) gain of the gene l added to S. Even if these two properties are remarkably close, they play a very different role in characterizing relevant indices, as it will be shown by Corollary 3.0.1 and Theorem 3.0.2.

To introduce the last property, we need some more notation. So, take a game $v \in \mathcal{M}^N$, and let M be the generating matrix. Let l be a null gene in v and $k \neq l$ another gene. Consider a new matrix M^{lk} defined by means of its rows:

$$m^{lk}_{i\cdot} = m_{i\cdot} \quad \text{if } i \neq l, \quad m^{lk}_{l\cdot} = m_{k\cdot}.$$

We call v_{lk} the game associated to the matrix M^{lk}.

We now introduce another property.

Property 11 *Let be given a finite set N of genes. The solution F has the* Pairwise Consistency *(PC) property on $\mathcal{M}^N$, if for every v, v_{lk} as above*

$$F_k(v) = F_l(v_{lk}) + F_k(v_{lk}) \tag{3.10}$$

Let us briefly comment on this property. The difference between the game v, where l is a null player, and the associated game v_{lk} is that in the second one the gene l is substituted by the gene k. In other words, the null gene is deleted in the new game, and the effect of the (non null) gene k is "doubled". Pairwise consistency thus requires that the power k has in the game v is now split into the players k and l, in the new game. This makes sense, since the player l in the old game had no power. Thus by changing it with another player would not affect the total sum of the powers of the genes. On the other hand, the power k had in the former game, should be split among k and l, since the other players do play the same role in the two games, and thus their relative power should not change.

Remark 3.0.1 It is not difficult to see that the game v_{lk} has the following form: for all $S \subseteq N$,

1. if $k \in S$ and $l \notin S$, then $v_{lk}(S) = v(S \setminus \{k\})$;
2. otherwise $v_{lk}(S) = v(S)$.

Moreover, it is clear that given $v = (v^1, \ldots, v^m)$ and the associated game $v_{lk} = (v^1_{lk}, \ldots, v^m_{lk})$, it holds that supp v^j = supp v^j_{lk} if $m_{kj} = 0$, supp v^j = supp $v^j_{lk} \cup \{l\}$ if $m_{kj} = 1$. ■

It is straightforward to see that the following relations among the properties hold: (PM) and (NG) together imply (S), (PR) and (PF) together imply (IC). The Banzhaf value satisfies the (NG), (ES), (PM), (PF) properties on the class of microarray games (see [Moretti et al. (2007)]). It does not satisfy (PR). We shall prove that it satisfies also the (IC) and (AL) properties.

The following is well known and easily seen to be true:

Proposition 3.0.1 *Given a coalition S, the Banzhaf and Shapley values for the unanimity game associated to S do assign 0 to the genes not in S, and $\frac{1}{2^{s-1}}$, $\frac{1}{s}$ respectively, to the genes in S.*

Proposition 3.0.2 *The Banzhaf value satisfies the properties (IC) and (AL). The Shapley value satisfies (TL).*

Proof Due to the fact that the Banzhaf value satisfies the equal splitting property, it is enough to show (IC) on a single patient j. But this readily follows from Proposition 3.0.1, applied to a single player coalition. As far as the (AL) property is concerned, let us consider a game $v = (v^1, \dots, v^m)$ and a partnership S of the game. Once again, the formula can be checked on the game v^j and the only interesting case is when $v^j(S) = 1$. Since S is a partnership of the game, this is the case if and only if $S = \text{supp } m_{\cdot j}$. Thus, for every $i \in S$, $\beta_i(v^j) = \frac{1}{2^{s-1}}$. Now, observe that the formula must be checked only in the case when $l \notin S$. Thus $\beta_i(v^j_{Sl}) = \frac{1}{2^s}$ for all $i \in S \cup \{l\}$. Thus

$$\frac{1}{s}\sum_{i\in S}[\beta_i(v^j) - \beta_i(v^j_{Sl})] = \frac{1}{s}\cdot s(\frac{1}{2^{s-1}} - \frac{1}{2^s}) = \frac{1}{2^s}.$$

On the other hand,

$$\beta_l(v^j_{Sl}) - \beta_l(v^j) = \frac{1}{2^s} - 0.$$

About the Shapley value:

$$\sum_{i\in S}[\sigma_i(v^j) - \sigma_i(v^j_{Sl})] = s(\frac{1}{s} - \frac{1}{s+1}) = \frac{1}{s+1}.$$

On the other hand,

$$\sigma_l(v^j_{Sl}) - \sigma_l(v^j) = \frac{1}{s+1} - 0.$$

This ends the proof. ∎

Remark 3.0.2 Suppose ϕ is a relevance index on $\mathcal{M}^N$ fulfilling (S) and (AL). Given a coalition S, and a gene $l \notin S$, consider the two unanimity games u_S and $u_{S\cup\{l\}}$. Then, for all $i \in S$ it holds that

$$\phi_i(u_{S\cup\{l\}}) = \frac{1}{2}\phi_i(u_S).$$

This readily follows from the following facts:

1. S is a partnership for u_S;

2. (S) implies
$$\phi_i(u_{S\cup\{l\}}) = \phi_k(u_{S\cup\{l\}})$$
for all $i, k \in S \cup \{l\}$,
$$\phi_i(u_S) = \phi_k(u_S)$$
for all $i, k \in S$;

3. thus
$$\phi_i(u_S) - \phi_i(u_{S\cup\{l\}}) = \phi_l(u_{S\cup\{l\}}) = \phi_i(u_{S\cup\{l\}}),$$

and we get the conclusion from the last equality.
Observe that a similar argument shows that if a relevance index ϕ satisfies (S) and (TL) then
$$\phi_i(u_{S\cup\{l\}}) = \frac{s}{s+1}\phi_i(u_S).$$

■

Theorem 3.0.2 *Let be given a finite set N. Then a relevance index ϕ on $\mathcal{M}^N$ satisfies the properties (S), (ES), (NG) and (AL) if and only if there is $a > 0$ such that $\phi = a\beta$.*

Proof In Proposition 3.0.2 we have already seen that the Banzhaf value satisfies the property (AL). (NG) and (ES) are obvious. (S) is shown in [Moretti et al. (2007)]. The same proof shows that a positive multiple of the Banzhaf value fulfills all properties above. Now we prove uniqueness, modulo a positive factor, of the relevance index. Consider a relevance index $\phi : \mathcal{M}^N \to \mathbb{R}^n$ satisfying the same properties.

We start by proving the statement for unanimity games. We claim that there is $a > 0$ such that, for the unanimity game (N, u_S), it holds that $\phi_i(u_S) = 0 \ \ \forall i \notin S$, and $\phi_i(u_S) = \frac{a}{2^{s-1}}$ for $i \in S$. First of all, remember that S is a partnership in the game u_S; then the first statement is immediate from (NG); about the second: let $S = \{1\}$; then from (NG), $\phi_i(u_S) = 0 \ \ \forall i \neq 1$; set $\phi_1(u_S) = a > 0$. Now applying (AL), once with $S = \{1\}$ and $l = i$, and successively with $S = \{i\}$ and $l = 1$, we see that

$$\phi_1(u_{\{1,i\}}) + \phi_i(u_{\{1,i\}}) = \phi_1(u_{\{1\}})$$

and

$$\phi_1(u_{\{1,i\}}) + \phi_i(u_{\{1,i\}}) = \phi_i(u_{\{i\}}).$$

So that $\phi_1(u_{\{1\}}) = \phi_i(u_{\{1\}})$ and the statement is proved for the one player coalitions. The argument now goes by induction on the cardinality of the coalitions. Suppose we have shown the claim for all coalitions of cardinality less or equal to s, and consider a coalition of the form $S \cup \{l\}$, with $l \notin S$. From Remark 3.0.2 we have that, for $i \in S$, $\phi_i(u_{S\cup\{l\}}) = \frac{1}{2}\phi_i(u_S) = \frac{a}{2^s}$. On the other hand, $\phi_l(u_{S\cup\{l\}}) = \frac{a}{2^s}$, by symmetry. Thus we have shown that

$\phi = a\beta$ on all unanimity games. Now the (ES) property allows us showing the claim, and the proof is complete. ■

From the theorem we easily get the following Corollary:

Corollary 3.0.1 *Let be given a finite set N. The Banzhaf value on the class $\mathcal{M}^N$ of microarray games is the unique relevance index which satisfies the properties (IC), (S), (ES), (NG) and (AL).*

A similar argument yields the following result:

Theorem 3.0.3 *Let be given a finite set N. Suppose a relevance index ϕ on $\mathcal{M}^N$ satisfies the properties (IC), (S), (ES), (NG) and ((TL)). Then ϕ is the Shapley value: $\phi = \sigma$.*

Note that properties (AL) and (TL) make the difference between the two sets of axioms used in Corollary 3.0.1 and Theorem 3.0.3 (see [Laruelle, Valenciano (2001)]). Both (AL) and (TL) have a clear meaning and are similarly compelling in the context of relevance indices. (TL) goes in the direction to give more relevance to single genes which have value 1 throughout the columns of $\mathcal{M}^N$ only occasionally, possibly due to chance. If $\mathcal{M}^N$ is generated from real data, a relevance index which satisfies the (TL) property faces the risk to overestimate the role of genes whose expression value is more sensitive to stochastic noise. On the other hand, a relevance index satisfying property (AL) seems to go in the direction to flatten the roles of genes within and between partnerships, especially in those data-sets where the set of genes is fragmented in several partnerships of similar size. From a practical perspective, the results provided by Corollary 3.0.1 and Theorem 3.0.3 seems to suggest to also look at relevance indices satisfying properties which involve an 'intermediate' loss-gain balance.

We now provide another characterization of the Banzhaf relevance index, more in line with that one proposed in [Moretti et al. (2007)] for the Shapley value. With respect to their list of properties, we know that the only difference between the two indices is that Shapley satisfies the partnership rationality, while Banzhaf does not. Thus the problem becomes how to substitute (PR) in a way to single out the Banzhaf value. The idea is to use the property of pairwise consistency.

Theorem 3.0.4 *There is one and only one index $\phi : \mathcal{M}^N \to \mathbb{R}^n$ fulfilling the properties (NG), (S), (ES), (IC), (PC). Then ϕ is the Banzhaf value: $\phi = \beta$.*

Proof We must show that the Banzhaf value fulfills (PC), and that it is the only one fulfilling the above list of properties. About the first point. Take a

game v with matrix M, and set $\mu_j = |\text{ supp } m_{\cdot j}|$. Observe that, for a game $v \in \mathcal{M}^N$, it holds that

$$m\beta_k(v) = \sum_{j:m_{kj}=1} \frac{1}{2^{\mu_j - 1}}.$$

Now, take a game v and l, k as in (PC), and write $M_{lk} = (\hat{m}_{ij})$ for the matrix associated to the game v_{lk}. Set finally $\hat{\mu}_j = |\text{ supp } \hat{m}_{\cdot j}|$. Observe that $\hat{\mu}_j = \mu_j$ if $m_{kj} = 0$, $\hat{\mu}_j = \mu_j + 1$ if $m_{kj} = 1$. It follows that

$$m\beta_k(v_{lk}) = m\beta_l(v_{lk}) = \sum_{j:m_{kj}=1} \frac{1}{2^{\mu_j}}.$$

Since

$$m\beta_k(v) = \sum_{j:m_{kj}=1} \frac{1}{2^{\mu_j - 1}},$$

the first part of the claim follows.

Now, let $\phi : \mathcal{M}^N \to \mathbb{R}^n$ be a relevance index fulfilling the above list of properties. It is clear that, thanks to (ES), it is enough to show that $\phi = \beta$ on unanimity games. To start with, observe that the (IC) property implies that $\phi_k(u_{\{l\}}) = \beta_k(u_{\{l\}})$ for all $k, l \in N$. Next, observe that l is a null gene in u_S and thus we can apply (PC) and (S) to get:

$$\phi_i(u_S) = \phi_i(u_{S\cup\{l\}}) + \phi_l(u_{S\cup\{l\}}) = 2\phi_i(u_{S\cup\{l\}}).$$

This allows to conclude the proof, since it implies, with a simple argument by induction, $\phi_i(u_S) = \beta_i(u_S) = \frac{1}{2^{s-1}}$, while the null gene property shows $\phi_i(u_S) = \beta_i(u_S) = 0$ for $i \notin S$. ■

Remark 3.0.3 In the spirit of Theorem 3.0.2, it can be shown that if we do not require (IC) in the above list of properties, the relevance index fulfilling all other ones must be a positive multiple of the Banzhaf value. ■

The following section will enhance our claim, by comparing the results given by the two indices on an interesting case study present in the literature.

3.1 Colon data analysis

Moretti et al. (2007) introduced a preliminary application of the Shapley value for a microarray game defined on a tumour/normal data-set published in [Alon et al. (1999)] [3] containing expression levels of a set N of 2000 genes

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

measured using *Affymetrix oligonucleotide* microarrays for a set of 40 tumour samples and a set of 22 normal samples, in total 62 samples from colon tissues. In that application, after the preprocessing stage performed by the Bioconductor[4] specific software for microarray analysis, a discriminant method was applied on tumour sample data in order to provide a boolean expression matrix which finally produces the corresponding microarray game (N, v_c).

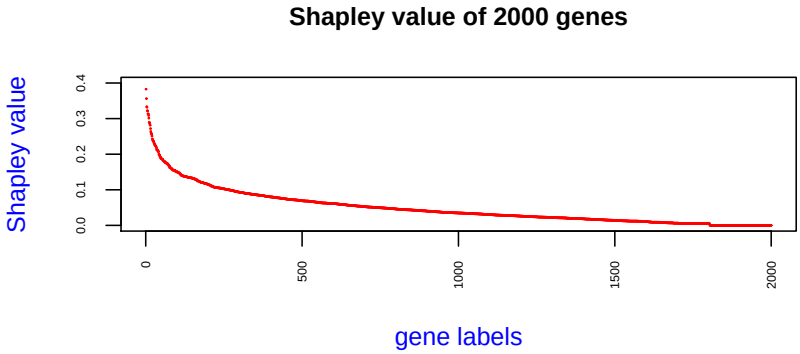
In this section we compare the results produced by the application of the Shapley value $\phi(v_c)$ with the results produced by the application of the Banzhaf value $\beta(v_c)$.

Both the Shapley value $\phi(v_c)$ and the Banzhaf value $\beta(v_c)$ are computed using functions implemented in the programming language R (R Development Core Team (2004)).

The Shapley value and the Banzhaf value of the 2000 genes are depicted in Figure 1. For each $k = 1, 2 \ldots 2000$, the number of genes which are among the first k with highest Shapley value and, at the same time, among the first k with highest Banzhaf value, is shown in Figure 2. If relevant genes are selected as the first k genes with highest Shapley value, these genes usually do not coincide with the first k genes with highest Banzhaf value, for each $k \in \{1, \ldots, 2000\}$, and an overlap of more than 50% is reached for $k \geq 260$ (see Figure 2). The first 40 genes with the highest Banzhaf value show the same value (approximately $\beta_i(v_c) \simeq 4.54\ 10^{-14}$), or at least no differences among these genes are detectable in terms of Banzhaf value. Most of their relevance according to the Banzhaf value was due to the contribution of a sample in which those 40 genes coincides with the support of the sample. Figure 3 shows the effect of the sample with smallest support on the Banzhaf value of top ranked genes. Note that all 40 top ranked genes according to Banzhaf value are abnormally expressed in the sample with precisely 40 genes abnormally expressed (triangles point-down and diamonds on the same vertical line for $x = 40$). Differently, the Shapley value is much less affected by the contribution of samples with small support, as it is also shown in Figure 3 (triangles point-up). As we noted earlier, the difference is in the way the indices change as long as the cardinality of supports grow in samples (i.e. columns of the binary matrix). This fact is also confirmed by the comparison with $\omega_i(v_c)$, which is the ratio of samples such that gene i takes value 1 in the Boolean matrix, for each gene $i \in N$. Figure 4 shows the number of genes among the first m genes with highest Shapley value (stair steps line labelled by 'Shapley') and the first m with highest Banzhaf value (stair steps line labelled by 'Banzhaf') which are also among the first m with highest $\omega(v_c)$, for each $m \in \{1, \ldots, 500\}$. The overlap of ω-based ranking with the Shapley value ranking is systematically larger than with the Banzhaf value ranking.

[4] http://www.bioconductor.org/

3.1.1 Figures


Shapley value of 2000 genes
Shapley value
0.0
0.1
0.2
0.3
0.4
0
500
1000
1500
2000
gene labels

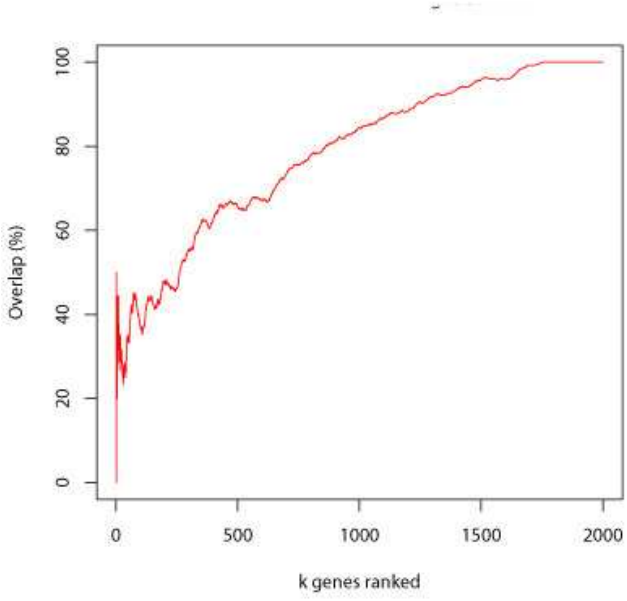


Figure 3.2: Overlap of genes ranked by the Shapley value and the Banzhaf value. For each $k = 1, 2, \ldots, 2000$ on the x-axis, the cardinality of the intersection between the set of k genes with highest Shapley value $\phi(v_c)$ and the set of k genes with highest Banzhaf value $\beta(v_c)$ is shown on the y-axis.

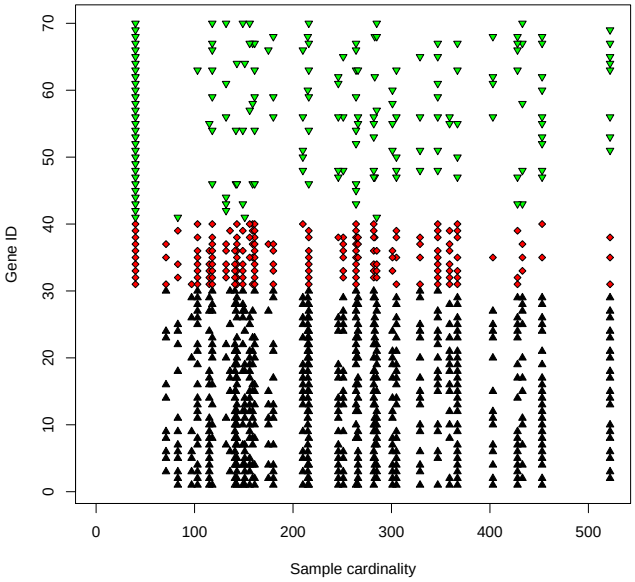


Figure 3.3: Comparison among the 40 genes with highest Shapley value (triangle point-up) and the 40 genes with highest Banzhaf value (triangle point-down). Ten genes which are top ranked by both Shapley and Banzhaf values are represented by diamond. Points on the same horizontal line belongs to the same gene. For each gene i represented on the y-axis, the x-coordinate of a point on the y_i-coordinate represents the cardinality of a sample in which gene i is abnormally expressed.

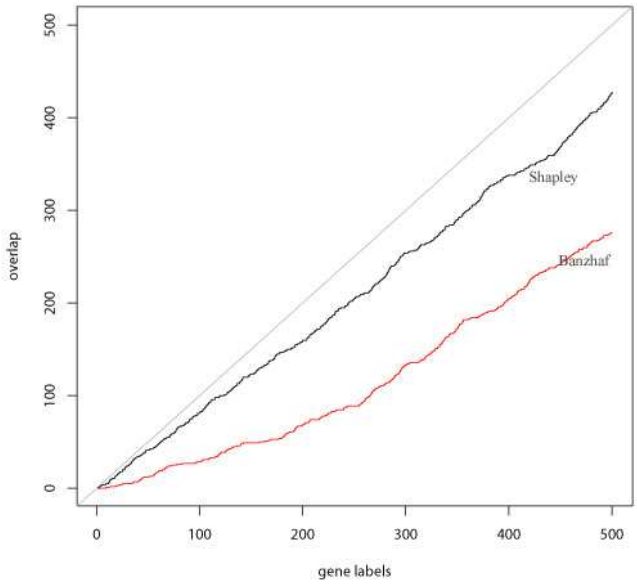


Figure 3.4: Genes are labelled on the x-axis. For each gene $m \in \{1, \ldots, 500\}$, the y_m-coordinate of each point on the stair steps line labelled by 'Shapley' equals the cardinality of the intersection between the set of m genes with highest Shapley value $\phi(v_c)$ and the set of m genes $\omega(v_c)$; the y_m-coordinate of each point on the stair steps line labelled by 'Banzhaf' equals the cardinality of the intersection between the set of m genes with highest Banzhaf value $\beta(v_c)$ and the set of m genes $\omega(v_c)$.

3.2 Some thoughts on Banzhaf versus Shapley

The two relevance indices are suitable to rank genes potentially responsible of a genetic disease. In general, they will give different ranking. How can we interpret this fact? In this section we briefly comment on this.

As it is clearly shown by the analysis of the previuos case study, the differences in the two indices arise from the differences of their behavior with respect to the unanimity games. So, what is the basic difference among them, when dealing with this type of games? Of course, they do assign zero to the players not belonging to the winning coalition, and the same amount to the players in the winning coalition. The difference is *in the way the relevance index changes as long as the cardinality of the coalition grows.* For, in the case of Banzhaf, for a coalition with s elements the value is $\frac{1}{2^{s-1}}$, while in the case of Shapley it is $\frac{1}{s}$. Then we see that the value decreases much more quickly for the Banzhaf value. This means that this relevance index gives much more importance to genes appearing in winning coalitions with *few* elements. Just to give an example, the contribution one gene has in a patient where it is the only one abnormally expressed counts as being in the support made by 10 elements in 10 patients as far as the Shapley value is concerned, while 2^9 patients are needed for Banzhaf's. Thus, we can expect that a better ranking for Banzhaf with respect to Shapley roughly indicated that the gene is abnormally expressed in patients having a relatively small group of abnormally expressed genes.

Thus the great difference of behavior of the two indices in microarray games is due to the fact that in the unanimity games, on one side the Shapley index depends linearly with respect to size of the winning coalition, while the Banzhaf index depends exponentially. Moreover, it can be noticed, by analyzing experimental results, that the Banzhaf index is unable to make a clear distinction between the various genes: in many cases the results divide the genes in few big groups, and within a given group all genes have the same index. So, it makes sense to consider what happens when we consider indices intermediate between Shapley's and Banzhaf's. The aim of the next chapter is to introduce new indices depending as a given power from the size of the winning coalition.

I conclude by mentioning that the results of this chapter are taken from the paper [Lucchetti et al.].

Chapter 4

A Family of New Indices

In this chapter, I introduce a family of indices on the set $\mathcal{G}^N$, just defining them on the set of the unanimity games, then extending them to the whole space by linearity.

4.1 Definition and main properties of the indices

Definition 4.1.1 *Let a be a natural number. We shall denote by σ^a and call* a-index *the one point solution defined on the unanimity game u_R as*

$$\sigma_i^a(u_R) = \begin{cases} \frac{1}{r^a} & \text{if } i \in R \\ 0 & \text{otherwise} \end{cases} .$$

On a generic game $v \in \mathcal{G}^N$, σ^a is extended by linearity.

It is clear that for $a = 1$ the index is the Shapley index. Even if the definition is given for all natural a, we mostly concentrate on the case $a = 2$. Among other things, I shall provide a general formula for the indices. It looks ugly, but at least for the case $a = 2$ can be simplified in a way that its computational complexity should be of the order of Shapley's index. Moreover, I shall show that σ^a is a probabilistic index for all a. The results are based on the following characterization of the probabilistic factor:

$$p_i(S) = \sigma_i^a(v_{S\cup\{i\}}). \tag{4.1}$$

In other words, in order to get the probabilistic coefficient, since the index σ^a is defined on the base of the unanimity games, it is necessary to find a formula of change of base. Before doing this, I prove that σ^a fulfills the dummy property (DP)[1].

[1] In [Weber (1988)] it is shown that an index is probabilistic if and only if it is linear, fulfills the (DP) property and the coefficients in (4.1) are positive.

Proposition 4.1.1 *Let v be any game, and let ϕ be a power index fulfilling the linearity, null player property and such that $\phi_j(u_{\{j\}}) = 1$ for all j, let i be a dummy player in the game v. Then*

$$\phi_i(v) = v(\{i\}), \tag{4.2}$$

for all a.

Proof Every game v can be written as a linear combination of unanimity games: $v = \sum_T c_T u_T$, where the coefficients c_T are inductively defined as: $c_{\{i\}} = v(\{i\})$ and, for $T \subseteq N$, $t \geq 2$

$$c_T = v(T) - \sum_{A \subset T} c_A. \tag{4.3}$$

It is clear that $c_{\{i\}} = v(\{i\})$; we now show that for every nonempty coalition T not containing i the coefficient $c_{T\cup\{i\}}$ is vanishing. Suppose $T = \{j\}$, with $j \neq i$. Then

$$c_{\{i,j\}} = v(\{i,j\}) - c_{\{i\}} - c_{\{j\}} = v(\{i\}) + v(\{j\}) - c_{\{i\}} - c_{\{j\}} = 0.$$

Suppose now we have proved the claim for all coalitions A containing i of size $2, 3, \ldots, s-1$ and consider a coalition of the form $T \cup \{i\}$, with $t = s-1$. We have

$$c_{T\cup\{i\}} = v(T) + v(\{i\}) - \sum_{A \subseteq T} c_A - \sum_{A: i \in A, A \subset T\cup\{i\}} c_A.$$

Thus

$$c_{T\cup\{i\}} = (v(T) - \sum_{A \subset T} c_A - c_T) + (v(\{i\}) - v(\{i\})) - \sum_{A: \{i\} \subset A, A \subset T\cup\{i\}} c_A = 0,$$

since the first parenthesis is vanishing by definition of the coefficient c_T, and the last sum is made by vanishing coefficients because of the inductive assumption. To conclude, use the linearity and the null player properties of σ^a. ■

I prove now a formula which provides the way to write the v_T games in terms of the unanimity games.

Proposition 4.1.2 *Let v_T be the family of games associated to the canonical base in $\mathbb{R}^{2^n-1}$ and u_A be the family of the unanimity games. Then the following formula holds:*

$$v_T = \sum_{k=0}^{n-t} (-1)^k \sum_{A: a=k, A \cap T = \emptyset} u_{A\cup T}. \tag{4.4}$$

Proof We need to prove that, for every coalition L, we have:

$$v_T(L) = \sum_{k=0}^{n-t}(-1)^k \sum_{A:a=k, A\cap T=\emptyset} u_{A\cup T}(L). \tag{4.5}$$

We distinguish three cases:

1. L does not contain T;
2. $L = T$;
3. $L \supset T$.

Case 1. This case is simple: all terms on the right part of equation (4.1) are vanishing, and the same is true for the left hand side;
Case 2. On the left we have 1, so we need to prove that the right hand side of equation (4.1) sums up to 1. Again, this is easy to see: for $k = 0$ one has the term $u_L(L)$ which is 1. For $k > 0$, i.e. if the coalition A is nonempty, the term $u_{L\cup A}(L)$ is vanishing;
Case 3. Finally, let us suppose L is of the form $L = T\cup H$, with H nonempty and not intersecting T. We need to prove that the sum on the the right hand side of equation (4.1) is vanishing. The term $v_{A\cup T}(H \cup T)$ is non vanishing (whence is 1), only in the case when $A \subseteq H$. Thus the right hand side becomes:

$$\sum_{k=0}^{h}(-1)^k \sum_{A:|A|=k, A\subseteq H} 1.$$

The number of coalitions A such that $|A| = k$ and $A \subseteq H$ is $\binom{h}{k}$. Thus

$$\sum_{k=0}^{h}(-1)^k \sum_{A:|A|=k, A\subseteq H} 1 = \sum_{k=0}^{h}(-1)^k \binom{h}{k} = (1-1)^h = 0.$$

■

The following proposition simplifies the above formula for symmetric indices.

Proposition 4.1.3 *Suppose, for each $s = 1, \ldots, n$, positive numbers a_s are given and suppose ϕ is a power index fulfilling the null player, linearity and symmetry axioms, and assigning a_t to all non null players in the unanimity game u_T, for all coalitions T such that $|T| = t$. Then, for a player i and for a coalition S such that $i \notin S$, it holds:*

$$\phi_i(v_{S\cup\{i\}}) = \sum_{k=0}^{n-s-1}(-1)^k \binom{n-s-1}{k} a_{s+1+k}. \tag{4.6}$$

Proof We can apply Proposition 4.1.2. Since the number of coalitions of size k, $k = 0, 1, \ldots, n-s-1$, contained in N and not intersecting $S \cup \{i\}$ is $\binom{n-s-1}{k}$, the thesis follows. ■

Theorem 4.1.1 *Let ϕ be an index fulfilling the symmetry, null player, linearity axioms and assigning a_s to all non null players in the unanimity game u_S, for all coalitions S such that $|S| = s$, where $a_1 = 1$ and $a_s > 0$ for $s = 2, \ldots, n$. Then ϕ fulfills the following formula:*

$$\phi_i(v) = \sum_{S \in 2^{N \setminus \{i\}}} \left(\sum_{k=0}^{n-s-1} (-1)^k \binom{n-s-1}{k} a_{s+k+1} \right) m_i(S). \tag{4.7}$$

Proof Writing

$$\phi_i(v) = \sum_{S \in 2^{N \setminus \{i\}}} p_i(S) m_i(v, S),$$

it holds that:

$$p_i(S) = \phi_i(v_{S \cup \{i\}}).$$

Now use Proposition 4.1.3 to conclude. ■

Remark 4.1.1 Of course, the above formula can be checked on Shapley and Banzhaf indices. In the second case, $a_{s+k} = \frac{1}{2^{s+k-1}}$ and the verification that the formula provides the probabilistic factor $\frac{1}{2^{n-1}}$ for all s is immediate. A little more involved is Shapley's case. To show that equation (4.7) provides the coefficient relative to Shapley index, one can appeal to the following general formula:

$$\sum_{k=0}^{l} \binom{l}{k} \frac{(-1)^k}{z+k} = \frac{l!}{z(z+1)\ldots(z+l)}$$

Thus, it is easy to get the formula just putting $l = n-s-1$ and $z = s+1$. ■

Theorem 4.1.2 *The* a-index σ^a *is a regular semivalue for all $a = 1, 2, \ldots$. The* 2-index *fulfills:*

$$\sigma_i^2(v) = \sum_{S \subseteq 2^{N \setminus \{i\}}} \left(\frac{s!(n-1-s)!}{n!} \sum_{k=s+1}^{n} \frac{1}{k} \right) m_i(S). \tag{4.8}$$

Proof From the formula:

$$\sum_{k=0}^{l} \binom{l}{k} \frac{(-1)^k}{z+k} = \frac{l!}{z(z+1)\dots(z+l)},$$

we easily get, by differentiating with respect to z, that

$$\sum_{k=0}^{l} \binom{l}{k} \frac{(-1)^k}{(z+k)^2} = \frac{l!}{z(z+1)\dots(z+l)} \sum_{k=0}^{l} \frac{1}{z+k}.$$

With the choice of $l = n - s - 1$ and $z = s + 1$ we then get the formula of equation (4.8). To prove that the index is probabilistic for all a, since a power index ϕ is a probabilistic value provided it is linear, fulfills the null property, assigns $v(\{i\})$ to every dummy player i and is such that all coefficients of $m_i(S)$ in equation (4.7) are positive (see [Weber (1988)]), what we need is to prove the last property. To see this, we see that the coefficients of σ^a can be obtained, when $a_s = \frac{1}{s}$, by differentiating $a - 1$ times, and taking into account that there is a change of sign (as shown when calculating the coefficient of σ^2). Thus, what we need is to show that all even derivatives of $\frac{l!}{z(z+1)\dots(z+l)}$ are positive, while the odd ones are negative. To see this, we use the fact that, given functions f_i, $i = 1, \dots, l$, the derivative

$$(f_1 \dots f_l)^{(n)} = \sum_{k_1,\dots,k_l} \binom{n}{k_1, \dots, k_l} f_1^{(k_1)} \dots f_l^{(k_l)},$$

where the summation is taken over all nonnegative integers $k_1, \dots, k_l$ such that their sum equals n. Now, setting $f_i(z) = \frac{1}{z+i-1}$, it is easy to see that if n is odd the term $f_1^{(k_1)} \dots f_l^{(k_l)}$ is negative, otherwise it is positive. ■

4.2 An application to a microarray game

I now provide an application of the σ^a indices to a microarray game. I consider, like in Lucchetti et al. (2008) [Lucchetti et al.], the microarray game defined on a tumour/normal data set published in [Alon et al. (1999)][2] and containing the expression levels of a set of 2000 genes measured using Affymetrix oligonucleotide microarray for a set of 40 tumour samples and 22 normal samples. As already remarked at the end of the previous chapter, the Banzhaf and Shapley indices provide different ranking and, moreover, the 40 genes with the highest Banzhaf value showed the same value, while the Shapley index allows a more refined analysis. In general, it can be expected

[2] microarray.princeton.edu/oncology/affydata/index.html

that the index of Banzhaf practically counts as zero the contribution of players of a winning coalition made by a large number of players. So that it is interesting to see what happens when using also the σ^a indices. We consider, in the following figures, the case of the genes common, among the first 100, to the rankings given by different indices. By the way, it is interesting to observe that this can be considered a good result, since there could be the suspect that by changing the index the ranking could be dramatically changed. Even more, when excluding the Banzhaf index the number of common genes significantly increases. Some more comment is below the figures.

4.2.1 Figures

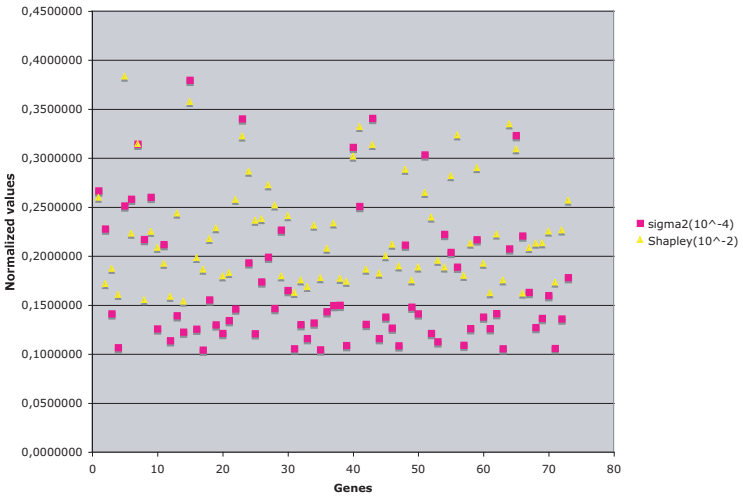


Figure 4.1: Comparison among the 73 genes with highest Shapley value and σ^2 value. Points on the same vertical line belong to the same gene.

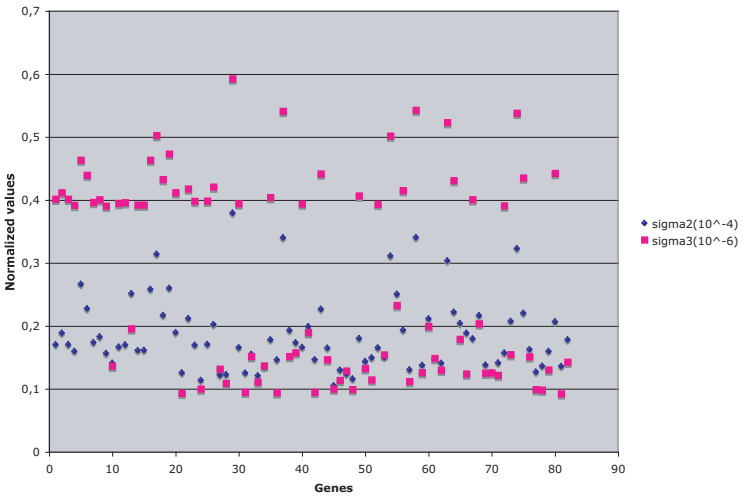


Figure 4.2: Comparison among the 56 genes with highest σ^2 value and σ^3 value. Points on the same vertical line belong to the same gene. We can observe that σ^3 tends to concentrate the genes in two strips, while σ^2 still maintains a degree of differentiation.

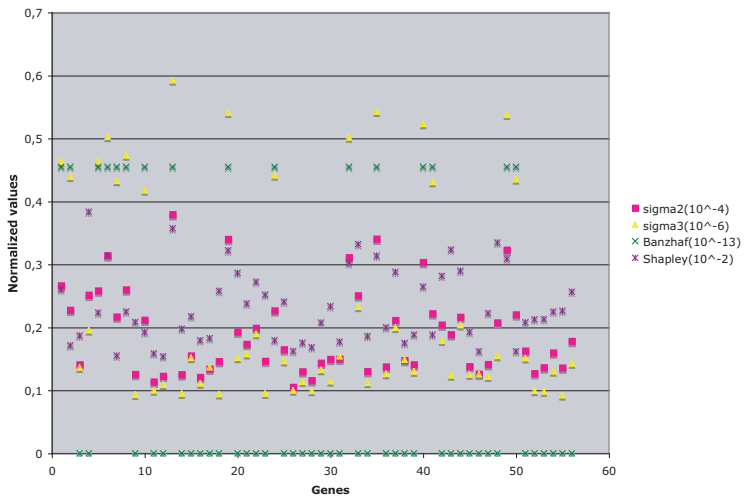


Figure 4.3: Comparison among the 56 genes with highest Shapley value, σ^2 value, σ^3 and Banzhaf value. Points on the same vertical line belong to the same gene. It is possible to appreciate the fact that the Shapley value makes possible to best differentiate genes. On the contrary, the Banzhaf value divides the genes in only two groups. The indices σ^2 and σ^2 have intermediate behavior.

4.3 Generating functions for computing power indices

In this section I present a combinatorial method based on the generating functions to compute (exactly) the power indices. The classical procedures to compute the power indices are based on the enumeration of all coalitions. If the input size of the problem is n, the function which measures the worst case running time for computing is $O(2^n)$. With the generating functions we can build algorithms to obtain these power indices with polynomial space complexity.

4.3.1 Formal power series

The *formal power series* are called formal because we ignore problems of convergence (see Stanley 1986) and we use them to have an algebraic representation of numeric successions. The formal geometric series are:

$$f(x) = \sum_{n \geq 0} f_n x^n = f_0 + f_1 x + \dots f_n x^n,$$

where f_n is a sequence on a field [3]. x^n is only a symbol that we use to point to the place of an element in a sequence.
If f_n is defined on the field $\mathbb{R}$, $f(x)$ is called *generating functions*.

A generating function approach to binomial coefficients may be obtained as follows.

Let be $S = \{x_1, x_2, \dots, x_n\}$ a set of n elements in which $x_1, x_2, \dots x_n$ are independent indeterminates. It is an immediate consequence of the process of multiplication that

$$(1 + x_1)(1 + x_2) \dots (1 + x_n) = \sum_{T \subseteq S} \prod_{x_i \in T} x_i.$$

If $T = \emptyset$ we obtain 1. If $x_i = x$ for all $i \in 1, 2, \dots n$, we have

$$(1 + x)^n = \sum_{T \subseteq S} \prod_{x \in T} x = \sum_{T \subseteq S} x^{|T|} = \sum_{k \geq 0} \binom{n}{k} x^k.$$

4.3.2 Generating Function for the Banzhaf Power Index

In this section I will consider the problem of efficiently compute the indices in the case of weighted majority games. For the special class of weighted majority games, the computational complexity is much lower, and thus the indices can be easily calculated for games with more than fifty players.

[3] To use formal series is enough that f_n is defined in a half ring.

The first result present in literature goes back to Brams-Affuso that used generating functions to compute the Banzhaf index in the case of weighted majority games. Now I briefly introduce and develop the idea, in order to apply it to the index introduced in the previous chapter.

For the weighted voting game $[q; w_1, \ldots, w_n]$ I shall denote by $w(S)$ the total weight of the coalition $S \subseteq N$:

$$w(S) = \sum_{i \in S} w_i.$$

I denote by b_k the number of coalitions whose total weight is k. I want to find the generating function of the sequence $\{b_k\}_{k \geq 0}$ because I need these coefficients to compute the number of swings of the players i. These swings for the player i are:

$$\eta_i(v) = |\{S \notin \mathcal{W} : S \cup i \in \mathcal{W}\}| = \sum_{k=q-w_i}^{q-1} b_k^i,$$

where b_k^i is the number of coalitions S with $i \notin S$ whose weight is k and $\mathcal{W}$ gives us the number of winning coalitions.

Proposition 4.3.1 *(Brams-Affuso)*
Let $[q; w_1, \ldots, w_n]$ be a weighted voting game. The generating function of the number b_k^i of coalitions S such that $i \notin S$ and $w(S) = k$, is given by:

$$f(x) = \prod_{j \neq i} (1 + x^{w_j}). \tag{4.9}$$

Proof Let $W = \{w_1, w_2 \ldots, w_n\}$. We consider the generating function

$$(1 + x^{w_1})(1 + x^{w_2}) \ldots (1 + x^{w_n}) = \sum_{V \subseteq W} \prod_{w_i \in V} x^{w_i} =$$

$$= \sum_{V \subseteq W} (x^{\sum_{w_i \in V} w_i}) =$$

$$= \sum_{k \geq 0} b_k x^k.$$

where b_k denotes the number of subsets of weights from W having total sum k. To obtain b_k^i we delete the factor $(1 + x^{w_i})$. ■

If $k = \sum_{i=1}^n w_i$ $\;0 \leq k \leq w(N)$:

$$f(x) = \prod_{j=1}^{n} (1 + x^{w_j}) = \sum_{k=0}^{w(N)} b_k x^k. \tag{4.10}$$

where b_k is the number of coalitions $S \subseteq N$ such that $w(S) = k$ that is the number of subsets S such that $w(S) = k \ \ \forall k = 0, 1, \ldots, w(N)$. There is only one subset (the empty set) such that $w(S) = 0$, so $b_0 = 1$.
How can we compute b_k in (4.10)?
We build the generating function through a sequence of multiplications.

$$f(x) = \prod_{j=1}^{n}(1 + x^{w_j}),$$

$$= (1 + x^{w_1}) \prod_{j=2}^{n}(1 + x^{w_j}),$$

$$= (1 + x^{w_1})(1 + x^{w_2}) \prod_{j=3}^{n}(1 + x^{w_j}),$$

$$= (1 + x^{w_1} + x^{w_2} + x^{w_1+w_2}) \prod_{j=3}^{n}(1 + x^{w_j}).$$

When we reach the step r with $r = 1, 2, \ldots, n$, we can write the polynomials in this way:

$$1 + b_1^{(r)}x + b_2^{(r)}x^2 + \cdots + b_w^{(r)}x^w$$

with $b_k^{(n)} = b_k \forall k$.
Now we can write $f(x)$:

$$f(x) = (1 + b_1^{(1)}x + \cdots + b_w^{(1)}x^w) \prod_{j=2}^{n}(1 + x^{w_j}) =$$

$$= (1 + b_1^{(2)}x + \cdots + b_w^{(2)}x^w) \prod_{j=3}^{n}(1 + x^{w_j}) =$$

$$\cdots$$

$$= 1 + b_1^{(n)}x + \cdots + b_w^{(n)}x^w,$$

in which $w = w(N)$.

We compute now b_k.
Let $b_k^{(0)} = 0 \ \ \forall k \neq 0$ e $b_0^{(r)} = 1$. The numbers $b_k^{(r)}$, at the step r, can be computed by means of the formula:

$$\begin{cases} b_k^{(r)} = b_k^{(r-1)} + b_{k-w_r}^{(r-1)} & \text{se } \ k = w_r, \ldots s_r, \\ b_k^{(r)} = b_k^{(r-1)} & \text{altrimenti} \end{cases} \tag{4.11}$$

where $s_r = w(\{1, 2, 3 \ldots, r\})$.
After n iterations, we will have all the coefficients b_k.

We can now use b_k to find the number of swings of each player i. To do that we look for all k such that $q - w_i \leq k < q$, the number b_k^i of coalitions S such that $i \notin S$ and we add on k. To obtain the numbers b_k^i we delete the factor $(1 + x^{w_i})$ in the generating function (4.10):

$$f(x) = (1 + b_1^i x + b_2^i x^2 + \dots b_v^i x^v)(1 + x^{w_i}) = 1 + b_1 x + \dots + b_w x^w,$$

where $v = w - w_i$. The b_k^i are expressed by the formula:

$$b_k^i = b_k - b_{k-w_i}^i \quad \forall k = 1, 2, \dots, v \tag{4.12}$$

where a coefficient with a negative index is zero.
The number of swings for the player i is:

$$\eta_i = \sum_{k=q-w_i}^{q-1} b_k^i$$

and $\beta_i = \frac{\eta_i}{2^{n-1}}$ is the Banzhaf power index of the player i.

4.3.3 Generating function for the Shapley index

David G.Cantor used generating functions to compute the Shapley index for large voting games. The Shapley index in this case can be written in the following fashion:

$$\sigma_i(v) = \sum_{S \not\in \mathcal{W}: S \bigcup i \in \mathcal{W}} \frac{s!(n-s-1)!}{n!} = \sum_{j=0}^{n-1} \frac{j!(n-j-1)!}{n!} \left(\sum_{k=q-w_i}^{q-1} a_{kj}^i \right), \tag{4.13}$$

where a_{kj}^i is the number of ways in which j players, different from player i, can sum up their weights to (exactly) k.

Proposition 4.3.2 *(Cantor)*
Let $[q; , w_1, w_2, \dots w_n]$ be a weighted voting games. The generating function of the number a_{kj}^i of coalitions S of j players with $i \notin S$ e $w(S) = k$, is given by:

$$\prod_{j \neq i} (1 + z x^{w_j}). \tag{4.14}$$

Proof Let $W = \{w_1, w_2, \dots w_n\}$ the set o the weights of all players. We consider the following generating function:

$$(1 + z x^{w_1}) \dots (1 + z w^{w_n}) = \sum_{T \subseteq W} \left(z^{|T|} x^{\sum_{w_i \in T} w_i} \right) =$$

$$= \sum_{k\geq 0}\sum_{j\geq 0} a_{kj}x^k z^j,$$

where a_{kj} is the number of coalitions of j players whose weight is k. To obtain a^i_{kj} we delete the factor $(1 + zx^{w_i})$.

■

The numbers a_{kj} are a vector whose size is $(w + 1)(n + 1)$ and we can compute them by the formula:

$$a^{(0)}_{kj} = \begin{cases} 1 & \text{per } \ j,k = 0 \\ 0 & \text{otherwise} \end{cases}$$

e, per $r = 1, 2, \ldots, n \quad j = 0, 1, \ldots, n \quad k = 0, 1, \ldots, w(N)$

$$a^{(r)}_{kj} = a^{(r-1)}_{kj} + a^{(r-1)}_{k-w_r,j-1}, \tag{4.15}$$

If an index is negative, its coefficient is zero.

The power index of the player i is obtained by the number of swings of j players with k votes.

$$a^i_{kj} = a_{kj} - a^i_{k-w_i,j-1}, \tag{4.16}$$

per $i = 1, 2, \ldots, n \quad j = 0, 1, \ldots, n-1 \quad k = 0, 1, \ldots, w(N) - w_i$.
The power index of each player is given by:

$$\sigma_i = \sum_{j=0}^{n-1} \frac{j!(n-1-j)!}{n!} \sum_{k=q-w_i}^{q-1} a^i_{kj}. \tag{4.17}$$

4.3.4 An algorithm for calculating the indices for weighted majority games

In this section we provide a formula for calculating in a fast way the indices in the case of weighted majority games, in the spirit of the previous section. Let a^i_{kj} count in how many ways the sum of the weights of j players different from i, can give k. Then the following proposition holds.

Proposition 4.3.3 *Let ϕ be an index fulfilling the symmetry, null player, linearity axioms and assigning a_s to all non null players in the unanimity game u_S, for all coalitions S such that $|S| = s$, where $a_1 = 1$ and $a_s > 0$ for $s = 2, \ldots, n$. Then the following formula holds:*

$$\phi_i(v) = \sum_{j=0}^{n-1} \left(\sum_{k=0}^{n-j-1} (-1)^k \binom{n-j-1}{k} a_{j+k+1} \right) \left(\sum_{k=q-w_i}^{q-1} a^i_{kj} \right). \tag{4.18}$$

Proof A coalition S made by j players different from i is not winning and such that $S \cup \{i\}$ is winning if and only if the sum k of weights of players in S lies between $q - w_i$ and $q - 1$. Thus the number of swings for the player i provided by coalitions of size j is exactly a^i_{kj}. The corresponding coefficient of such a coalition S is

$$p_i(S) = \sum_{k=0}^{n-j-1} (-1)^k \binom{n-j-1}{k} a_{j+k+1}.$$

Therefore the formula follows. ■

Theorem 4.3.1 *The 2-index σ^2 satisfies the following formula, valid for a weighted majority game:*

$$\sigma^2_i(v) = \sum_{S \notin \mathcal{W}: S \bigcup i \in \mathcal{W}} \frac{s!(n-s-1)!}{n!} \sum_{k=s+1}^{n} \frac{1}{k} = \sum_{j=0}^{n-1} \left(\frac{j!(n-j-1)!}{n!} \sum_{h=j+1}^{n} \frac{1}{h} \right) \left(\sum_{k=q-w_i}^{q-1} a^i_{kj} \right), \tag{4.19}$$

Proof It readily follows from equations (4.8) and (4.18). ■

Thus the problem becomes now to evaluate the term a^i_{kj}.

We have seen that the power index relative to the player i can be calculated by counting the number of swings of j players with weight equal to k.

$$a^i_{kj} = a_{kj} - a^i_{k-w_i,j-1}, \tag{4.20}$$

for $i = 1, 2, \ldots, n$; for all $j = 0, 1, \ldots, n-1$ and $k = 0, 1, \ldots, w(N) - w_i$. The (4.20) is repeated for each player.

Remark 4.3.1 The above formula applies also when some player has null weight. As already mentioned somewhere else, one reason why considering players with null weight is that sometimes it can be useful, like in microarray games, to consider averages of weighted games. In this case it can happen that in one game a player does not have positive weight. We have already introduced a simplified formula for Banzhaf's index: in that case it is not necessary to count the elements of the coalition. We can write:

$$\beta_i(v) = \sum_{S \notin \mathcal{W}: S \bigcup i \in \mathcal{W}} \frac{1}{2^{n-1}} = \frac{1}{2^{n-1}} \sum_{k=q-w_i}^{q-1} a^i_k, \tag{4.21}$$

where a_k^i is the number of coalitions of total weight k, not containing i. In this case however the formula must be changed in the following way:

$$a_k^{(0)} = 0 \text{ if } k \neq 0, a_0^{(0)} = 1, a_0^{(r)} = a_0^{(r-1)} \text{ if } w_r > 0, a_0^{(r)} = 2a_0^{(r-1)} \text{ if } w_r = 0, \tag{4.22}$$

$$a_k^{(r)} = a_k^{(r-1)} + a_{k-w_r}^{(r-1)}, \tag{4.23}$$

where a negative index implies that the corresponding coefficient vanishes. Then, the coefficient a_k^i can be easily calculated as:

$$a_k^i = a_k - a_{k-w_i}^i, \tag{4.24}$$

for $i = 1, 2, \ldots, n$; for all $j = 0, 1, \ldots, n-1$ and $k = 0, 1, \ldots, w(N) - w_i$. ■

4.4 An application of the indices: the EU Council

In this section we present an application of the indices to a very classical setting: the EU Council.

The treaty of Nice has established, after long discussions, a new weighting system for the European council. This is of course one of the most natural situations where the power indices give useful information. Here we compare the results provided by Banzhaf, Shapley, σ^2.
The set N of players is given by:

$N = \{$Malta, Latvia, Cyprus, Slovenia, Estonia, Luxembourg, Finland, Denmark,

Slovakia, Ireland, Lithuania, Sweden, Austria, Bulgaria, Belgium, CzechRepublic,

Greece, Hungary, Portugal, The Netherlands,

Romania, Spain, Poland, Germany, France, Italy, United Kingdom$\}$.

The game is defined as

$$v = [q; 3, 4, 4, 4, 4, 4, 7, 7, 7, 7, 7, 10, 10, 10, 12, 12, 12, 12, 12, 13, 14, 27, 27, 29, 29, 29, 29].$$

The quota q is $q = 255$. In the next table are displayed the following indices: Banzhaf (B), Shapley (S) and σ^2 of each state. We also show the ratio of the values (B (i)), (S (i)), ($\sigma 2(i)$) compared to the value of the state with the lower power index, Malta (MT) .

STATES	WEIGHTS	B	S	σ^2	B(i)/B(MT)	S(i)/S(MT)	$\sigma^2(i)/\sigma^2(MT)$
GE	29	0,032688	0,086738	0,02797	8,260800	10,606260	9,815720
FR	29	0,032688	0,086738	0,02797	8,260800	10,606260	9,815720
IT	29	0,032688	0,086738	0,02797	8,260800	10,606260	9,815720
UK	29	0,032688	0,086738	0,02797	8,260800	10,606260	9,815720
SP	27	0,031164	0,079975	0,025999	7,875660	9,779280	9,123380
PL	27	0,031164	0,079975	0,025999	7,875660	9,779280	9,123380
RO	14	0,017889	0,039937	0,013476	4,520850	4,883468	4,729164
NL	13	0,016691	0,036825	0,012476	4,218090	4,502940	4,378370
BE	12	0,015475	0,034068	0,011555	3,910791	4,165810	4,055050
CZ	12	0,015475	0,034068	0,011555	3,910791	4,165810	4,055050
GR	12	0,015475	0,034068	0,011555	3,910791	4,165810	4,055050
HU	12	0,015475	0,034068	0,011555	3,910791	4,165810	4,055050
PT	12	0,015475	0,034068	0,011555	3,910791	4,165810	4,055050
SE	10	0,012989	0,028193	0,00961	3,282540	3,447420	3,372390
AU	10	0,012989	0,028193	0,00961	3,282540	3,447420	3,372390
BG	10	0,012989	0,028193	0,00961	3,282540	3,447420	3,372390
FI	7	0,00916	0,019606	0,006721	2,314890	2,397410	2,358600
DK	7	0,00916	0,019606	0,006721	2,314890	2,397410	2,358600
SK	7	0,00916	0,019606	0,006721	2,314890	2,397410	2,358600
IR	7	0,00916	0,019606	0,006721	2,314890	2,397410	2,358600
LT	7	0,00916	0,019606	0,006721	2,314890	2,397410	2,358600
LV	4	0,005251	0,011042	0,003813	1,327020	1,350210	1,338030
CY	4	0,005251	0,011042	0,003813	1,327020	1,350210	1,338030
SLO	4	0,005251	0,011042	0,003813	1,327020	1,350210	1,338030
ES	4	0,005251	0,011042	0,003813	1,327020	1,350210	1,338030
LU	4	0,005251	0,011042	0,003813	1,327020	1,350210	1,338030
MT	3	0,003957	0,008178	0,00285	1,000000	1,000000	1,000000

As a comment, we can see that σ^2 is intermediate between Shapley, which enhances the power of more powerful players, and Banzhaf, which tends to diminish that power. in this sense, σ^2 looks like a compromise between the two.

Chapter 5

Weighted Indices

Looking at the data elaborated in the previous chapters, we can notice that the power indices usually have difficulties in distinguishing the genes. This can be structural, in that few samples versus so many genes can cause the fact that several of them could be grouped in families of symmetric players. On the other hand, it is quite possible that round off errors do not allow evaluating very small differences. By elaborating some (real) data it turned out that some patients presented around 200 genes abnormally expressed. In such a case, the Banzhaf index is simply useless, since it attaches to each gene approximatively the value $\frac{1}{2^{200}}$. I.e., zero for any computer. This means that actually the patient does not provide useful data, since it considers all genes as null genes. This in principle cannot be considered totally useless: in some sense, it indicates that the patient could be considered not meaningful for the analysis, since its abnormally expressed genes are too many. But on the other hand, especially when treating data with few patients, it is of interest to avoid the risk of having a partition of the set of genes made by few elements (i.e. few subsets with a large number of genes). For this reason, it is interesting to try to better differentiate the contribution that each gene could give to the disease.

Thus, it seems to be a promising idea to try to differentiate the genes, by considering indices better differentiating the contribution of the players. This in some sense goes in the opposite direction with respect to what we did in the previous chapter, but it is natural to think that both approaches make sense, for several reasons: one of them, the great variety of data- sets available in literature. It is conceivable that having the data of many patients it is enough to use the Shapley to differentiate the genes in a significant way, while with few data probably different indices are needed. Moreover, the data available are not homogeneous: actually sometimes patients do present very many abnormally expressed genes while in other cases do not. In short, the idea is that the great variety of the data set justifies the idea of having different indices.

This chapter thus deals with the introduction of a variant of the microarray game, in order to use the so called *weighted indices*. Then, we consider a new model of game, derived from the results of the (modified) microarray game. In short, we consider a weighted majority game, by considering a much restricted set of genes, selected by means of the ranking of the indices. It is clear that all of this must be considered, at the current state of the art, only experimental. Several facts do not have, at the present, strong theoretical motivations. For instance, which index should be used to select a group of genes to analyze further with the weighted majority game. Then, how many genes should be used in the subsequent game. Of course, we must take into account the complexity of the calculations. Fortunately, as we have seen, for this type of games the evaluation of the indices is much easier, thanks to the algorithm presented above. However, it makes no sense, and it is impossible, to consider thousand of players. Furthermore, it is not clear how to attach weights to the players, and it is not clear too what should be the majority quota.
Despite the above remarks, we believe that these data are interesting, at least for one strong reason: it seems that there is some form of stability on the experimental results. Looking at the first 100 ranked genes with respect to the various indices, we find that a great percentage of them are present in all ranking made the different indices (with the exception of Banzhaf's, for the reason explained above that it does not differentiate enough the genes). Thus we shall perform the subsequent weighted majority game with the genes we find in the intersection of the rankings made by the indices.
Finally, a check made in the medical literature showed that some of the selected genes by our methods in particular experiments are considered to be of great importance from the medical point of view, in the onset of the considered disease. Thus as a conclusion we believe that further interactions with researchers in medical groups should be enhanced in order to suggest new development of this approach.

5.1 An extended version of the microarray game

The idea underlying the new version of the microarray game is to allow the matrix at the core of the game to contain not only zeroes and ones. In other words, we do not classify the genes only in two big categories, normally and abnormally expressed, but we take also into account "how much" the genes are abnormally expressed, by giving them a weight gradually increasing depending on how much the gene is far from the normality interval. Of course, this can be done in several ways. A natural one is to consider, for each gene i, the normality interval, let us call it $N_i = [m_i, M_i]$(where m_i and M_i are respectively the minimum and maximum value of genes in the expression of the genes in the reference group), to evaluate the standard

deviation s_i relative to the data of the gene, to set $N_i^k = [m_i - ks_i, M_i + ks_i]$, $k = 1, \ldots, n$, and to assign the value k to the gene falling in the set $N_i^k \setminus N_i^{k-1}$ (n if it falls outside all these sets). In this way, we can rank the genes according to another type of index, called in the literature *weighted index*.

Thus, suppose we are given a $n \times m$ matrix M such that $m_{ij} \geq 0$ for all i, j. Observe that when M represents a classical microarray game, i.e. $m_{ij} \in \{0, 1\}$, due to the equal splitting property the Shapley index of the player i fulfills the formula

$$\sigma_i(v) = \frac{1}{m} \sum_{j=1}^{m} \frac{m_{ij}}{\sum_{i=1}^{n} m_{ij}}.$$

It seems to be very natural then, to use exactly the same formula also when the coefficient m_{ij} is not only valued in $\{0, 1\}$. It turns out that the index so obtained is already known on the literature, since the resulting index is the so called *weighted (Shapley) index*. We address the interested reader to the survey article [Kalai and Samet (1987)], for more about these indices.

Our first attempts of processing data showed a kind of stability with respect to the ranking of the genes, even though, as expected, taking into account more intervals resulted in a better differentiation of the genes. Thus, we decided to avoid binding the number of intervals, in order to have a more fragmented ranking between the genes.
The (extended) microarray matrix well serves also to build a weighted majority game. Quite naturally, the weight of the player i in the game j is given by the coefficient m_{ij}.
I want to add one remark. Even if the various indices give the same ranking in weighted games (this is well known), this is no longer true in microarray games, as simple examples show. Thus in the microarray game also the ranking between genes, and not only the ratio of the power of the players is relevant. Nevertheless, I notice that again a strong stability is shown in the ranking of the genes, as far as we use different indices.

To conclude, I perform some tests with different data sets, i.e. Stroma Rich and Stroma Poor Neuroblastic tumours, Ductal and Lobular breast tumour, two different types of Colon tumour.

5.2 Data analysis

5.2.1 Data from early onset colon rectal cancer

Gene expression analysis was performed by using Human Genome U133A-Plus 2.0 GeneChip arrays (Affymetrix, Inc., Calif). This data set contains 10 healthy samples and 12 derived from tumour tissues. In the following table

we can see the ranking of the first 50 genes labelled with the Shapley value (SY) by using the weighted indices and Banzhaf power index (WMGBa) and Shapley power index (WMGSh) of the same 50 genes after they have played a weighted majority game. The weights in the game are the values of n that we have given to the genes in the cooperative game played with the weighted indices, the share q is $50\% + 1$.

	SY	Power index	WMGBa	Power index	WMGSh	Power index		SY	Power index	WMGBa	Power index	WMGSh	Power index
1	FOSB	0,013406	FOSB	0,43783	FOSB	0,25665	30	TAGLN	0,00068224	JUN	0,026731	RHOB	0,0095588
2	CYR61	0,010749	CYR61	0,4308	CYR61	0,164	31	MGC52498	0,00066903	AI492388	0,0266	TCEAL7	0,0090184
3	FOS	0,0040232	FXYD6	0,14946	FOS	0,046882	32	FABP4	0,00066456	TUBA1A	0,026464	AI492388	0,0087509
4	SFRP2	0,0030742	FOS	0,12135	FXYD6	0,046606	33	PLN	0,00066026	TUBB6	0,02621	MLLT11	0,0085647
5	VIP	0,0028325	VIP	0,11096	VIP	0,039589	34	KRT24	0,00065687	AHNAK2	0,025788	TUBA1A	0,0085303
6	FXYD6	0,0027964	SFRP2	0,062859	SFRP2	0,022553	35	MLLT11	0,00065618	TCEAL7	0,025705	MGP	0,0082503
7	ADAMTS1	0,0013196	DMN	0,061413	DMN	0,018164	36	TPM2	0,0006431	MLLT11	0,025094	TUBB6	0,008098
8	EGR1	0,0012059	DES	0,045702	S59049	0,015231	37	ATF3	0,00062799	HSPB6	0,024802	HSPB6	0,0078001
9	S59049	0,0010926	S59049	0,043233	EGR1	0,014095	38	GAL	0,00062255	MGP	0,023548	AHNAK2	0,0076697
10	DMN	0,0010618	CNN1	0,04316	DES	0,013667	39	PCP4	0,00061682	PDLIM7	0,022843	NR4A2	0,0075324
11	CTGF	0,0010598	MYL9	0,041688	ADAMTS1	0,013292	40	TCEAL7	0,00061188	NR4A2	0,021293	CCDC3	0,0071738
12	PRPH	0,0010598	RBPMS2	0,040835	RERGL	0,01297	41	AA889653	0,00060942	ADIPOQ	0,019666	PDLIM7	0,0070993
13	RERGL	0,00099813	AI969945	0,040284	MGC52498	0,01287	42	MAB21L2	0,00058727	CCDC3	0,019588	ADIPOQ	0,0070134
14	MGP	0,00094337	EGR1	0,039211	CNN1	0,012751	43	W72348	0,00058676	GAL	0,01923	GAL	0,0068292
15	DUSP1	0,0009388	ADAMTS1	0,037699	RBPMS2	0,01273	44	PDLIM7	0,00058529	FILIP1L	0,018355	FILIP1L	0,0061772
16	ADIPOQ	0,00091756	TPM2	0,036275	MYL9	0,012595	45	HSPB6	0,00058023	ATF3	0,017132	ATF3	0,0059581
17	JUN	0,00088653	PCP4	0,035583	AI969945	0,012539	46	TUBA1A	0,00057738	KRT24	0,015136	KRT24	0,0056429
18	hCG_1776018	0,00084162	AA889653	0,035282	CTGF	0,012041	47	AHNAK2	0,0005708	FABP4	0,01425	FABP4	0,0050061
19	AI969945	0,00083569	CTGF	0,035238	PRPH	0,011947	48	FILIP1L	0,00056639	BE044614	0,011158	BE044614	0,0039619
20	AI492388	0,00082685	MGC52498	0,034634	hCG_1776018	0,011142	49	TUBB6	0,00056406	W72348	0,0078261	W72348	0,0030981
21	SCG2	0,00082385	RERGL	0,034025	TPM2	0,010931	50	BE044614	0,00055078	BC038379	0,0074452	BC038379	0,002648
22	DES	0,00080315	MAB21L2	0,033892	SCG2	0,010925							
23	RBPMS2	0,00079997	PRPH	0,033589	DUSP1	0,010648							
24	MYL9	0,00078826	TAGLN	0,032999	AA889653	0,010379							
25	NR4A2	0,00076649	PLN	0,03249	PCP4	0,010292							
26	CNN1	0,00075588	hCG_1776018	0,030335	PLN	0,010142							
27	RHOB	0,00074379	SCG2	0,029826	TAGLN	0,010129							
28	BC038379	0,00070971	RHOB	0,028782	JUN	0,0099649							
29	CCDC3	0,00070223	DUSP1	0,028085	MAB21L2	0,0098938							

Figure 5.1: Early Onset Colon rectal Cancer. I take the first 50 genes classified by Shapley value in weighted indices. The Banzhaf, Shapley values in the weighted majority game are displayed.

Remark 5.2.1 Seven genes, CYR61, UCHL1, FOS, FOSB, EGR1, VIP, KRT24, all present in our rankings function in a multitude of biological processes ranging from transcription , angiogenesis, adhesion and inflammatory regulation to protein catabolism in various cellular compartments, from extracellular to the nucleo. The over expression of these was already identified as a potentially prediction of early onset colorectal cancer([Yi Hong et al.(2007)]). These genes are all present in our rankings and show a certain stability. ■

FOSB	0,43783	AI492388	0,0266
CYR61	0,4308	TUBA1A	0,026464
FXYD6	0,14946	TUBB6	0,02621
FOS	0,12135	AHNAK2	0,025788
VIP	0,11096	TCEAL7	0,025705
SFRP2	0,062859	MLLT11	0,025094
DMN	0,061413	HSPB6	0,024802
DES	0,045702	MGP	0,023548
S59049	0,043233	PDLIM7	0,022843
CNN1	0,04316	NR4A2	0,021293
MYL9	0,041688	ADIPOQ	0,019666
RBPMS2	0,040835	CCDC3	0,019588
AI969945	0,040284	GAL	0,01923
EGR1	0,039211	FILIP1L	0,018355
ADAMTS1	0,037699	ATF3	0,017132
TPM2	0,036275	KRT24	0,015136
PCP4	0,035583	FABP4	0,01425
AA889653	0,035282	BE044614	0,011158
CTGF	0,035238	W72348	0,007826
MGC52498	0,034634	BC038379	0,007445
RERGL	0,034025		
MAB21L2	0,033892		
PRPH	0,033589		
TAGLN	0,032999		
PLN	0,03249		
hCG_1776018	0,030335		
SCG2	0,029826		
RHOB	0,028782		
DUSP1	0,028085		
JUN	0,026731		

Figure 5.2: Early Onset Colon rectal Cancer. I take the first 50 genes classified by Shapley value using the weighted indices, and the Banzhaf value in the weighted majority game is displayed.

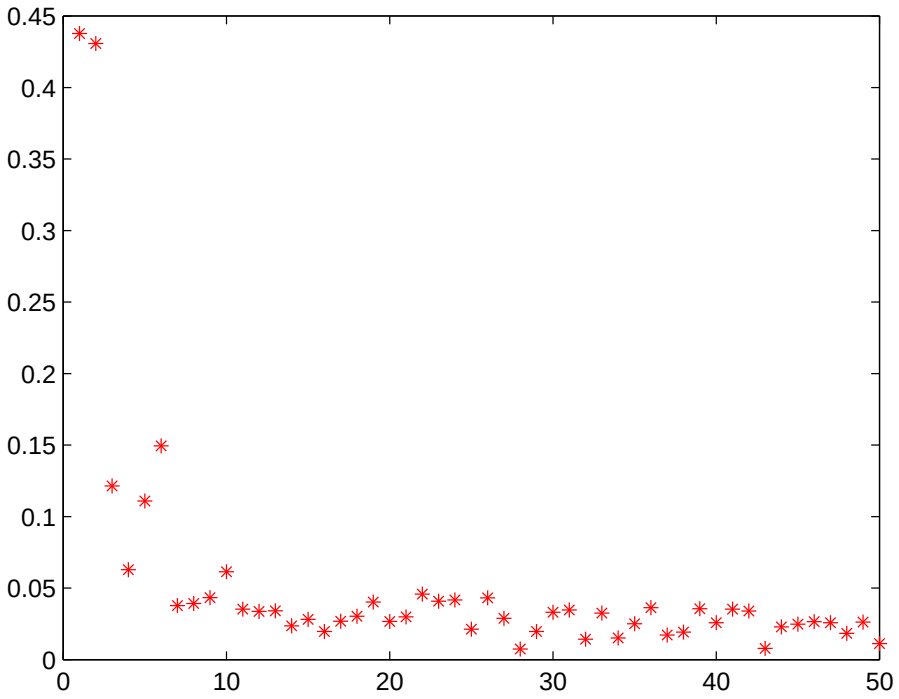


Figure 5.3: Early Onset Colon rectal Cancer. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Banzhaf power index.

FOSB	0,25665	TCEAL7	0,009018
CYR61	0,164	AI492388	0,008751
FOS	0,046882	MLLT11	0,008565
FXYD6	0,046606	TUBA1A	0,00853
VIP	0,039589	MGP	0,00825
SFRP2	0,022553	TUBB6	0,008098
DMN	0,018164	HSPB6	0,0078
S59049	0,015231	AHNAK2	0,00767
EGR1	0,014095	NR4A2	0,007532
DES	0,013667	CCDC3	0,007174
ADAMTS1	0,013292	PDLIM7	0,007099
RERGL	0,01297	ADIPOQ	0,007013
MGC52498	0,01287	GAL	0,006829
CNN1	0,012751	FILIP1L	0,006177
RBPMS2	0,01273	ATF3	0,005958
MYL9	0,012595	KRT24	0,005643
AI969945	0,012539	FABP4	0,005006
CTGF	0,012041	BE044614	0,003962
PRPH	0,011947	W72348	0,003098
hCG_1776018	0,011142	BC038379	0,002648
TPM2	0,010931		
SCG2	0,010925		
DUSP1	0,010648		
AA889653	0,010379		
PCP4	0,010292		
PLN	0,010142		
TAGLN	0,010129		
JUN	0,009965		
MAB21L2	0,009894		
RHOB	0,009559		

Figure 5.4: Early Onset Colon rectal Cancer. Shapley value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

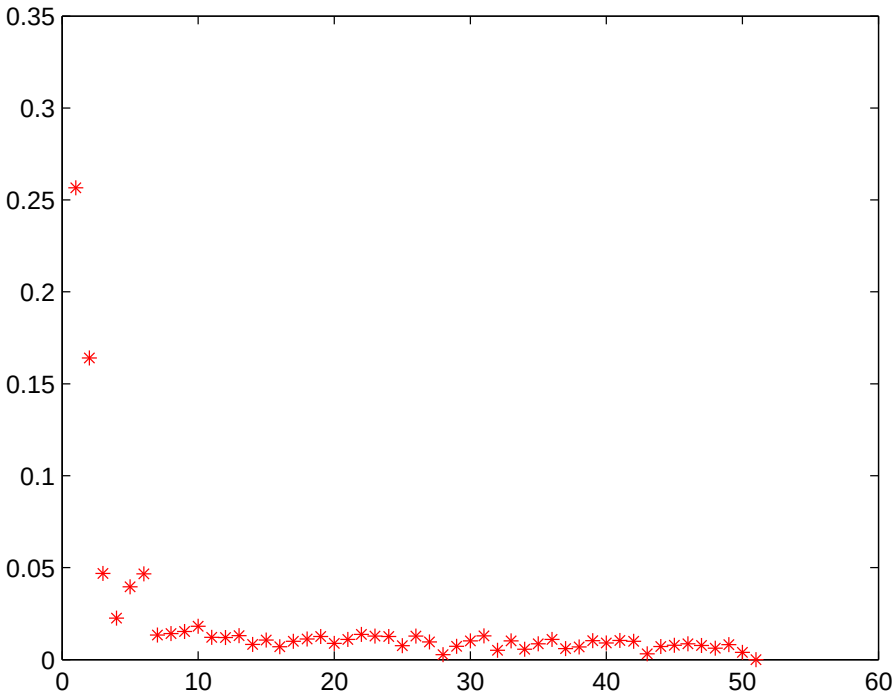


Figure 5.5: Early Onset Colon rectal Cancer. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Shapley power index.

5.2.2 Data from neuroblastic tumours

Gene expression analysis was performed by using Human Genome U133A GeneChip arrays (Affymetrix, Inc., Calif). This data set contains 10 healthy samples and 9 derived from tumour tissues of two different types of tumour: neuroblastic tumours stroma poor (SP) and neuroblastic tumours stroma rich (SR). In this case we have highlighted genes over or under expressed in a form of cancer compared to the other one. The following tables contain the ranking of the first 50 genes labelled with the Shapley value (SY) by using the weighted indices. The first one contains the ranking obtained analyzing data related to the tumour SR compared with SP used to identify the range of normality (in the table marked with SR / SP), the second one the ranking obtained analyzing data related to the tumour SP compared with SR used to identify the range of normality (in the table marked with SP / SR). As in the previous section, in the tables we can see also the Banzhaf power index (WMGBa) and Shapley power index (WMGSh) of the same 50 genes after they have played a weighted majority game. The weights in the game are the values of n that we have given to the genes in the cooperative game played with the weighted indices, the share q is $50\% + 1$.

	SY	Power index	WMGBa	Power index	WMGSh	Power index	SY	Power index	WMGBa	Power index	WMGSh	Power index
1	ITGB8	0,0021972	ITGB8	0,18111	ITGB8	0,037546	30 NBLA00301	0,0010748	UTS2	0,081471	ANGPTL7	0,017831
2	PDZRN4	0,0021116	PDZRN4	0,16469	PDZRN4	0,03405	31 C1orf76	0,00098634	C1orf76	0,080401	DAG1	0,017829
3	PMP2	0,0018277	CALCA	0,14958	PMP2	0,030634	32 TFPI2	0,00098564	ADAM28	0,079995	ISL1	0,017514
4	NDP	0,0018028	PMP2	0,14606	TCL1A	0,03033	33 GFAP	0,00097911	STAP1	0,079337	TFPI2	0,017371
5	CALCA	0,0017725	NDP	0,14591	NDP	0,030329	34 ITPR3	0,00096842	CYP4B1	0,077544	MAST1	0,017027
6	SERPINA3	0,0017664	SERPINA3	0,13877	CALCA	0,030234	35 PLEKHB1	0,00096679	PRKD1	0,077529	UTS2	0,016355
7	KLF5	0,0016495	TCL1A	0,12628	SERPINA3	0,028513	36 PRKD1	0,00093585	MAST1	0,077351	PRKD1	0,016305
8	MGC39900	0,0015321	KLF5	0,12586	MGC39900	0,027313	37 AASS	0,00093032	PLEKHB1	0,076944	ITPR3	0,015985
9	TMSL8	0,0015321	MGC39900	0,12486	TMSL8	0,027313	38 MPZ	0,00092516	ITPR3	0,07673	VGLL3	0,015884
10	ST6GALNAC2	0,0015132	TMSL8	0,12486	KLF5	0,026083	39 MAST1	0,00092437	ISL1	0,076487	AASS	0,015765
11	TNNC1	0,001484	ST6GALNAC2	0,11963	CDH1	0,02521	40 VGLL3	0,0009126	VGLL3	0,07569	PLEKHB1	0,015705
12	CAB39L	0,0014384	CDH1	0,11656	ST6GALNAC2	0,024818	41 ADAM28	0,00090855	AASS	0,074856	CYP4B1	0,01558
13	TCL1A	0,0013945	CAB39L	0,11536	CAB39L	0,024136	42 OLFM4	0,00090614	GFAP	0,073608	GFAP	0,015209
14	CDH1	0,0013833	TNNC1	0,11472	TNNC1	0,023777	43 GAS7	0,00089997	MPZ	0,072223	MPZ	0,014787
15	ALLC	0,0013368	ADAMTS8	0,10098	NBLA00301	0,02067	44 STAP1	0,00089545	MGC87042	0,070209	MGC87042	0,014595
16	ADAMTS8	0,0013115	ALLC	0,098107	ADAMTS8	0,020564	45 CAPN6	0,00089392	GAS7	0,069895	HNT	0,014275
17	CTDSPL	0,0012303	MBP	0,095578	TSPAN8	0,020059	46 ISL1	0,00089056	HNT	0,06977	GAS7	0,014243
18	TSPAN8	0,001195	CTDSPL	0,095135	MBP	0,020053	47 HNT	0,00088862	CAPN6	0,06863	CAPN6	0,014133
19	MBP	0,0011811	TSPAN8	0,094831	ALLC	0,019592	48 MGC87042	0,00088692	OLFM4	0,068557	OLFM4	0,013892
20	NGFR	0,0011669	NBLA00301	0,090417	STAP1	0,01959	49 COBL	0,00086458	COBL	0,067226	COBL	0,013666
21	SEMA3B	0,0011512	NGFR	0,090173	CTDSPL	0,019462	50 FOXD1	0,00084731	FOXD1	0,067212	FOXD1	0,012596
22	SLC22A3	0,0011482	SEMA3B	0,08814	CIITA	0,018531						
23	DAG1	0,0011257	LGI1	0,088138	SLC22A3	0,018285						
24	ANGPTL7	0,0011224	SLC22A3	0,088011	NGFR	0,018267						
25	LGI1	0,0011058	DAG1	0,087284	LGI1	0,018198						
26	SDC4	0,0010914	SDC4	0,087148	SEMA3B	0,018169						
27	CIITA	0,001085	ANGPTL7	0,086316	SDC4	0,018144						
28	UTS2	0,00108	CIITA	0,085446	C1orf76	0,018138						
29	CYP4B1	0,0010753	TFPI2	0,082401	ADAM28	0,01804						

Figure 5.6: Neuroblastic tumour: SR/SP. I take the first 50 genes classified by Shapley value in weighted indices. The Banzhaf, Shapley values in the weighted majority game are displayed.

Remark 5.2.2 Eight genes, ANGPTL7, PMP2, TSPAN8, CENPF, EYA1,PBK, TOP2A, TFAP2B are present in our rankings and five of them (CENPF, EYA1,PBK, TOP2A, TFAP2B) encode for nuclear proteins. The over expression of these genes was already identified in([Albino et al. (2008)]). These genes show a certain stability in our rankings.

■

ITGB8	0,18111	NBLA00301	0,090417
PDZRN4	0,16469	C1orf76	0,080401
PMP2	0,14606	TFPI2	0,082401
NDP	0,14591	GFAP	0,073608
CALCA	0,14958	ITPR3	0,07673
SERPINA3	0,13877	PLEKHB1	0,076944
KLF5	0,12586	PRKD1	0,077529
MGC39900	0,12486	AASS	0,074856
TMSL8	0,12486	MPZ	0,072223
ST6GALNAC2	0,11963	MAST1	0,077351
TNNC1	0,11472	VGLL3	0,07569
CAB39L	0,11536	ADAM28	0,079995
TCL1A	0,12628	OLFM4	0,068557
CDH1	0,11656	GAS7	0,069895
ALLC	0,098107	STAP1	0,079337
ADAMTS8	0,10098	CAPN6	0,06863
CTDSPL	0,095135	ISL1	0,076487
TSPAN8	0,094831	HNT	0,06977
MBP	0,095578	MGC87042	0,070209
NGFR	0,090173	COBL	0,067226
SEMA3B	0,08814	FOXD1	0,067212
SLC22A3	0,088011		
DAG1	0,087284		
ANGPTL7	0,086316		
LGI1	0,088138		
SDC4	0,087148		
CIITA	0,085446		
UTS2	0,081471		
CYP4B1	0,077544		

Figure 5.7: Neuroblastic tumour: SR/SP. Banzhaf value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

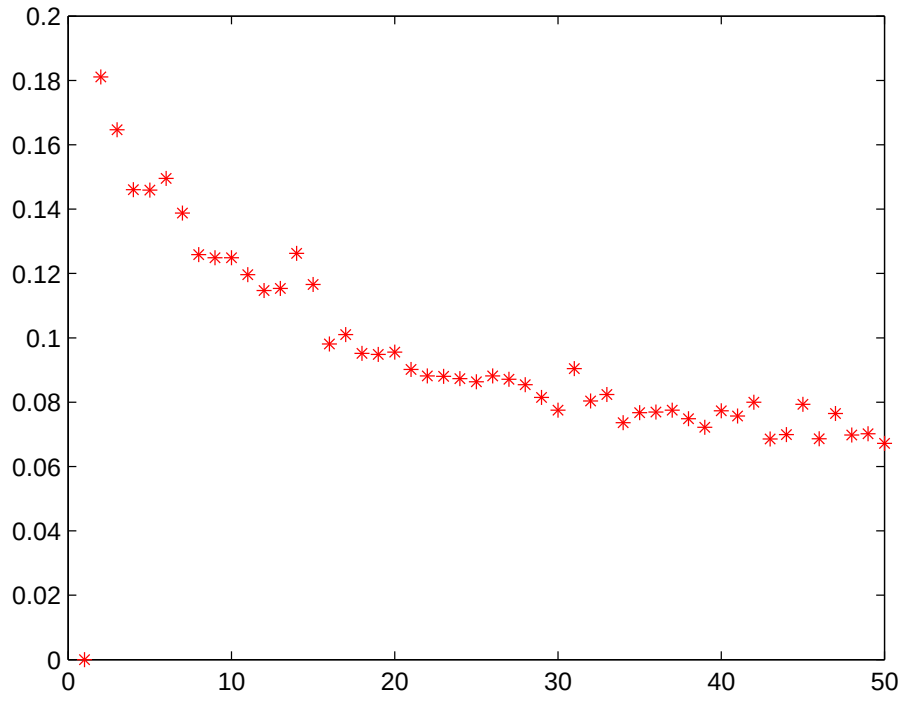


Figure 5.8: Neuroblastic tumour: SR/SP. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Banzhaf power index.

ITGB8	0,037546	NBLA00301	0,02067
PDZRN4	0,03405	C1orf76	0,018138
PMP2	0,030634	TFPI2	0,017371
NDP	0,030329	GFAP	0,015209
CALCA	0,030234	ITPR3	0,015985
SERPINA3	0,028513	PLEKHB1	0,015705
KLF5	0,026083	PRKD1	0,016305
MGC39900	0,027313	AASS	0,015765
TMSL8	0,027313	MPZ	0,014787
ST6GALNAC2	0,024818	MAST1	0,017027
TNNC1	0,023777	VGLL3	0,015884
CAB39L	0,024136	ADAM28	0,01804
TCL1A	0,03033	OLFM4	0,013892
CDH1	0,02521	GAS7	0,014243
ALLC	0,019592	STAP1	0,01959
ADAMTS8	0,020564	CAPN6	0,014133
CTDSPL	0,019462	ISL1	0,017514
TSPAN8	0,020059	HNT	0,014275
MBP	0,020053	MGC87042	0,014595
NGFR	0,018267	COBL	0,013666
SEMA3B	0,018169	FOXD1	0,012596
SLC22A3	0,018285		
DAG1	0,017829		
ANGPTL7	0,017831		
LGI1	0,018198		
SDC4	0,018144		
CIITA	0,018531		
UTS2	0,016355		
CYP4B1	0,01558		

Figure 5.9: Neuroblastic tumour: SR/SP. Shapley value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

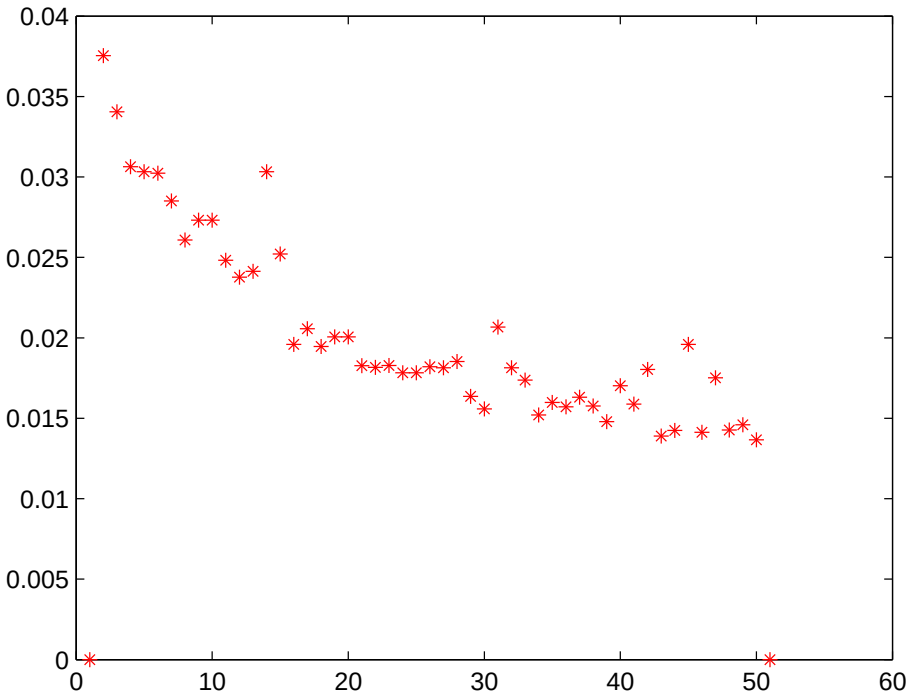


Figure 5.10: Neuroblastic tumour: SR/SP. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Shapley power index.

	SY	Power index	WMGBa	Power index	WMGSh	Power index		SY	Power index	WMGBa	Power index	WMGSh	Power index
1	UBE2C	0,0029603	UBE2C	0,37953	UBE2C	0,076685	29	MGC39900	0,00069808	CENPA	0,085599	TMSL8	0,017562
2	VASH2	0,0011515	VASH2	0,14108	VASH2	0,029082	30	TMSL8	0,00069808	TFAP2B	0,08508	TFAP2B	0,017384
3	CENPF	0,001052	CENPF	0,1264	CENPF	0,026105	31	CDC20	0,00069502	KLHL23	0,083241	KLHL23	0,01707
4	MCM10	0,0010153	CXXC4	0,12143	CXXC4	0,024891	32	MLF1IP	0,00069459	GDAP1L1	0,082828	GDAP1L1	0,017033
5	CXXC4	0,00097964	TGFBR2	0,11617	TGFBR2	0,023944	33	MCM4	0,00069399	MCM4	0,081924	MCM4	0,01691
6	BIRC5	0,0009709	ZNF821	0,11381	BIRC5	0,023378	34	TUSC4	0,00068942	CDC20	0,081875	CDC20	0,016876
7	FOXM1	0,00097027	BIRC5	0,11283	ZNF821	0,023179	35	FAM64A	0,00068869	LPAR2	0,081498	HJURP	0,016797
8	ZNF821	0,00090364	FOXM1	0,11187	FOXM1	0,023158	36	GDAP1L1	0,00068173	FAM64A	0,081348	MLF1IP	0,016775
9	NCAPG	0,00087407	MCM10	0,10977	MCM10	0,022801	37	ARID3B	0,00066655	MLF1IP	0,081238	FAM64A	0,016695
10	TGFBR2	0,00087197	DCN	0,10288	NCAPG	0,021348	38	GRM8	0,00066268	HJURP	0,080887	LPAR2	0,016684
11	TPX2	0,00086789	NCAPG	0,1026	DCN	0,021125	39	TFAP2B	0,00066144	ICA1	0,080251	KIF22	0,016577
12	MKI67	0,00083887	TPX2	0,099536	TPX2	0,020627	40	GTSE1	0,00065314	KIF22	0,079968	ICA1	0,01653
13	ESPL1	0,00083773	ESPL1	0,09937	ESPL1	0,020428	41	MMP12	0,00065187	ARID3B	0,079597	ARID3B	0,016269
14	KIF20A	0,00082464	IGF2BP3	0,097603	KIF20A	0,020081	42	ICA1	0,00064484	FLJ22184	0,078721	FLJ22184	0,016175
15	EIF4EBP1	0,00082064	KIF20A	0,097162	IGF2BP3	0,020061	43	GINS2	0,00063776	GRM8	0,078091	GRM8	0,016036
16	IGF2BP3	0,00080548	EIF4EBP1	0,094094	EIF4EBP1	0,01972	44	LPAR2	0,00063766	GTSE1	0,077673	GTSE1	0,01602
17	DTL	0,00079153	FEV	0,09358	MKI67	0,019426	45	FLJ22184	0,00063747	GINS2	0,077582	MMP12	0,015909
18	TTK	0,00077759	MKI67	0,093233	FEV	0,019218	46	PXMP2	0,00063551	MMP12	0,077217	GINS2	0,015904
19	PBK	0,00077254	LOC157627	0,092917	DTL	0,019036	47	TRAP1	0,00063517	ARC	0,077127	TRAP1	0,015742
20	FEV	0,00076269	EYA1	0,092815	EYA1	0,01902	48	KIF14	0,00063496	TRAP1	0,075046	ARC	0,015642
21	PHOX2A	0,00075755	DTL	0,092507	LOC157627	0,018996	49	ARC	0,00063075	KIF14	0,074853	KIF14	0,015321
22	DCN	0,0007492	TTK	0,091287	TTK	0,01888	50	PXMP2	0,00063551	PXMP2	0,074801	PXMP2	0,015263
23	TOP2A	0,00073944	PHOX2A	0,089891	PHOX2A	0,01842							
24	LOC157627	0,00073602	PBK	0,089613	PBK	0,018411							
25	EYA1	0,00073289	TOP2A	0,087734	TOP2A	0,018001							
26	CENPA	0,00073287	TUSC4	0,086041	CENPA	0,017678							
27	HJURP	0,00072763	MGC39900	0,085727	TUSC4	0,017564							
28	KIF22	0,0007002	TMSL8	0,085727	MGC39900	0,017562							

Figure 5.11: Neuroblastic tumour: SP/SR. I take the first 50 genes classified by Shapley value in weighted indices. The Banzhaf, Shapley values in the weighted majority game are displayed.

UBE2C	0,37953	CDC20	0,081875
VASH2	0,14108	MLF1IP	0,081238
CENPF	0,1264	MCM4	0,081924
MCM10	0,10977	TUSC4	0,086041
CXXC4	0,12143	FAM64A	0,081348
BIRC5	0,11283	GDAP1L1	0,082828
FOXM1	0,11187	ARID3B	0,079597
ZNF821	0,11381	GRM8	0,078091
NCAPG	0,1026	TFAP2B	0,08508
TGFBR2	0,11617	GTSE1	0,077673
TPX2	0,099536	MMP12	0,077217
MKI67	0,093233	ICA1	0,080251
ESPL1	0,09937	GINS2	0,077582
KIF20A	0,097162	LPAR2	0,081498
EIF4EBP1	0,094094	FLJ22184	0,078721
IGF2BP3	0,097603	PXMP2	0,074801
DTL	0,092507	TRAP1	0,075046
TTK	0,091287	KIF14	0,074853
PBK	0,089613	ARC	0,077127
FEV	0,09358	KLHL23	0,083241
PHOX2A	0,089891		
DCN	0,10288		
TOP2A	0,087734		
LOC157627	0,092917		
EYA1	0,092815		
CENPA	0,085599		
HJURP	0,080887		
KIF22	0,079968		
MGC39900	0,085727		
TMSL8	0,085727		

Figure 5.12: Neuroblastic tumour: SP/SR.Banzhaf value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

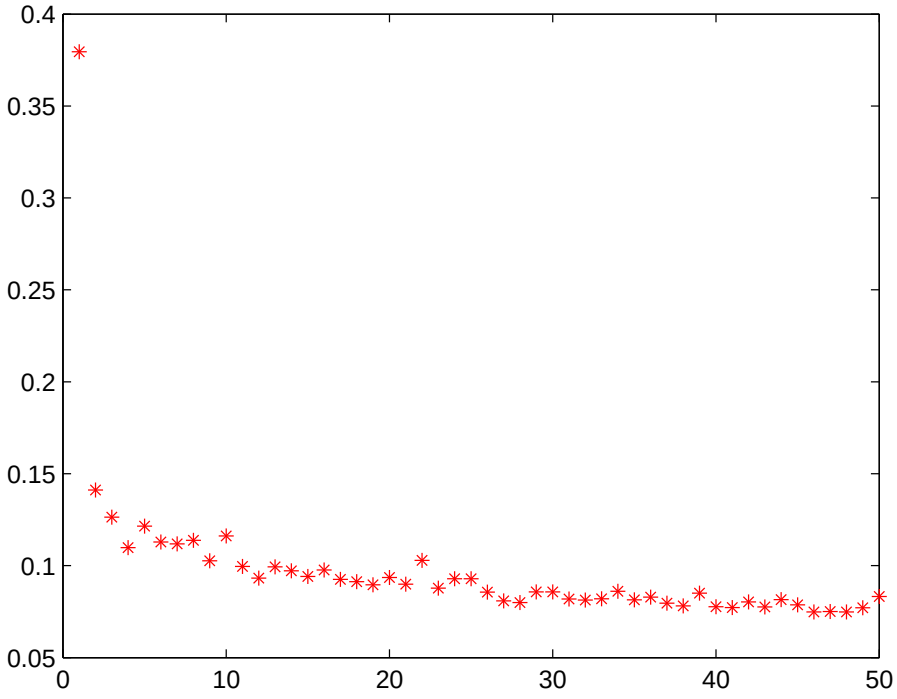


Figure 5.13: Neuroblastic tumour: SP/SR. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Banzhaf power index.

UBE2C	0,076685	CDC20	0,016876
VASH2	0,029082	MLF1IP	0,016775
CENPF	0,026105	MCM4	0,01691
MCM10	0,022801	TUSC4	0,017564
CXXC4	0,024891	FAM64A	0,016695
BIRC5	0,023378	GDAP1L1	0,017033
FOXM1	0,023158	ARID3B	0,016269
ZNF821	0,023179	GRM8	0,016036
NCAPG	0,021348	TFAP2B	0,017384
TGFBR2	0,023944	GTSE1	0,01602
TPX2	0,020627	MMP12	0,015909
MKI67	0,019426	ICA1	0,01653
ESPL1	0,020428	GINS2	0,015904
KIF20A	0,020081	LPAR2	0,016684
EIF4EBP1	0,01972	FLJ22184	0,016175
IGF2BP3	0,020061	PXMP2	0,015263
DTL	0,019036	TRAP1	0,015742
TTK	0,01888	KIF14	0,015321
PBK	0,018411	ARC	0,015642
FEV	0,019218	KLHL23	0,01707
PHOX2A	0,01842		
DCN	0,021125		
TOP2A	0,018001		
LOC157627	0,018996		
EYA1	0,01902		
CENPA	0,017678		
HJURP	0,016797		
KIF22	0,016577		
MGC39900	0,017562		
TMSL8	0,017562		

Figure 5.14: Neuroblastic tumour: SP/SR. Shapley value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

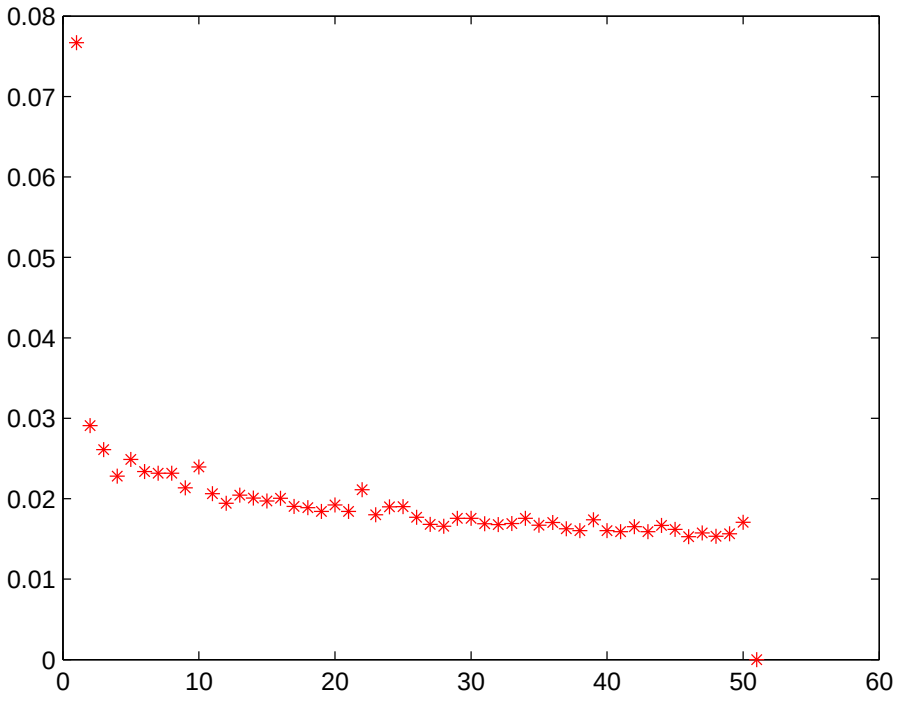


Figure 5.15: Neuroblastic tumour: SP/SR. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Shapley power index.

5.2.3 Data from lobular and ductal invasive breast carcinomas

Gene expression analysis was performed by using Human Genome U133-Plus 2.0 GeneChip arrays (Affymetrix, Inc., Calif). This data set contains 10 healthy samples of ductal and lobular cells, 5 samples of ductal cells and 5 samples of lobular cells derived from tumour tissues. In the following table we can see the ranking of the first 50 genes labelled with the Shapley value (SY) by using the weighted indices and Banzhaf power index (WMGBa) and Shapley power index (WMGSh) of the same 50 genes after they have played a weighted majority game. The weights in the game are the values of n that we have given to the genes in the cooperative game played with the weighted indices, the share q is $50\% + 1$.

	SY	Power index	WMGBa	Power index	WMGSh	Power index		SY	Power index	WMGBa	Power index	WMGSh	Power index
1	KCNU1	0,0017369	CYP7B1	0,14203	CYP7B1	0,060945	30	MPV17	0,00063948	C5orf29	0,045413	CD207	0,014674
2	TRAT1	0,001725	TRAT1	0,12173	TRAT1	0,044342	31	TMEFF1	0,00062415	SPA17	0,045072	FLNA	0,014621
3	CYP7B1	0,0016046	KCNU1	0,12051	CRTAM	0,042187	32	HIST1H2AG	0,00061281	MKKS	0,045063	KIF4A	0,014596
4	hCG_2032978	0,0013923	DACH2	0,10382	KCNU1	0,040051	33	FLJ35816	0,00061196	NRK	0,044512	KIF4B	0,014596
5	ZNF675	0,0012282	CRISP3	0,095383	DACH2	0,033444	34	UGT2B4	0,00060918	MMP1	0,044442	SPA17	0,014209
6	DACH2	0,001219	hCG_2032978	0,092778	CRISP3	0,032362	35	APOBEC3A	0,00060457	MPV17	0,044366	MPV17	0,014103
7	CRTAM	0,0012087	FMR1NB	0,090091	FMR1NB	0,030744	36	HIST1H2AM	0,00059883	SPESP1	0,042324	LOC285033	0,014065
8	ZFY	0,001087	CRTAM	0,090081	ZFY	0,029403	37	MMP1	0,00058417	KRT17	0,041871	MMP1	0,013931
9	GPR128	0,0010328	ZFY	0,087583	hCG_2032978	0,026666	38	FLNA	0,00058355	LSM11	0,041408	SPESP1	0,013926
10	CRISP3	0,0010273	ZNF675	0,082429	TEX14	0,026406	39	KIF4A	0,00058009	DMN	0,041028	DMN	0,013823
11	PDE6C	0,00096831	DEPDC7	0,079213	DEPDC7	0,02594	40	KIF4B	0,00058009	RIT2	0,040285	LSM11	0,013816
12	TRAM1L1	0,00093223	TEX14	0,077749	ZNF675	0,025458	41	MKKS	0,00057925	LOC285033	0,039512	CST4	0,0125
13	DEPDC7	0,00092639	TRAM1L1	0,077708	TRAM1L1	0,025077	42	GABRB1	0,00057586	LOC651721	0,039163	TMEFF1	0,012229
14	ZNF750	0,00088785	KRT14	0,066076	KRT14	0,024191	43	SPESP1	0,00056432	CST4	0,038783	LOC651721	0,012201
15	KRT14	0,00084938	HIST1H2AG	0,065322	HIST1H2AG	0,022628	44	KRT17	0,000553	TMEFF1	0,038338	RIT2	0,011842
16	TEX14	0,00081284	GPR128	0,064358	FLJ40473	0,020597	45	HABP2	0,00055263	APOBEC3A	0,037465	HIST1H2AM	0,011397
17	ZNF28	0,00080345	PDE6C	0,0578	PDE6C	0,019807	46	STYX	0,00055009	C20orf103	0,035827	C20orf103	0,011354
18	CD207	0,00080195	FLJ40473	0,054813	GPR128	0,019005	47	NRK	0,00054823	HIST1H2AM	0,035684	GABRB1	0,011247
19	FMR1NB	0,00079665	HMMR	0,054299	ZNF750	0,01727	48	C20orf103	0,00054602	STYX	0,034744	STYX	0,011105
20	FLJ40473	0,00077918	EEF1G	0,053303	HMMR	0,017262	49	DMN	0,0005454	GABRB1	0,033254	APOBEC3A	0,011001
21	HMMR	0,0007486	LOC729998	0,053303	EEF1G	0,016994	50	LOC285033	0,00054043	HABP2	0,032788	HABP2	0,01087
22	EEF1G	0,00069973	ZNF750	0,051903	LOC729998	0,016994							
23	LOC729998	0,00069973	FLJ35816	0,051054	FLJ35816	0,016485							
24	LSM11	0,00067243	UGT2B4	0,050269	C5orf29	0,016303							
25	LOC651721	0,00066891	ZNF28	0,050252	UGT2B4	0,016223							
26	SPA17	0,00066873	CD207	0,049877	NRK	0,01572							
27	CST4	0,000667	KIF4A	0,045947	MKKS	0,0153							
28	C5orf29	0,00066365	KIF4B	0,045947	KRT17	0,015065							
29	RIT2	0,0006495	FLNA	0,045426	ZNF28	0,015025							

Figure 5.16: Lobular. I take the first 50 genes classified by Shapley value in weighted indices. The Banzhaf, Shapley values in the weighted majority game are displayed.

Remark 5.2.3 In our ranking of genes, we identified an important gene HMMR which is already known to be associated with higher risk of breast cancer in humans([Pujana et al.(2007)]). In this paper the authors, starting with four known genes encoding tumour suppressors of breast cancer, combined gene expression profiling with functional genomic and proteomic (or 'omic') data from various species to generate a network containing 118 genes linked by 866 potential functional associations. This network shows higher connectivity than expected by chance, suggesting that its components function in biologically related pathways. One of the components of the network is HMMR, encoding a centrosome subunit. Two case-control studies of incident breast cancer indicate that the HMMR locus is associated with higher risk of breast cancer in humans. ■

KCNU1	0,12051	MPV17	0,044366
TRAT1	0,12173	TMEFF1	0,038338
CYP7B1	0,14203	HIST1H2AG	0,065322
hCG_2032978	0,092778	FLJ35816	0,051054
ZNF675	0,082429	UGT2B4	0,050269
DACH2	0,10382	APOBEC3A	0,037465
CRTAM	0,090081	HIST1H2AM	0,035684
ZFY	0,087583	MMP1	0,044442
GPR128	0,064358	FLNA	0,045426
CRISP3	0,095383	KIF4A	0,045947
PDE6C	0,0578	KIF4B	0,045947
TRAM1L1	0,077708	MKKS	0,045063
DEPDC7	0,079213	GABRB1	0,033254
ZNF750	0,051903	SPESP1	0,042324
KRT14	0,066076	KRT17	0,041871
TEX14	0,077749	HABP2	0,032788
ZNF28	0,050252	STYX	0,034744
CD207	0,049877	NRK	0,044512
FMR1NB	0,090091	C20orf103	0,035827
FLJ40473	0,054813	DMN	0,041028
HMMR	0,054299	LOC285033	0,039512
EEF1G	0,053303		
LOC729998	0,053303		
LSM11	0,041408		
LOC651721	0,039163		
SPA17	0,045072		
CST4	0,038783		
C5orf29	0,045413		
RIT2	0,040285		

Figure 5.17: Lobular. Banzhaf value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

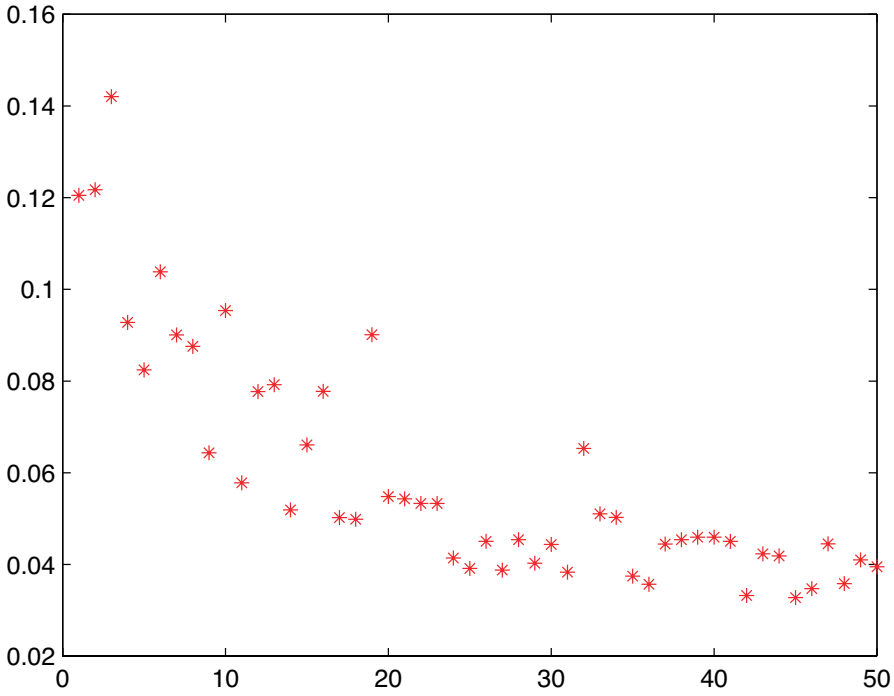


Figure 5.18: Lobular. Comparison of the previous 50 genes classified by the Shapley value on weightedindices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Banzhaf power index.

KCNU1	0,040051	MPV17	0,014103
TRAT1	0,044342	TMEFF1	0,012229
CYP7B1	0,060945	HIST1H2AG	0,022628
hCG_2032978	0,026666	FLJ35816	0,016485
ZNF675	0,025458	UGT2B4	0,016223
DACH2	0,033444	APOBEC3A	0,011001
CRTAM	0,042187	HIST1H2AM	0,011397
ZFY	0,029403	MMP1	0,013931
GPR128	0,019005	FLNA	0,014621
CRISP3	0,032362	KIF4A	0,014596
PDE6C	0,019807	KIF4B	0,014596
TRAM1L1	0,025077	MKKS	0,0153
DEPDC7	0,02594	GABRB1	0,011247
ZNF750	0,01727	SPESP1	0,013926
KRT14	0,024191	KRT17	0,015065
TEX14	0,026406	HABP2	0,01087
ZNF28	0,015025	STYX	0,011105
CD207	0,014674	NRK	0,01572
FMR1NB	0,030744	C20orf103	0,011354
FLJ40473	0,020597	DMN	0,013823
HMMR	0,017262	LOC285033	0,014065
EEF1G	0,016994		
LOC729998	0,016994		
LSM11	0,013816		
LOC651721	0,012201		
SPA17	0,014209		
CST4	0,0125		
C5orf29	0,016303		
RIT2	0,011842		

Figure 5.19: Lobular. Shapley value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

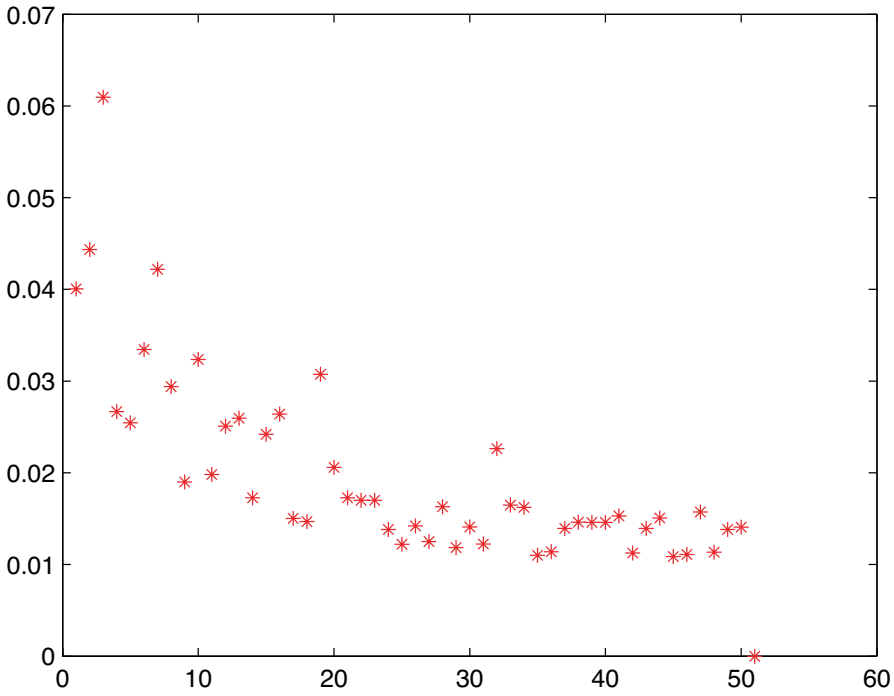


Figure 5.20: Lobular. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Shapley power index.

	SY	Power index	WMGBa	Power index	WMGSh	Power index		SY	Poer index	WMGBa	Power index	WMGSh	Power index
1	CRISP3	0,001578	EPYC	0,1289	EPYC	0,053485	30	GAGE6	0,00078691	SPANXA2	0,047292	MMP1	0,015442
2	FLJ30672	0,0015734	FLJ30672	0,12786	FLJ30672	0,053144	31	C20orf197	0,00076646	LOC285389	0,04531	HIST1H3D	0,014797
3	EPYC	0,001495	CRISP3	0,11985	CRISP3	0,052176	32	AHRR	0,00075993	AHRR	0,044629	RIMS2	0,01399
4	SPANXB1	0,0013188	IKZF3	0,11154	IKZF3	0,052036	33	SOX2OT	0,00075932	GAGE4	0,043721	SPANXA1	0,013976
5	SPANXB2	0,0013188	TRAT1	0,086493	TRAT1	0,041803	34	SCGB2A2	0,00075215	GAGE5	0,043721	SPANXA2	0,013976
6	GPM6A	0,0012047	CACNA2D3	0,073796	CACNA2D3	0,032335	35	C14orf25	0,00074636	ARHGAP6	0,043529	SPANXC	0,01262
7	FLJ33534	0,0011462	SPANXB1	0,071642	TMEFF1	0,027349	36	GAGE1	0,0007455	SPANXC	0,042698	LINGO1	0,012331
8	PCSK1	0,0011059	SPANXB2	0,071642	KRT5	0,022428	37	GAGE12J	0,0007455	GAGE12G	0,041428	GAGE4	0,012147
9	MMP1	0,0010354	GPM6A	0,065237	MCF2L2	0,021927	38	MAGEA6	0,0007455	GAGE12I	0,041428	GAGE5	0,012147
10	TMEFF1	0,00097972	TMEFF1	0,064083	PABPC5	0,021927	39	LOC728342	0,00073615	GAGE6	0,041428	LOC643300	0,011678
11	KRT5	0,00096731	KRT5	0,063182	CST4	0,021644	40	LOC643300	0,00072739	HIST1H3D	0,040284	LOC644745	0,011678
12	RIMS2	0,00096376	FLJ33534	0,060922	C20orf197	0,021318	41	LOC644745	0,00072739	LINGO1	0,039685	GAGE12G	0,011503
13	CACNA2D3	0,00094862	PCSK1	0,059158	CA6	0,02124	42	LINGO1	0,00072497	GAGE1	0,039154	GAGE12I	0,011503
14	KRT14	0,00090252	KRT14	0,057898	SPANXB1	0,021077	43	LOC285389	0,00072131	GAGE12J	0,039154	GAGE6	0,011503
15	SPANXA1	0,00089349	ABCC4	0,056175	SPANXB2	0,021077	44	ANXA3	0,00070911	MAGEA6	0,039154	LOC728342	0,011303
16	SPANXA2	0,00089349	MMP1	0,055522	SOX2OT	0,020711	45	TRAT1	0,00069184	LOC728342	0,039012	LOC153328	0,011257
17	CST4	0,00088635	KRT17	0,054617	ABCC4	0,020161	46	LOC153328	0,00068821	LOC643300	0,037933	GAGE1	0,010862
18	CA6	0,00085997	CST4	0,053303	KRT14	0,020144	47	HIST1H3D	0,00067501	LOC644745	0,037933	GAGE12J	0,010862
19	ABCC4	0,00083706	CENPA	0,052876	LOC285389	0,020051	48	ARHGAP6	0,0006649	LOC153328	0,035996	MAGEA6	0,010862
20	IKZF3	0,00083569	CA6	0,052257	KRT17	0,019666	49	GAGE2A	0,00066266	GAGE2A	0,034655	GAGE2A	0,0095942
21	GAGE4	0,00082833	SCGB2A2	0,051237	SCGB2A2	0,018578	50	GAGE7	0,00066266	GAGE7	0,034655	GAGE7	0,0095942
22	GAGE5	0,00082833	RIMS2	0,049883	GPM6A	0,018406							
23	KRT17	0,00082218	MCF2L2	0,049806	CENPA	0,018336							
24	CENPA	0,00081033	PABPC5	0,049806	C14orf25	0,018236							
25	MCF2L2	0,00080899	SOX2OT	0,049077	ARHGAP6	0,018136							
26	PABPC5	0,00080899	C20orf197	0,047632	FLJ33534	0,01764							
27	SPANXC	0,00080842	C14orf25	0,047353	AHRR	0,017345							
28	GAGE12G	0,00078691	ANXA3	0,047335	PCSK1	0,017213							
29	GAGE12I	0,00078691	SPANXA1	0,047292	ANXA3	0,01679							

Figure 5.21: Ductal. I take the first 50 genes classified by Shapley value in weighted indices. The Banzhaf, Shapley values in the weighted majority game are displayed.

CRISP3	0,11985	GAGE6	0,041428
FLJ30672	0,12786	C20orf197	0,047632
EPYC	0,1289	AHRR	0,044629
SPANXB1	0,071642	SOX2OT	0,049077
SPANXB2	0,071642	SCGB2A2	0,051237
GPM6A	0,065237	C14orf25	0,047353
FLJ33534	0,060922	GAGE1	0,039154
PCSK1	0,059158	GAGE12J	0,039154
MMP1	0,055522	MAGEA6	0,039154
TMEFF1	0,064083	LOC728342	0,039012
KRT5	0,063182	LOC643300	0,037933
RIMS2	0,049883	LOC644745	0,037933
CACNA2D3	0,073796	LINGO1	0,039685
KRT14	0,057898	LOC285389	0,04531
SPANXA1	0,047292	ANXA3	0,047335
SPANXA2	0,047292	TRAT1	0,086493
CST4	0,053303	LOC153328	0,035996
CA6	0,052257	HIST1H3D	0,040284
ABCC4	0,056175	ARHGAP6	0,043529
IKZF3	0,11154	GAGE2A	0,034655
GAGE4	0,043721	GAGE7	0,034655
GAGE5	0,043721		
KRT17	0,054617		
CENPA	0,052876		
MCF2L2	0,049806		
PABPC5	0,049806		
SPANXC	0,042698		
GAGE12G	0,041428		
GAGE12I	0,041428		

Figure 5.22: Ductal. Banzhaf value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

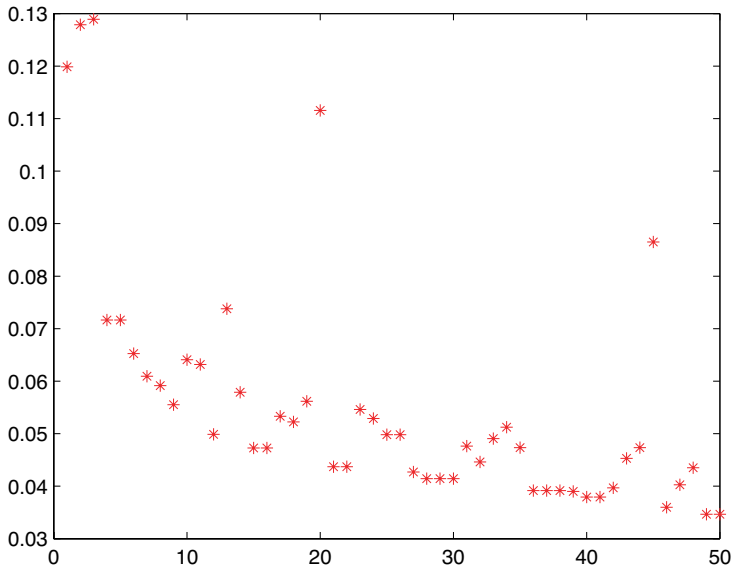


Figure 5.23: Ductal. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Banzhaf power index.

CRISP3	0,052176	GAGE6	0,011503
FLJ30672	0,053144	C20orf197	0,021318
EPYC	0,053485	AHRR	0,017345
SPANXB1	0,021077	SOX2OT	0,020711
SPANXB2	0,021077	SCGB2A2	0,018578
GPM6A	0,018406	C14orf25	0,018236
FLJ33534	0,01764	GAGE1	0,010862
PCSK1	0,017213	GAGE12J	0,010862
MMP1	0,015442	MAGEA6	0,010862
TMEFF1	0,027349	LOC728342	0,011303
KRT5	0,022428	LOC643300	0,011678
RIMS2	0,01399	LOC644745	0,011678
CACNA2D3	0,032335	LINGO1	0,012331
KRT14	0,020144	LOC285389	0,020051
SPANXA1	0,013976	ANXA3	0,01679
SPANXA2	0,013976	TRAT1	0,041803
CST4	0,021644	LOC153328	0,011257
CA6	0,02124	HIST1H3D	0,014797
ABCC4	0,020161	ARHGAP6	0,018136
IKZF3	0,052036	GAGE2A	0,0095942
GAGE4	0,012147	GAGE7	0,0095942
GAGE5	0,012147		
KRT17	0,019666		
CENPA	0,018336		
MCF2L2	0,021927		
PABPC5	0,021927		
SPANXC	0,01262		
GAGE12G	0,011503		
GAGE12I	0,011503		

Figure 5.24: Ductal. Shapley value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

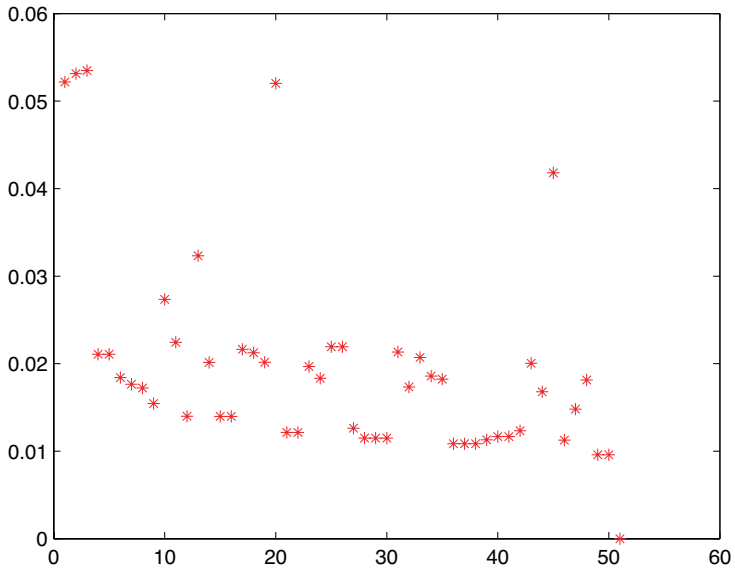


Figure 5.25: Ductal. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Shapley power index.

5.2.4 Data from colon tumour

Gene expression analysis was performed by using Affymetrix oligonucleotide (http://microarray.princeton.edu/oncology/affydata/index.html) microarrays for a set of 40 tumour samples and a set of 22 normal samples. In the following table we can see the ranking of the first 50 genes labelled with the Shapley value (SY) by using the weighted indices and Banzhaf power index (WMGBa) and Shapley power index (WMGSh) of the same 50 genes after they have played a weighted majority game. The weights in the game are the values of n that we have given to the genes in the cooperative game played with the weighted indices, the share q is $50\% + 1$.

	SY	Power index	WMGBa	Power index	WMGSh	Power index		SY	Power index	WMGBa	Power index	WMGSh
1 Hsa.8831	T49941	0,0098248	R36977	0,1544	R36977	0,058615	29 Hsa.1588	U09587	0,0029307	H40137	0,041392	R08021
2 Hsa.549	R36977	0,0096155	T51261	0,11963	T51261	0,046732	30 Hsa.1701	M86934	0,0028796	R41873	0,040628	X53586
3 Hsa.22762	H17434	0,0083611	H72234	0,0977	T49941	0,038656	31 Hsa.471	M29277	0,0028438	U09587	0,040576	H40269
4 Hsa.9972	T51261	0,0070922	H17434	0,091886	H17434	0,03489	32 Hsa.9218	T51858	0,0028349	R84411	0,040466	R10066
5 Hsa.7	H72234	0,0067444	M15841	0,087466	H72234	0,033965	33 Hsa.2157	R20554	0,0028175	Y00971	0,040255	R41873
6 Hsa.2196	M58050	0,0067087	T49941	0,084495	T41204	0,032903	34 Hsa.2959	K03124	0,0028066	T83368	0,03997	J05032
7 Hsa.9353	M15841	0,005824	M58050	0,081495	M15841	0,031709	35 Hsa.1209	T41204	0,0027826	T51858	0,039666	M83751
8 Hsa.1143	T49941	0,0049608	R16156	0,066671	M58050	0,029906	36 Hsa.23824	R41873	0,0027485	R54097	0,039199	Y00971
9 Hsa.6814	H08393	0,0047805	H65355	0,064892	T64885	0,026204	37 Hsa.31500	R62945	0,0027448	H40269	0,039148	M19045
10 Hsa.831	M22382	0,004575	M22382	0,064317	T58731	0,025426	38 Hsa.594	M83751	0,0027186	D00762	0,039094	D00762
11 Hsa.7652	R16156	0,0041367	T41204	0,064019	H65355	0,024696	39 Hsa.462	U09564	0,002674	U09564	0,037141	T51023
12 Hsa.42625	H65355	0,004034	T84049	0,059947	M15841	0,023511	40 Hsa.832	T51023	0,0026239	T65740	0,036635	R67999
13 Hsa.1047	R84411	0,00372	R05145	0,058056	T84049	0,023058	41 Hsa.60	D00762	0,002618	U28686	0,035313	T51858
14 Hsa.1410	R54097	0,0036787	T58731	0,058028	M58050	0,02302	42 Hsa.37553	H40269	0,0026011	T51023	0,035277	U28686
15 Hsa.3306	X12671	0,0035641	R43914	0,057617	R05145	0,02274	43 Hsa.3230	U28686	0,0025839	R20554	0,034599	U09564
16 Hsa.21562	R08021	0,0033905	R62945	0,05704	M22382	0,022082	44 Hsa.11240	T58731	0,0025839	T64885	0,034127	T65740
17 Hsa.10664	T83368	0,0033898	M15841	0,051913	R43914	0,021361	45 Hsa.36689	Z50753	0,0025717	R49416	0,03177	R20554
18 Hsa.3141	R05145	0,0033638	X12671	0,0519	R62945	0,020995	46 Hsa.5908	R67999	0,0025217	M29277	0,028788	R49416
19 Hsa.13628	T64885	0,0033302	T58731	0,051511	R43914	0,020332	47 Hsa.42186	H61410	0,0024757	M86934	0,02798	M29277
20 Hsa.2821	X53586	0,0033198	K03124	0,0491	K03124	0,018091	48 Hsa.2964	Y00971	0,0024685	M28373	0,022027	M86934
21 Hsa.4937	R43914	0,0032856	X53586	0,046683	R84411	0,01761	49 Hsa.5756	T65740	0,0024557	R54097	0,019352	M28373
22 Hsa.31630	R64115	0,0032422	M83751	0,044293	X12671	0,017532	50 Hsa.891	M19045	0,0024397	H40269	0,019257	M83751
23 Hsa.7395	R10066	0,0031524	R64115	0,043793	U28686	0,017491						
24 Hsa.37541	H40137	0,0030939	M19045	0,042937	H40137	0,01675						
25 Hsa.2280	R49416	0,0030479	R10066	0,0423	R54097	0,016599						
26 Hsa.601	J05032	0,0030449	R08021	0,041969	T83368	0,016303						
27 Hsa.1731	M28373	0,0030042	R67999	0,041715	U09587	0,01627						
28 Hsa.6288	T84049	0,0029694	J05032	0,04153	R64115	0,016237						

Figure 5.26: Colon. I take the first 50 genes classified by Shapley value in weighted indices. The Banzhaf, Shapley values in the weighted majority game are displayed.

Remark 5.2.4 Some of the genes selected were previously observed in association with the colon cancer ([Fujarewicz K., Wiench M. (2003)]): the vasoactive intestinal peptide (M36634: Human vasoactive intestinal peptide (VIP)), has been suggested to promote the growth and proliferation of tumour cells; the membrane cofactor protein (M58050; Human membrane cofactor protein (MCP)) represents a possible mechanism of the ability of the tumour to evade destruction by the immune system. H72234: DNA-(APURINIC OR APYRIMIDINIC SITE) LYASE (HUMAN) plays an important role in DNA repair and in resistance of cancer cells to radiotherapy ([Moler E.J., Chow M.L, Mian I.S. (2000)]). ■

T49941	0,084495	T51858	0,022027
R36977	0,1544	R20554	0,039666
H17434	0,091886	H65355	0,03177
T51261	0,11963	K03124	0,0491
H72234	0,0977	T41204	0,064019
M58050	0,081495	R41873	0,040628
M15841	0,087466	R62945	0,051913
M15841	0,0519	M83751	0,044293
H08393	0,057617	U09564	0,037141
M22382	0,064317	T51023	0,034599
R16156	0,066671	D00762	0,039094
H65355	0,064892	H40269	0,039148
R84411	0,040466	T58731	0,058028
R54097	0,039199	U28686	0,035277
X12671	0,051511	Z50753	0,034127
R08021	0,041969	R67999	0,041715
T83368	0,03997	H61410	0,035313
T64885	0,058056	Y00971	0,040255
X53586	0,028788	T65740	0,036635
R43914	0,046683	M19045	0,042937
R64115	0,05704		
R10066	0,043793		
H40137	0,0423		
R49416	0,041392		
J05032	0,02798		
M28373	0,04153		
T84049	0,019257		
U09587	0,059947		
M86934	0,040576		
M29277	0,019352		

Figure 5.27: Colon. Banzhaf value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

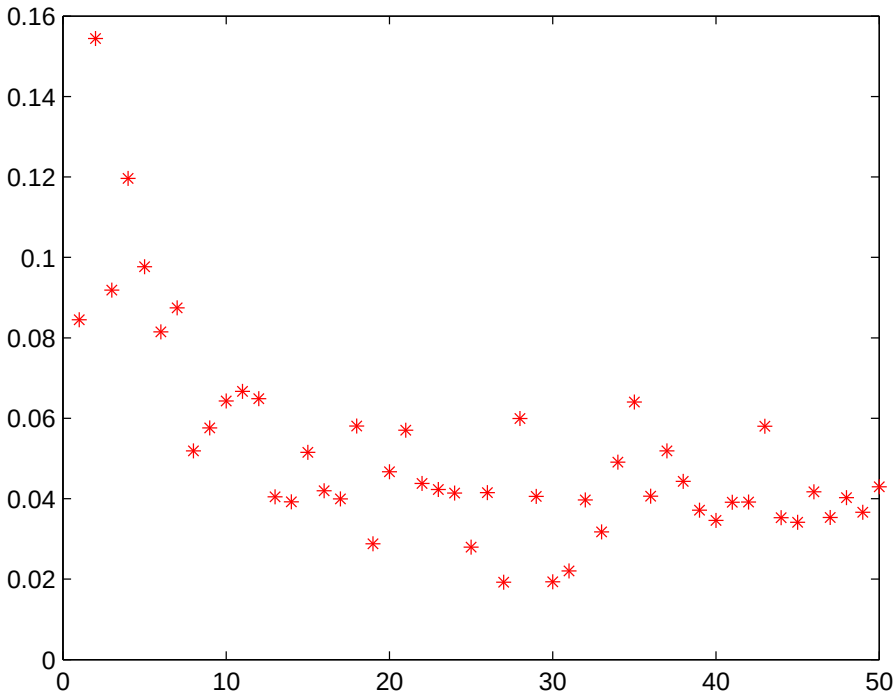


Figure 5.28: Colon. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Banzhaf power index.

T49941	0,038656	M29277	0,0099332
R36977	0,058615	T51858	0,012449
H17434	0,03489	R20554	0,012
T51261	0,046732	K03124	0,018091
H72234	0,033965	T41204	0,032903
M58050	0,029906	R41873	0,014771
M15841	0,031709	R62945	0,020995
M15841	0,023058	M83751	0,0144
H08393	0,020332	U09564	0,012224
M22382	0,022082	T51023	0,012786
R16156	0,023511	D00762	0,012924
H65355	0,024696	H40269	0,015697
R84411	0,01761	T58731	0,025426
R54097	0,016599	U28686	0,01225
X12671	0,017532	R67999	0,012431
R08021	0,016227	T41204	0,012773
T83368	0,016303	H61410	0,017491
R05145	0,02274	Y00971	0,013992
T64885	0,026204	T65740	0,012046
X53586	0,015891	M19045	0,013446
R43914	0,021361		
R64115	0,016237		
R10066	0,015124		
H40137	0,01675		
R49416	0,011106		
J05032	0,014685		
M28373	0,0095128		
T84049	0,02302		
U09587	0,01627		
M86934	0,0096472		

Figure 5.29: Colon. Shapley value of the first 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game.

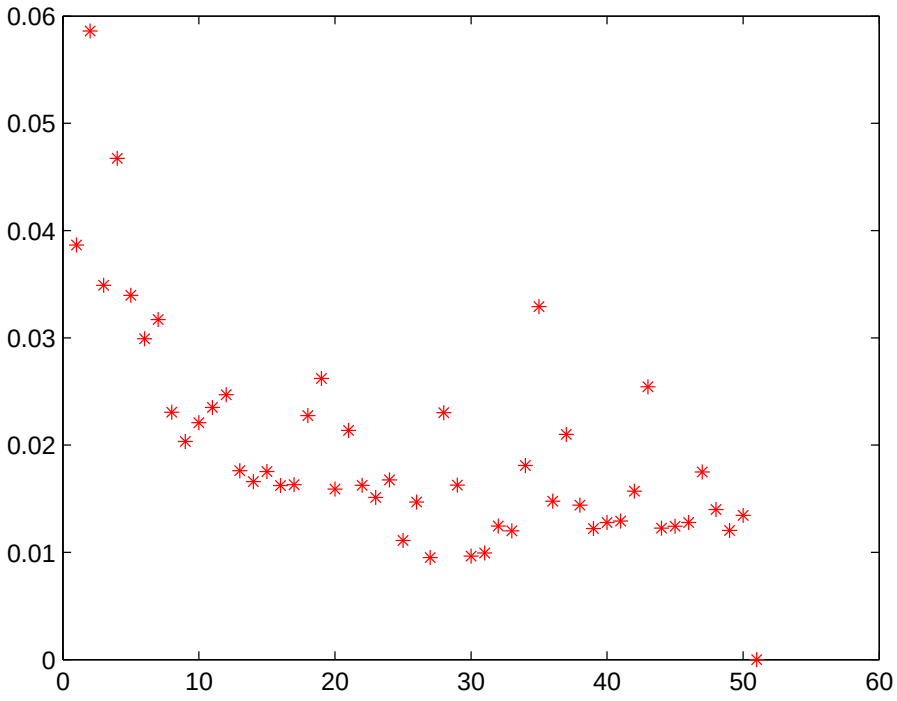


Figure 5.30: Colon. Comparison of the previous 50 genes classified by the Shapley value on weighted indices now classified in the weighted majority game. Genes are labelled on the x-axis; on y-axis we have their Shapley power index.

Bibliography

[Albino et al. (2008)] Albino D., Scaruffi P., Moretti S., Coco S., Di Cristofano C., Cavazzana A., Truini M, Stigliani S., Bonassi S., Tonini G.P. (2008) Stroma poor and stroma rich gene signatures show a low intratumoural gene expression heterogeneity in Neuroblastic tumours, *Cancer*,113,1412-22

[Alon et al. (1999)] 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 tumour and normal colon tissue probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America, 96, 6745-6750.

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

[Bilbao et al. (2000)] Bilbao J.M., Fernandez J.R., Jimenez Losada A., Lopez J.J (2000). Generating Functions for Computing Power Indices efficiently, TOP8 **2**, 191-213.

[Carreras and Freixas (2008)] Carreras F. and Freixas J. On ordinal equivalence of power measures given by regular semivalues, Mathemmatical Social Sciences, **55**, 221-234.

[Chin et al. (2006)] Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM et al. (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 10:529-541.

[Fearon (1997)] Fearon ER.(1997) Human cancer syndromes: clues to the origin and nature of cancer. Science 278:1043-1050.

[Fragnelli and Moretti (2008)] Fragnelli V., Moretti S. (2008) A game theoretical approach to the classification problem in gene expres-

sion data analysis, *Computers & Mathematics with Applications*, **55**(5), 950-959.

[Ge et al. (2003)] Ge H, Walhout AJ, Vidal M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. Trends Genet 19:551-560.

[Hanahan and Weinberg (2000)] Hanahan D, Weinberg RA. (2000) The hallmarks of cancer. Cell 100:57-70.

[Kalai and Samet (1987)] Kalai E., Samet D. (1987) On weighted Shapley values, *International Journal of Game Theory*, **24**, 179-186. 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.

[Lander et al. (2001)] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K et al.(2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

[Laruelle, Valenciano (2001)] Laruelle A., Valenciano F. (2001) Shapley-Shubik and Banzhaf indices revisited, *Mathematics of Operations Research*, **26**, 89-104.

[Leech (2002)] Leech D. (2002). Computation of Power Indices. Worwich Economic Research Paper 644.

[Liu et al. (2006)] Liu ET, Kuznetsov VA, Miller LD. (2006) In the pursuit of complexity: systems medicine in cancer biology. Cancer Cell 9:245-247.

[Lucchetti et al.] Lucchetti R., Moretti S., Patrone F. and Radrizzani P. The Shapley and Banzhaf indices in microarray games, *to appear*, .

[Monderer and Samet (2001)] Monderer D., Samet D. (2002) Variations on the Shapley value, *In Handbook of Game Theory, R.J. Aumann and S. Hart editors*, Elsevier Science, Amsterdam, **54**.

[Moretti (2006)] Moretti S. (2006) Minimum cost spanning tree games and gene expression data analysis, *GameNets '06: Proceeding from the 2006 workshop on Game theory for communications and networks* (Pisa, Italy), ACM International Conference Proceeding Series, New York, NY, USA, **199**, pp.8.

[Moretti et al. (2007)] Moretti S., Patrone F., Bonassi S. (2007) The class of microarray games and the relevance index for genes, *TOP*, 15, 256-280.

[Moretti and Patrone (2008)] Moretti S., Patrone F. (2008) Transversality of the Shapley value, *Top*, DOI: 10.1007/s11750-008-0044-5.

[Moretti et al. (2008)] Moretti S., van Leeuwen D., Gmuender H., Bonassi S., van Delft J., Kleinjans J., Patrone F., Merlo D.F. (2008) Combining Shapley value and statistics to the analysis of gene expression data in children exposed to air pollution. (submitted)

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

[Parmigiani et al. (2003)] Parmigiani G. , Garrett E.S., Irizarry R.A., Zeger. S.L.(ed.) (2003) The analysis of gene expression data: methods and software, Springer, New York.

[Peto (2001)] Peto J. (2001) Cancer epidemiology in the last century and the next decade. Nature 411:390-395.

[Pujana et al.(2007)] Pujana M.A., J Han J.D., Starita L.M., Stevens K.N., Tewari M., Sook Ahn J., Rennert G., Moreno V., Kirchhoff T., Gold B., Assmann V., ElShamy W., Rual J.F., Levine D., Rozek L.S., Gelman R.S., Gunsalus K.C., Greenberg R.A., Sobhian B., Bertin N., Venkatesan K., Ayivi-Guedehoussou N., Sol X., Herndez P., Lzaro C., Nathanson K.L., Weber B.L., Cusick M.E., Hill D.E., Offit K., Livingston D.M., Gruber S.B., Parvin J.D., Vidal M.(2007) Network modeling links breast cancer susceptibility and centrosome dysfunction, Nature Genetics 39, 1338 - 1349 DOI:10.1038/ng.2007.2

[Shapley (1953)] 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.

[Sjöblom et al. (2006)] Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JKV, Gazdar AF, Hartigan J et al. (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314:268-274.

[Turashvili et al. (2007)] Turashvili G., Bouchal J., Baumforth K., Wei W., Dziechciarkova M., Ehrmann J., Klein J., Fridman E., Skarda J.,

Srovnal J., Hajduch M., Murray P., Kolar Z.(2007) Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. BMC Cancer DOI: 10.1186/1471-2407/7/55.

[Venter et al. (2001)] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M et al.(2001) The sequence of the human genome. Science 291:1304-1351.

[Volgelstein and Kinzler (2004)] Vogelstein B, Kinzler KW.(2004) Cancer genes and the pathways they control. Nat Med 10:789-799.

[Weber (1988)] Weber R. J. (1988). Probabilistic Values for Games, in The Shapley Value. A. E. Roth (ed.), Cambridge University Press, 307-317.

[Weinberg (2006)] Weinberg RA.(2006) Biology of cancer. Garland Science, Taylor and Francis Group, New York, 864.

[Fujarewicz K., Wiench M. (2003)] . Fujarewicz K., Wiench M. (2003). Selecting differentially expressed genes for colon tumour classification. International Journal of Applied Mathematics and Computer Science, 13(3), 327-335.

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

[Yi Hong et al.(2007)] Yi H., Kok S.H., Kong W. E., Peh Y. C. (2007) A Susceptibility Gene Set for Early Onset Colorectal Cancer That Integrates Diverse Signaling Paqthways: Implication for Tumorigenesis. Clin Cancer Res, 13(4).

Contents