Using Variable Precision Rough Set for Selection and Classification of Biological Knowledge Integrated in DNA Gene Expression

D. Calvo-Dmgz, J. F. Gálvez*, D. Glez-Peña, S. Gómez-Meire, F. Fdez-Riverola

ESEI: Escuela Superior de Enxeñería Informática, University of Vigo,
Ed. Politécnico, Campus Universitario As Lagoas s/n 32004 Ourense, Spain,
http://www.esei.uvigo.es

Summary

DNA microarrays have contributed to the exponential growth of genomic and experimental data in the last decade. This large amount of gene expression data has been used by researchers seeking diagnosis of diseases like cancer using machine learning methods. In turn, explicit biological knowledge about gene functions has also grown tremendously over the last decade. This work integrates explicit biological knowledge, provided as gene sets, into the classication process by means of Variable Precision Rough Set Theory (VPRS). The proposed model is able to highlight which part of the provided biological knowledge has been important for classification. This paper presents a novel model for microarray data classification which is able to incorporate prior biological knowledge in the form of gene sets. Based on this knowledge, we transform the input microarray data into supergenes, and then we apply rough set theory to select the most promising supergenes and to derive a set of easy interpretable classification rules. The proposed model is evaluated over three breast cancer microarrays datasets obtaining successful results compared to classical classification techniques. The experimental results shows that there are not significat differences between our model and classical techniques but it is able to provide a biological-interpretable explanation of how it classifies new samples.

1 Introduction

During the last decade, DNA microarrays have been used for aswering many biological questions. Some of the most frequently applications of microarrays study genes expression in diferent situations (healthy/diseased), molecular classification of complex disease, prediction of response to medication, among others. The availability of public biological knowledge allows researchers to extract biological conclusions and to interpretate their experimental results. Sources of knowledge include genomic databases, ontologies, public experimental datasets, metabolic pathways, gene-disease association registries, etc. This biological knowledge could be applied to statistical and machine learning techniques to improve global results when they are applied to microarray data.

There are some interesting proposals in this line. Some of them represent the knowledge as networks, whereas others make use gene sets. Among those using sets of genes, there are recent works such as *supergenes* [1], *Nonparametric pathway-based regression* (NPR) [2], *Gene regularized discriminant analysis* (GRDA) [3] and mPAM/mPPLS [4]. Tai & Pan proposed

*To whom correspondence should be addressed. Email: galvez@uvigo.es

a modification of classification methods *Nearest shrunken centroids* (PAM) [5] and *Penalized partial least squares* (PPLS) [6] called *mPAM* and *mPLS*, respectively [4]. Both methods implicity contain a mechanism for selecting genes based on a penalty applied according to the discriminatory power of the gene. Authors suggest that the penalty depends on a parameter λ that is global and arbitrary for all genes, so therefore they propose that this parameter be different for genes belonging to different groups, for example, genes known as marker in cancer. Authors also have presented a modification of LDA-based methods called GRDA [3]. This method incorporates information from KEGG [7] metabolic pathways in their experiments. Wei & Li developed NPR [2], a modification of the *boosting* scheme [8]. This paper proposes that, in each step of *boosting*, a classifier be trained for each predefined pathway. After the training process, a classifier based on several models with biological criteria is obtained. In addition, those metabolic pathways with greater success in training are highlighted. Chen & Wang focused on regression problems with microarray data applied patient survival prediction, rather than in classification [1]. However, the proposed model could be easily extended to classification. They propose a new framework that takes prior information in form of gene sets representing metabolic pathways. The expression levels of genes belonging to each pathway are summarized in a single variable called *supergene*, by means of *Supervised Principal Component Analysis* (SPCA) [9].

In addition, there are some publications focused on applying rough set theory to improve classification techniques over DNA microarray data. Most of these works try to select features that provide valuable information for classificacion and, to achieve this, they define a set of metrics for determining which features are most important and which must be discarded. However, as far as we know, none of the techniques has used rough set theory for the classification step, since they are limited to the feature selection. Zhou, Liu & Zhu propose a feature selection step using *Mutual Information* and *Rough set* (MIRS). The idea is to select those features that have the highest mutual information with the target class to predict [10]. Then, rough set theory is applied to remove redundancy among the selected features. Another recent method that uses rough set theory for classication of DNA microarray data was proposed by Maji & Paul [11]. In this paper, the *Max-Dependency* is studied as a feature selection criterion that uses the feature *dependence* measure based on rough set theory. In addition, a new criterion for feature selection called *Maximum Relevance-Maximum Significance* (MRMS) is proposed. This method uses the measures of *relevance* and *significance* of the rough set theory.

In this paper we present a new model divided into five steps, including (*i*) supergene generation, (*ii*) attribute discretization, (*iii*) feature selection, (*iv*) decision rule generation and (*v*) rule application during classification.

The paper is structured as follows. The second section introduces the rough set theory concepts we have used. The third section describes how the proposed model represents and incorporates prior biological knowledge. The fourth section describes the global model architecture and details each step. The fifth section shows the experimental results and finally the last section includes the conclusions and further work.

2 Rough Sets, Variable Precision Rough Sets and CAI Model

Rough Set Model was introduced by Z. Pawlak in 80's to satisfy the need for a formal framework to manage imprecise knowledge expressed in terms of data acquired from experiments

[12]. *Imprecise* refers to the fact that the granularity of knowledge causes *indiscernibility*. These imprecise concepts can be defined approximately with available knowledge using two precise concepts called *lower approximation* ($\underline{R}X$) and *upper approximation* ($\overline{R}X$).

Let $I = (\mathbb{U}, \mathbb{A})$ be an *information system* (attribute-value sytem), where $\mathbb{U}$ is a non-empty set of finite objects and $\mathbb{A}$ is a non-empty, finite set of attributes such that $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. V_a is the set of values that attribute a may take. The information table assigns a value $a(x)$ from V_a to each attribute a and object x in the universe $\mathbb{U}$. With any $R \subseteq \mathbb{A}$ there is an associated *equivalenve relation* $IND(R) = \{(x, y) \in \mathbb{U}^2 | \forall a \in R, a(x) = a(y)\}$. The relation $IND(R)$ is called a ***R-indiscernibility relation***. The partition of $\mathbb{U}$ is a family of all equivalence classes of $IND(R)$ and is denoted by $\mathbb{U}/IND(R)$.

Let $X \subseteq \mathbb{U}$ be a target set that we wish to represent using attribute subset P; that is, we are told that an arbitrary set of objects X comprises a single class, and we wish to express this class (i.e., this subset) using the equivalence classes induced by attribute subset R. In general, X cannot be expressed exactly, because the set may include and exclude objects which are indistinguishable on the basis of attributes R.

However, the target set X can be *approximated* using only the information contained within R by constructing the *lower* and *upper approximations* of X. The ***R-lower approximation*** $\underline{R}X = \bigcup\{[x]_R \in \mathbb{U}/R : [x]_R \subseteq X\}$ or positive region, is the union of all equivalence classes in $[x]_R$ which are contained by (i.e., are subsets of) the target set. The ***R-upper approximation*** $\overline{R}X = \bigcup\{[x]_R \in \mathbb{U}/R : [x]_R \cap X \neq \emptyset\}$ is the union of all equivalence classes in $[x]_R$ which have non-empty intersection with the target set. Based on these concepts, the reference universe $\mathbb{U}$ can be divided in three regions: the *positive region* $POS_R(X) = \underline{R}X$; the *negative region* $NEG_R(X) = \mathbb{U} - \overline{R}X$; and the *boundary region* $BN_R(X) = \overline{R}X - \underline{R}X$.

The boundary region consists of those objects that can neither be ruled in nor ruled out as members of the target set X. The lower approximation contains objects that are members of the target set with certainty (probability = 1), while the upper approximation contains objects that are members of the target set with non-zero probability. The tuple $\langle \underline{R}X, \overline{R}X \rangle$ is called a *rough set*. Boundary region can be considered also as an area there classification is not possible under a certain level of error. With this in mind, Rough Set model can be extended to characterize a set in terms of uncertain information under some levels of certainty. This idea is based in *Variable Precision Rough Set* [13].

In data analysis, *Variable Precision Rough Set* (VPRS) is very useful for addressing problems where data sets have lots of boundary objects. In addition, this model allows identifying data patterns that otherwise would be lost. The standard definition of the set inclusion relation is too rigoruos to represent any *almost* complete set inclusion. So, the extended notion should be able to allow for some degree of misclassification in the large correct classification. Before a more general definition is presented, it is convenient to introduce the measure $c(X, Y)$ of the *relative degree of misclassification* (1) of the set X with respect to set Y defined as:

$$\begin{array}{ll} c(X,Y) = 1 - |X \cap Y| \,/\, |X| & \text{if } |X| > 0 \text{ or} \\ c(X,Y) = 0 & \text{if } |X| = 0 \end{array} \tag{1}$$

The *majority inclusion relation* (2) under an admissible umbral of classification error β (which must be within the range $0 \leq \beta \leq 0.5$ is defined as:

$$X \subseteq_\beta Y \Leftrightarrow c(X, Y) \leq \beta \tag{2}$$

By replacing the inclusion relation with majority inclusion relation in the original definition of *lower approximation* and *upper approximation*, the generalized notion of β*-lower approximation* $\underline{R}_\beta X = \bigcup\{[x]_R \in \mathbb{U}/R : [x]_R \subseteq_\beta X\}$ and β*-upper approximation* $\underline{R}_\beta X = \bigcup\{[x]_R \in \mathbb{U}/R : [x]_R \subseteq_\beta X\}$.

Alike in rough set model, the universe $\mathbb{U}$ can be divided in three different regions: the β*-positive region* $POS_{R,\beta}(X) = \underline{R}_\beta X$; the β*-negative region* $NEG_{R\beta}(X) = \mathbb{U} - \overline{R}_\beta X$; and the β*-boundary region* $BN_{R\beta}(X) = \overline{R}_\beta X - \underline{R}_\beta X$

The *Conjuntos Aproximados con Incertidumbre* (CAI) or *Uncertainty Rough Sets* [14] model is derived from the VPRS model. As the VPRS model, CAI works also with uncertain information but with the aim of improve the classification power in order to introduce stronger rules. In the CAI model, uncertainy is introduced at two different levels: the constituting blocks of knowledge (elementary categories) and the overall knowledge, through the relationship of majority inclusion. So that, two different knowledge bases P and Q are equivalent or approximately equal, and denoted by $P \approx_\beta Q$, if the majority of their constituting blocks are similar.

3 Introducing Biological Knowledge

One of the problems with the high-dimensional microarray data is that not all the genes (attributes, variables) are useful for classifying a sample into a class (phenomena of interest) [15]. As we stated before, introducing biological knowledge in microarray data analysis can (*i*) reduce data dimensionality, (*ii*) improve the model interpretability, by targeting only at genes that are involved or related to biological concepts of interest, and (*iii*) enhance the model robustness when mixing samples coming from different experiments.

In the proposed model, the biological knowledge is represented as follows. Given a universe of discourse $\mathbb{U}$ (e.g. composed by the genes measured with a microarray), a concept of interest is often not explicitly expressed by the expert, but defined by joining a series of subsets of the universe of interest, defined independently. We call *interpretation context* [16] any family of subsets $F = \{F_1, \ldots, F_i, \ldots, F_n\}$, with $F_i \subseteq \mathbb{U}$, where all interpretation context defines a concept (subset) of interest, formed by the union of all categories of F and denoted by $\bigcup F$. Any interpretation context F imposes a structure on the concept of interest given by $\bigcup F$, formed by basic categories which are non-overlapping and constitute a cover of $\bigcup F$. Formally, given an interpretation context F and given $N = \{1, 2, \ldots, n\}$, it is called *basic category* to any set constructed from F as follows:

$$m_S = \left(\bigcap_{i \in S} F_i\right) - \left(\bigcup_{i \in N-S} F_i\right), \text{ with } \emptyset \neq S \in \wp(N) \tag{3}$$

4 Classification Process

The classification process is divided into five steps as shown in Figure 1. Firstly, supergenes are created, which summarize the information gene sets intersections (called basic categories), by means of Principal Component Analysis (PCA). Then, continuous values of supergenes are discretized using Discriminant Fuzzy Patterns. In the third step the most relevant supergenes

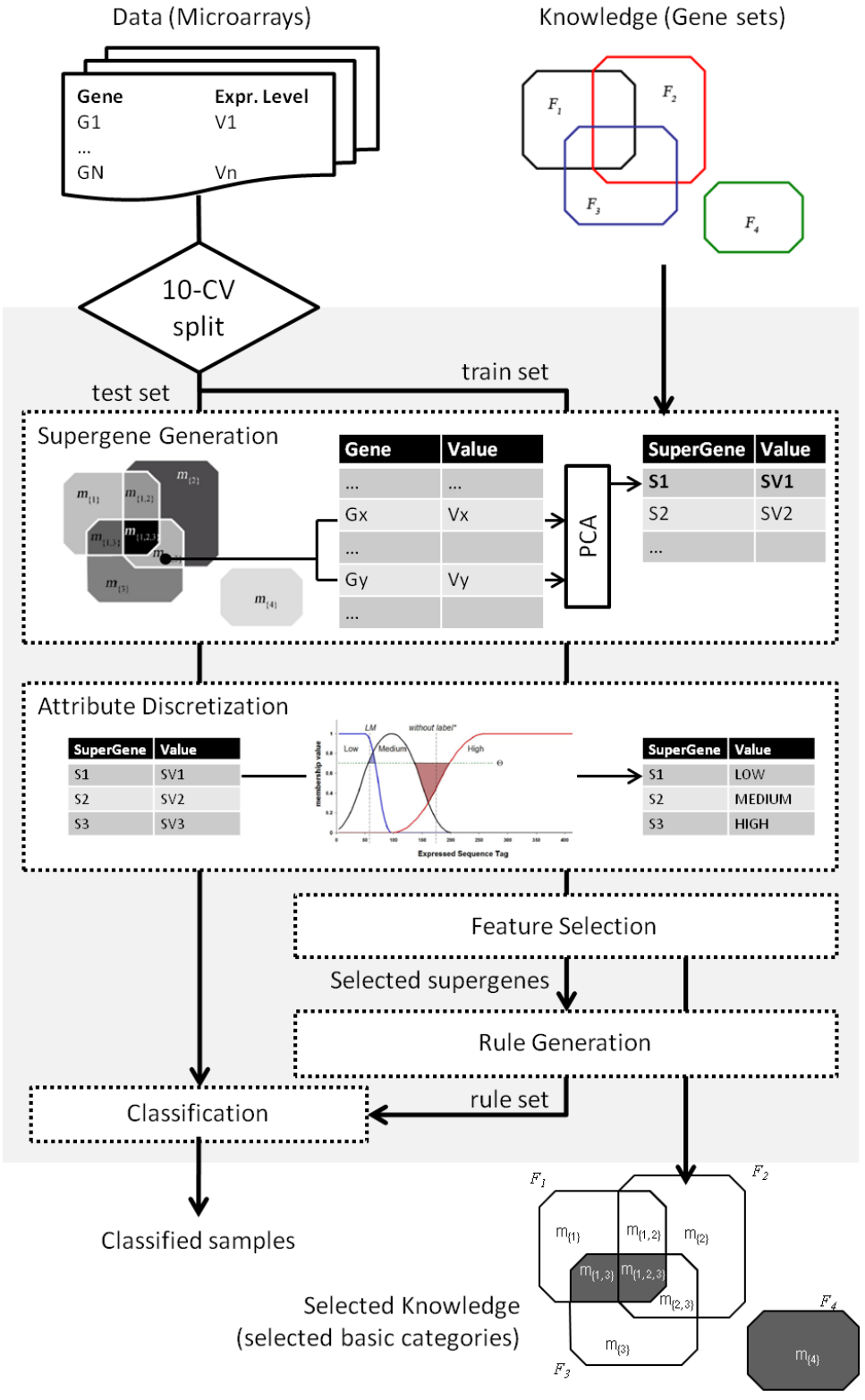

Figure 1: Classification process

are selected by using two methods, the criterion of maximum β-relevance and the VPRS-Q algorithm based on Quickreduct, both supported by VPRS. Then, decision rules are generated using the CAI model. Finally, a classifier is built using the decision rules generated in the pre-

vious step, giving they an order of application based on a score. These steps will be explained in the following sections.

4.1 Supergene generation

The idea of supergenes was introduced by X. Chen and L. Wang [1] and it is a construction that summarizes information from a set of genes like *gene categories*, *pathways*, *gene sets*, or, in this case, *basic categories*. The information summarized from genes is generated using the *principal component analysis* (PCA) method [17].

PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called *principal components*. The number of principal components is less than or equal to the number of original variables but to define a supergene it is only necessary to get the first principal component because it seeks to reduce the dimensionality to a minimum.

Once supergenes representing information from each basic cateogory have been generated, they are used as predictors of the sample class instead using genes.

Let $m_j = \{A_1^j, \ldots, A_i^j, \ldots, A_n^j\}$ denotes the set of *n* features or genes of a given basic category and $\mathbb{M} = \{m_1, \ldots, m_j, \ldots, m_p\}$ denotes the set of *p* basic categories relevant to the class of interest. $\mathbb{S}$ is the set of supergenes generated by the algorithm.

1. Initialize $\mathbb{M} \leftarrow \{m_1, \ldots, m_j, \ldots, m_p\}, \mathbb{S} \leftarrow \emptyset$.
2. Repeat the following two steps for each $m_j \in \mathbb{M}$.
3. If $|m_j| = 1$ then basic category m_j only has one gene a_1^j, so it is not necessary to create a supergene. In effect, $s_j = a_1^j$ and $s_j \in \mathbb{S}$.
4. If $|m_j| > 1$ then basic category m_j has more than one gene ant it is necessary to create a supergene. In effect, $s_j \leftarrow PCA(m_j)$ and $s_j \in \mathbb{S}$.

PCA is the function that implements the principal component analysis and returns the first principal component. PCA is used to find causes of the variability of a data set and sort by relevance.

4.2 Attribute discretization

Rough Sets algorithms works with *Nominal* attributes but gene expression levels are *floating point* values, so it is necessary to discretize values to make Rough Sets applicable to data from DNA microarray. Discretization transforms a continuous range of values in a defined number of bins. Each bin will contain all values of a subrange of values and will represent this range with a discrete value.

Discriminant fuzzy pattern (DFP) is used to discretize supergene values of the output from the previous step. DFP is an extension package for the programming language and statistical environment R [18]. The software has been developed to perform fuzzy analysis and gene reduction

using microarray data. It employs object classes and functions that are also standard in other packages of the Bioconductor project [19]. The whole algorithm comprises of three main steps. First, it represents each gene value in terms of one from the following linguistic labels: Low, Medium, High and their intersections LowMedium and MediumHigh. The output is a *fuzzy microarray descriptor* (FMD) for each existing sample (microarray) containing the discretized gene expression values. The second phase aims to find all genes that best explain each class, constructing a supervised *fuzzy pattern* (FP) for each class (pathology). Starting from the previous generated fuzzy patterns, the package is able to discriminate those genes that can provide a substantial discernibility between existing classes, generating a unique *discriminant fuzzy pattern* (DFP). For this method, only the first phase of DFP algorithm is used to obtain discretized gene expression values from supergenes generated in supergene generation step.

4.3 Feature selection

In real data analysis such as microarray data, the presence of such irrelevant and insignificant features may lead to a reduction in the useful information. Ideally, the selected features should have high relevance with the classes and high significance in the feature set. The features with high relevance are expected to be able to predict the classes of the samples. However, if insignificant features are present in the subset, they may reduce the prediction capability. A feature set with high relevance and high significance enhances the predictive capability [11]. Rough set theory is used to select the most relevant features from supergene Data set. The method of *Max β-relevance* based on CAI model [20] has been defined after failing to apply the method of MRMS because the apply the significance and relevance of all supergenes was zero.

Also another method of feature selection called *Quickreduct* [21] tries to find reducts in a decision table. Intuitively, a *β-reduct* of the set of supergenes $\mathbb{C}$ is its essential part, which is sufficient to define all the basic concepts of data considered, with an classification error less than or equal to β. Family $S \subseteq \mathbb{C}$ is denominated *β-reduct* of $\mathbb{C}$, if and only if S does not contain any dispensable attribute (supergen) and $IND(S) \approx_\beta IND(\mathbb{C})$. The process of determining the reducts of an information system is know to be very expensive in terms of execution time. Its variant called *VPRS-Quickreduct* (VPRS-Q) [22] is applied to the model. The Quickreduct algorithm attemps to get a reduct without generating all posible subsets of attributes. But the output of Quickreduct is not guaranted to be a reduct so this method is also used as a feature selection method. VPRS-Q adds to Quickreduct the capacity of working with β parameter of VPRS.

4.3.1 Maximum beta-relevance

Define $\widehat{r}_\beta(s_i, \mathbb{D})$ as the β-relevance of the supergene s_i with respect to the class labels $\mathbb{D}$. The β-relevance of s_i with respect to $\mathbb{D}$ can be calculated as:

$$\widehat{r}_\beta(s_i, \mathbb{D}) = \frac{|POS_{s_i,\beta}(\mathbb{D})|}{|\mathbb{U}|}, \text{ where } 0 \leq \widehat{r}_\beta(s_i, \mathbb{D}) \leq 1 \quad (4)$$

The purpose of maximum β-relevance is to select the features with classification ability and reject those that interfere with the process.

Let $\mathbb{C} = \{s_1, \ldots, s_i, \ldots, s_n\}$ denotes the set of *m* features of a given supergene data set and $\mathbb{S}$ is the selected genes.

1. Initialize $\mathbb{C} \leftarrow \{s_1, \ldots, s_i, \ldots, s_n\}, \mathbb{S} \leftarrow \emptyset$.
2. Repeat the following two steps until the desired of supergenes is selected or all remaining supergenes have $\widehat{r}_\beta(s_i, \mathbb{D}) = 0$.
3. Calculate the β-relevance $\widehat{r}_\beta(s_i, \mathbb{D})$ of each feature or supergene $s_i \in \mathbb{C}$.
4. Select the feature s_i as the most relevant feature that has the highest value $\widehat{r}_\beta(s_i, \mathbb{D})$. In effect, $s_i \in \mathbb{S}$ and $\mathbb{C} = \mathbb{C} \setminus s_i$.

The β-relevance of a supergene is calculated based on CAI model. The β-relevance $\widehat{r}_\beta(s_i, \mathbb{D})$ of a supergene s_i with respect to the class labels $\mathbb{D}$ is calculated using (4).

4.3.2 VPRS-Q

Define $\gamma_\beta(S, \mathbb{D})$ as the β-dependency of a subset of supergenes S with respect to the decision class label $\mathbb{D}$. The β-dependency of S with respect to $\mathbb{D}$ can be calculated as:

$$\gamma_\beta(S, \mathbb{D}) = \frac{|POS_{S,\beta}(\mathbb{D})|}{|\mathbb{U}|}, \text{ where } 0 \leq \gamma_\beta(S, \mathbb{D}) \leq 1 \tag{5}$$

The goal of ***VPRS-Q*** is to obtain a minimum subset of supergenes with the same or the most approximately β-dependency than the full set S. When the subset has the same β-dependency as the full supergene set then the subset is a β-reduct.

4.4 Decision rule generation

The method used to simplify decision tables under CAI model [23] with the method of *Max β-Relevance* consists of the following steps: Firstly, β-reducts of condition attributes (supergenes) are computated, i.e., remove superfluous supergenes; then, superfluous attribute values are eliminated (β-reducts of categories). This is equivalent to calculate reducts of categories. In the case of VPRS-Q method only second step is performed. After the development of these steps, a set of decision rules is obtained.

4.5 Rule application in classification process

Let $Cover(R_i)$ denote the number of objects of the training set that *support* the decision rule R_i, let $NOC(R_i)$ denote the number of objects with the same decision label as the decision rule R_i, and let $E(R_i)$ denote the classification error (in training set) of the decision rule R_i. Define $s(R_i)$ as the score of the decision rule R_i. The score of R_i can be calculated as:

$$s(R_i) = \frac{Cover(R_i)}{|NOC(R_i)|^2}(1 - E(R_i)) \tag{6}$$

The purpose of decision rule score is to sort rules, placing first rules with more coverage of objects (samples) and with less error.

Let $\mathbb{B} = \{R_1, \ldots, R_i, \ldots, R_n\}$ denotes the set of *n* rules generated from train data, $R_i = \{(s_j^i, v_j^i), \ldots, (s_m^i, v_m^i), (d^i, c^i)\}$ denotes one rule of the set $\mathbb{B}$, and $o = \{(s_k, v_k), \ldots (s_p, v_p)\}$ denotes a sample to be classified with a class label from $\mathbb{D}$.

1. Initialize $\mathbb{B} \leftarrow \{R_1, \ldots, R_i, \ldots, R_n\}$, $\mathbb{S} \leftarrow \emptyset$.
2. Repeat the following step for each $R_i \in \mathbb{B}$.
3. If $\forall (s_j^i, v_j^i) \in R_i, (s_k, v_k) \in o : (s_j^i, v_j^i) = (s_k, v_k)$, the rule R_i matches with the sample o. In effect $R_i \in \mathbb{S}$.
4. Calculate the score $s(R_i)$ of each rule $R_i \in \mathbb{S}$
5. Select the rule R_i as the most scored rule that has the highest value $s(R_i)$. In effect, class of sample o is c^i, where $c^i \in (d^i, c^i)$ and $(d^i, c^i) \in R_i$.

The score $s(R_i)$ of a rule R_i is calculated using (6).

5 Experimental results

5.1 Data sets and experimental process

The performance of the proposed classification techniques are studied and compared with some existing classic classification methods: *Sequential Minimal Optimization* for *Support Vector Machines* [24], *K-nearest neighbors* [25], and *Random Forests* [26]. Proposed classification techniques are implemented in different languages: introduction of biological knowledge and attribute discretization are implemented in R; feature selection and calculation of reducts and rules are implemented in C for the case of *Max β -Relevance* and a Weka (*Waikato Environment for Knowledge Analysis*) [27] Java extension for the case of *VPRS-Q*; and final classifier is implemented in Java. Classification methods for comparison are implemented in Weka. Experiments run in LINUX environment having machine configuration Intel Core i7, 2.80 GHz, 8 MB cache, and 4 GB RAM.

The performance of different algorithms is analyzed doing the experimentation on three data set from microarrays of *breast cancer* samples. One set of basic categories, with one set of genes for each basic category is introduced. The metrics for evaluating the performance of different algorithms are the classification accuracy and the Cohen's kappa coefficient. Kappa coefficient is an agreement measure between classes predicted by a classifier and expected classes [28]. To compute the prediction accuracy of all methods, 10-fold cross-validation is performed.

Different methods are compared using breast cancer data set from microarrays. Breast cancer data set contains expression levels of 12650 genes. Data sets (GSE2034) [29], GSE2990 [30], GSE3494 [31] has been extracted from public database GEO [32]. Samples are classified according to their estrogen receptor (ER) status: active (ER+) or inactive (ER-), an interesting factor in determining the aggressiveness necessary during treatment. The data set has been

normalized and used to evaluate and compare classification methods using a 10-fold cross validation. Data set (GSE2034) [29] has been used to determine the optimal value of β for *Max β-Relevance* method.

The basic category data set was created using some different sources. The first consulted source was SABioscience enterprise (http://www.sabiosciences.com), which has identified relationships of 33 metabolic pathways with cancer. Also ONIM® database has been used, that lists those diseases with a genetic component and their associated genes [33]. 5 genes related to breast cancer have been selected from this source. All this gene sets are combined to create 130 basic categories as one of the inputs of the rough set method for classification. In addition, union of all gene expression levels sets selected from the above sources will restrict the genes expression levels used for classification in the rest of techniques.

At first, an attempt to adapt SPCA [9] was performed. But this method is proposed for samples containing data about the survival of individuals and it did not fit this case. Therefore, the classical PCA method was chosen to be used in place of SPCA. Once the set of 130 supergenes and their respective values for each sample have been obtained, it is necessary discretize to apply rough set theory. The discretization was done using DFP [18] defining 3 ranges of values: *High*, *Medium*, *Low*.

At this point we have a decision table with 130 attributes of condition (supergenes) and 1 decision attribute (ER+ or ER-).

With the method of *Max β-Relevance* defined, it was necessary to establish an optimum value for β (using data set GSE2034). β is a value of imprecision such that $0 \leq \beta \leq 0.5$, and it introduces an error in classification. Therefore, the optimal value for β will be the minimum one that will allow to obtain β-relevant attributes. The optimal value in this case is $\beta = 0.1$. With this value, you get about 6 or 7 relevant attributes. These supergenes, which correspond to some of the basic categories, are the most relevant to the class. If $\beta = 0.15$ the average number of selected supergenes increases to more than 15 and if $\beta = 0.05$ the average number of selected supergenes decreases to less than 1.

Once a few supergenes have been selected (most β-relevant supergenes, or VPRS-Q output), reducts (only with max β-relevance method) and decision rules are computed in the last step. Decision rules serve as a fundamental core of the classifier using the score proposed as a method for sorting rules. Performance of this classifier is evaluated and compared against other conventional methods.

5.2 Analysis of results

As stated above, the classification methods were evaluated using a 10-fold stratified cross-validation scheme, measuring their *accuracy* (well classified samples) and and their *Cohen's kappa coefficient*. Table 1 shows average values of accuracy and kappa for each of the classification methods (using $\beta = 0.1$ for proposed VPRS methods). It also shows the variability of each measure by its standard deviation. Regarding accuracy, the classic methods seem to overperform the Rough sets models, being the SMO the best performing model in the GSE2034 (0.86), and the K-NN the most accurate one in GSE2990 (0.86) and GSE3494 (0.87). However, if we take into account the kappa coefficient, the Rough sets VPRS-Q method was the best one in the GSE2990 (kappa=0.52 and β=0.10) and in GSE3494 (kappa=0.41 and β=0.20),

Table 1: Accuracy and kappa of classifications methods

GSE2034				
	Accuracy		Kappa	
	Average	Std. Dev.	Average	Std. Dev.
Rough Sets (Max β-Relevance)	0.84044	$\pm$ 0.0709	0.44928	$\pm$ 0.1672
Rough Sets (VPRS-Q)	0.78695	$\pm$ 0.0466	0.43783	$\pm$ 0.1421
k-NN	0.82832	$\pm$ 0.0758	0.52627	$\pm$ 0.2365
SMO	**0.86367**	$\pm$ 0.0731	**0.65002**	$\pm$ 0.1941
Random Forests	0.81453	$\pm$ 0.0695	0.48568	$\pm$ 0.2031
GSE2990				
	Accuracy		Kappa	
	Average	Std. Dev.	Average	Std. Dev.
Rough Sets (Max β-Relevance)	0.83158	$\pm$ 0.0804	0.49898	$\pm$ 0.1878
Rough Sets (VPRS-Q)	0.85381	$\pm$ 0.0739	**0.52423**	$\pm$ 0.2430
k-NN	**0.86374**	$\pm$ 0.0813	0.48083	$\pm$ 0.2593
SMO	0.83041	$\pm$ 0.0664	0.41626	$\pm$ 0.1989
Random Forests	0.86174	$\pm$ 0.0709	0.47729	$\pm$ 0.1878
GSE3494				
	Accuracy		Kappa	
	Average	Std. Dev.	Average	Std. Dev.
Rough Sets (Max β-Relevance)	0.82217	$\pm$ 0.0542	**0.39905**	$\pm$ 0.1507
Rough Sets (VPRS-Q)	0.82650	$\pm$ 0.0669	0.26413	$\pm$ 0.2328
k-NN	**0.87083**	$\pm$ 0.0580	0.26239	$\pm$ 0.2358
SMO	0.84617	$\pm$ 0.0652	0.31589	$\pm$ 0.3120
Random Forests	0.86667	$\pm$ 0.0498	0.25666	$\pm$ 0.1729

whereas the Max β-Relevance also outperformed the classical models in the GSE2990 dataset (kappa=0.49) and in the GSE3494 dataset (kappa=0.39).

However, we have carried out an ANOVA test in order to find significant differences among the compared models. Tests have shown that there are not any significant differences between models regarding the accuracy (p-value = 0.3392) nor regarding the kappa coefficient (p-value = 0.8192). Figures 2 and 3 show the accuracy and kappa for each model in each dataset, as well as their general behaviour in all datasets.

Regarding the selected knowledge, Table 2 shows the most frequent supergenes selected after the feature selection phase, over the three breast cancer datasets. The frequency represents how many times a given supergene has been selected in the 10 * 3 = 30 tran-test cycles that have been performed.

We have found in bibliography that many of these genes have been reported as being related to the estrogen receptor status. For example, the gene CCND2 has been reported in [34] as being significantly more methylated (under-expressed) in ER-positive than in ER-negative tumors. The gene TFDP2 has been found to be highly expressed in ER- tumors in [35]. The IGF1R and his interaction with the estrogen receptor signaling pathway has been also studied in [36]. The RHO proteins were shown in [37] to have a dramatic impact on ER transcriptional activity. The

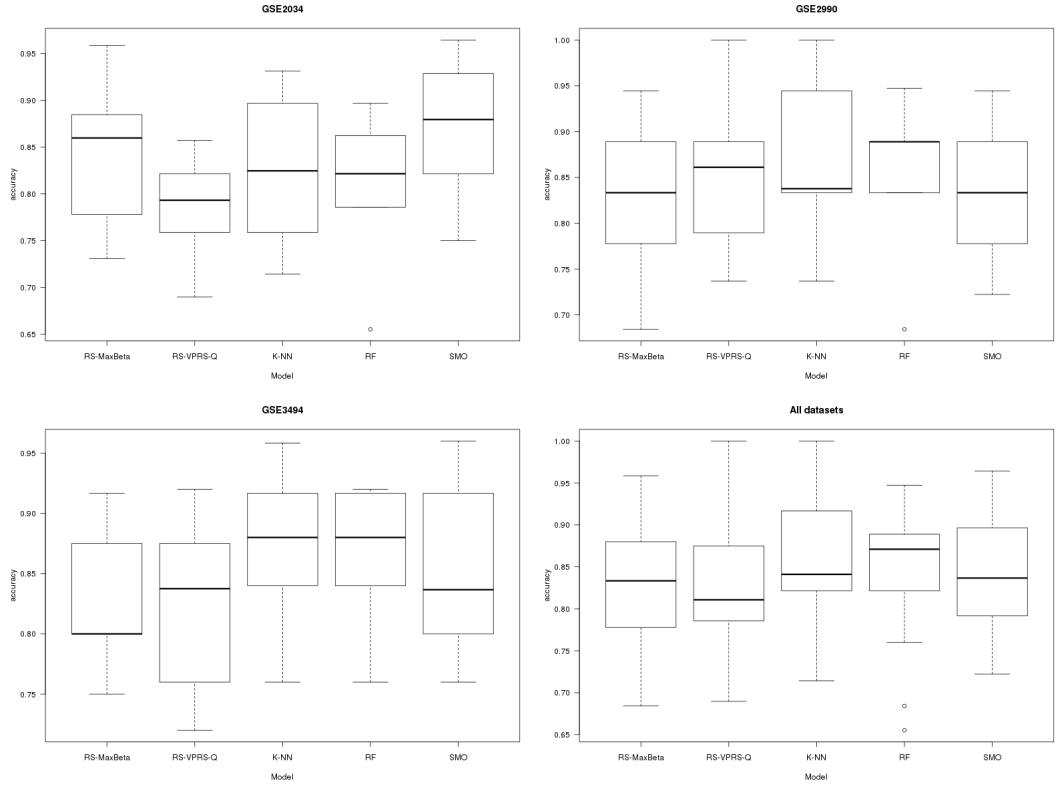

Figure 2: Comparison of the models accuracy in each dataset and in all datasets

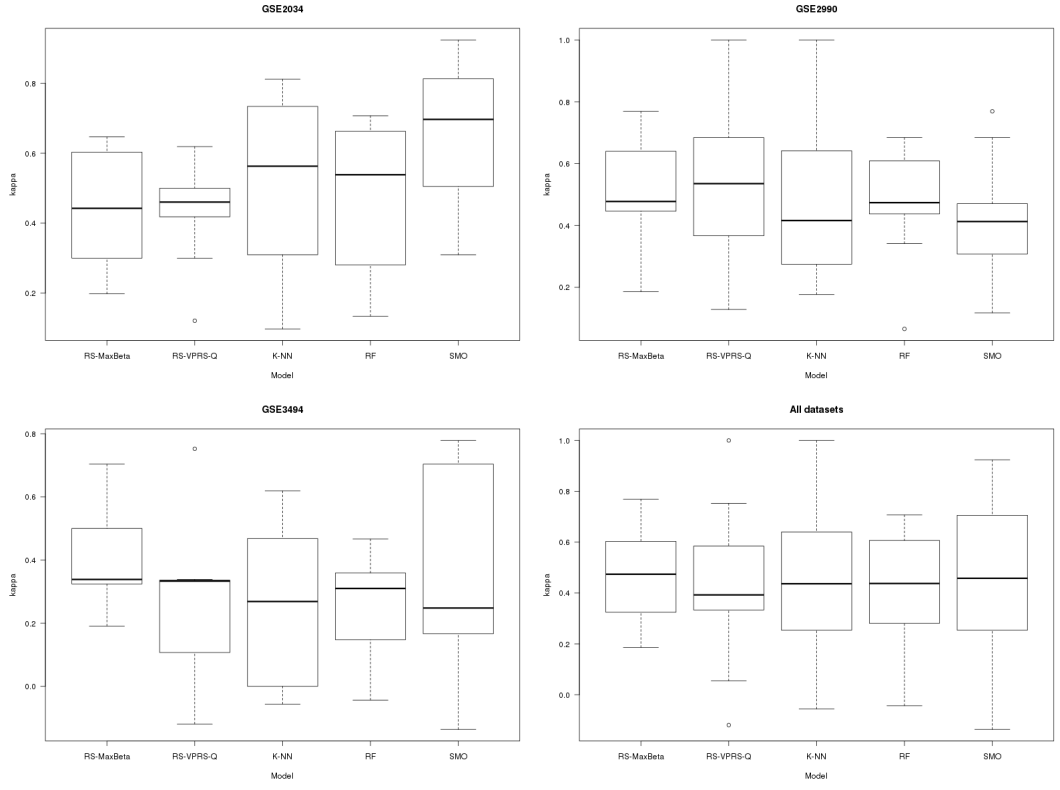

Figure 3: Comparison of the models kappa coefficient in each dataset and in all datasets

Table 2: Top selected supergenes

Supergene	Genes	Frequency
Regulation of the Cell Cycle ⋂ Factors Involved in Other Aspects of the Cell Cycle	CCND2 CCND3 CCNE1 CDC20 E2F3 RBL2 TFDP2	63.33%
Insulin Receptor Signaling Pathway ⋂ Regulation of the Cell Cycle ⋂ Anti Apoptosis ⋂ Regulation of Cell Proliferation Differentiation Growth and Volume	IGF1R	36.66%
Growth Factors	GDF15	36.66%
Rho Protein Signal Transduction ⋂ Small GTPase-Mediated Signal Transduction	RHOD	20.00%
Regulation of the Cell Cycle ⋂ Cell Growth and Maintenance	AXL	16.66%
Insulin Receptor Signaling Pathway	PDPK1	16.66%
Factors Involved in Other Aspects of Apoptosis ⋂ Regulation of the Cell Cycle ⋂ Regulation of Apoptosis ⋂ Anti Apoptosis ⋂ Cell Growth and Maintenance ⋂ Regulation of Cell Proliferation Differentiation Growth and Volume	BCL2	16.66%
Intracellular Signaling Cascade ⋂ Cell Proliferation ⋂ Cell Growth and Maintenance	FES	16.66%
Regulation of the Cell Cycle ⋂ OMIM Breast Cancer ⋂ OMIM Breast Cancer susceptibility to	BRCA2	16.66%

AXL has been recently reported in [38] as being overexpressed in lapatinib-resistant ER+ tumor cells. The overexpression of PDPK1 conferres resistance to chemotherapy in breast cancer as it has been shown in [39], which is a typical phenotype of negative ER status. Finally, it has been shown that the BCL-2 expression levels correlates with ER positivity [40]. The FES gene interacts with the BCAR1 gene [41], which has been shown to be involved with the antiestrogen resistance in breast cancer cells [42]. Finally, BRCA2 is one of well-known the breast cancer related genes.

It should also be pointed out that the most frequent supergenes were not those derived from the breast-cancer related gene sets taken from the OMIM database. This could be explained from the fact that those gene sets are general breast cancer related genes, but not those that are differentially expressed when the ER status is the studied condition.

6 Conclusions and Future Work

In this paper, we have presented a novel model which integrates explicit biological knowledge into classication process using Variable Precision Rough Set Theory (VPRS). The knowledge is given by the user in the form of gene sets, configuring an interpretation context for the

microarray data.

The interpretation context is divided into basic categories allowing us to transform the genes of the input data into supergenes, via PCA, whose values are also discretized. The most promising supergenes are then selected via two alternative methods: Max β-relevance and VPRS-Quickreduct. Finally a set of decision rules is generated from the selected supergenes. These set of rules can be given to the user. Since they are a set of conditions over the selected supergenes or basic categories, they can be easily interpretated by the biomedical expert.

We have tested our models over three breast cancer datasets, aiming at predicting the ER status of tumors. Having 33 cancer-related pathways and a few breast-cancer gene sets taken from OMIM as explicit biological knowledge, we have trained and tested the model over each dataset. We have concluded that there are not significant differences between our model and three classical classification models (KNN, Random Forest and SMO). However, our model is able to provide a more biological-interpretable explanation of how it classifies new samples. In addition, we have found that most of the genes contained in the frequently selected supergenes are reported in the literature as being ER status-related genes.

Our future work will be focused at (*i*) including a mechanism to automatically select the best β value, (*ii*) testing the model in a inter-dataset scenario, i.e. train the model with one dataset and test it over another, aiming at assessing the robustness capabilities derived from the introduction of biological knowledge and (*iii*) using a non parametric test to analyze the differences among groups, i.e. Kruskal-Wallis test.

Acknowledgements

This work was partially funded by the (*i*) TIN2009-14057-C03-02 project from the Spanish Ministry of Science and Innovation, the Plan E from the Spanish Government and the European Union from the ERDF and (*ii*) the integrated action AIB2010PT-00353 from the Spanish Ministry of Science and Innovation.

References

[1] X. Chen and L. Wang. Integrating biological knowledge with gene expression profiles for survival prediction of cancer. *Computational Biology*, 16(2):265–278, 2009.

[2] Z. Wei and H. Li. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*, 8:265–284, 2007.

[3] F. Tai and W. Pan. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23(23):3170–3177, 2007.

[4] F. Tai and W. Pan. Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics*, 23(14):1775–1782, 2007.

[5] R. Tibshirani et al. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science*, 18(1):104–117, 2003.

[6] S. Wold et al. Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

[7] M. Kanehisa et al. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* 36:480–484, 2008

[8] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. *International Conference on Machine Learning*, pages 148–156, 1996.

[9] X. Chen et al. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics*, 24(21):2474–2481, 2008.

[10] W. Zhou et al. Feature selection for microarray data analysis using mutual information and rough set theory. *AIAI*, 204:492–499. Springer, 2006.

[11] P. Maji and S. Paul. Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. *Int. J. Approx. Reasoning*, 52(3):408–426, 2011.

[12] Z. Pawlak. Rough sets. *International Journal of Computer and Information Sciences*, 11(5):341–356, 1982.

[13] W. Ziarko. Variable precision rough set model. *Computer and System Sciences*, 46:39–59, 1993.

[14] J. F. Galvez et al. An application for knowledge discovery based on a revision of vprs model. *Rough Sets and Current Trends in Computing*, 2005:296–303, 2001.

[15] I. Fodor. A survey of dimension reduction techniques. *Tech. Rep., Lawrence Livermore National Laboratory*, 2002.

[16] D. Glez-Pena. Modelo para la integración de conocimiento biológico explícito en técnicas de clasificación aplicadas a datos procedentes de microarrays de ADN. *PhD thesis*, University of Vigo, 2009.

[17] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[18] D. Glez-Pena et al. Dfp: a bioconductor package for fuzzy profile identification and gene reduction of microarray data. *BMC Bioinformatics*, 10(1):37, 2009.

[19] R. Gentleman et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):80, 2004.

[20] D. Calvo-Dmgz et al. Biologial knowledge integration in DNA microarray gene expression classification. *Advances in Soft Computing* 154:53–61, 2012

[21] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorization. *Applied Artifcial Intelligence*, 15(9):843–873, 2001.

[22] X. Pan et al. A variable precision rough set approach to the remote sensing land use/cover classification. *Computer & Geosciences* 36(12):1466–1473, 2010.

[23] J. F. Galvez et al. An improved algorithm for determining reducts in rough set models. *Proceedings of the IASTED*, 2003.

[24] J. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods — Support Vector Learning* (Cambridge, MA), pages 185–208, MIT Press, 1999.

[25] E. Fix and J. L. Hodges. Discriminatory analysis – nonparametric discrimination: Consistency propertie. *Tech. Rep. Project 21-49-004, Report No. 4, pages 261–279, USAF School of Aviation Medicine, Randolph Field, Texas* 1951.

[26] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[27] M. Hall et al. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[28] A. Ben-David. Comparison of classification accuracy using cohen's weighted kappa. *Expert Syst. Appl.*, 34(2):825–832, 2008.

[29] Y. Wang et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365:671–679, 2005.

[30] C. Sotiriou et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98:262–272, 2006.

[31] L. D. Miller et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. U.S.A.* 102(38):13550-13555, 2005.

[32] R. Edgar et al. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[33] J. S. Amberger et al. Mckusick's online mendelian inheritance in man (OMIM®). *Nucleic Acids Research*, 37:793–796, 2009.

[34] E. Sunami et al. Estrogen receptor and HER2/neu status affect epigenetic differences of tumor-related genes in primary breast tumors. *Breast Cancer Res.*, 10(3):R46, 2008.

[35] M. C. Alles et al. Meta-analysis and gene set enrichment relative to er status reveal elevated activity of MYC and E2F in the "basal" breast cancer subgroup. *PLoS One*, 4(3):e4710, 2009.

[36] D. H Fagan and D. Yee. Crosstalk between IGF1R and estrogen receptor signaling in breast cancer. *J. Mammary Gland Biol. Neoplasia*, 13(4):423–429, 2008.

[37] G. Huet et al. Repression of the estrogen receptor-alpha transcriptional activity by the Rho/megakaryoblastic leukemia 1 signaling pathway. *J. Biol. Chem.*, 284(49):33729-22739, 2009.

[38] L. Liu et al. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. *Cancer Res.*, 49(17):6871–6878, 2009.

[39] K. Liang et al. Differential roles of phosphoinositide-dependent protein kinase-1 and akt1 expression and phosphorylation in breast cancer cell resistance to Paclitaxel, Doxorubicin, and gemcitabine. *Mol Pharmacol.*, 70(3):1045–1052, 2006.

[40] R. D Leek et al. bcl-2 in normal human breast and carcinoma, association with oestrogen receptor-positive, epidermal growth factor receptor-negative tumours and in situ cancer. *Br. J. Cancer*, 69(1):135–139, 1994

[41] M. Jücker et al. The Fes protein-tyrosine kinase phosphorylates a subset of macrophage proteins that are involved in cell adhesion and cell-cell signaling. *J. Biol. Chem.*, 272(4):2104–2109, 1997.

[42] A. Brinkman et al. BCAR1, a human homologue of the adapter protein p130Cas, and antiestrogen resistance in breast cancer cells. *J. Natl. Cancer. Inst.*, 92(2):112–120, 2000.