Classification of cancer cell death with spectral dimensionality reduction and generalized eigenvalues

Mario R. Guarracino[a], Petros Xanthopoulos[b], Georgios Pyrgiotakis[c], Vera Tomaino[b,d], Brij M. Moudgil[c], Panos M. Pardalos[b]

[a] *High Performance Computing and Networking Institute - National Research Council (ICAR-CNR), Via P. Castellino, 111 - 80131 Naples, Italy*
[b] *Center for Applied Optimization, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611-6595 USA*
[c] *Particle Engineering Research Center, University of Florida, Gainesville, FL, 32611 USA*
[d] *Department of Experimental Medicine and Clinic, University Magna Græcia of Catanzaro, Italy*

Abstract

Objective Accurate cell death discrimination is a time consuming and expensive process that can only be performed in biological laboratories. Nevertheless, it is very useful and arises in many biological and medical applications.

Materials and methods Raman spectra are collected for 84 samples of A549 cell line (human lung cancer epithelia cells) that has been exposed to toxins to simulate the necrotic and apoptotic death. The proposed data mining approach for the multiclass cell death discrimination problem uses a generalized eigenvalue algorithm for classification, together with a novel dimensionality reduction algorithm based on spectral clustering.

Email addresses: mario.guarracino@cnr.it (Mario R. Guarracino), petrosx@ufl.edu (Petros Xanthopoulos), gpyrgiotakis@perc.ufl.edu (Georgios Pyrgiotakis), vera.tomaino@ufl.edu (Vera Tomaino), bmoudgil@perc.ufl.edu (Brij M. Moudgil), pardalos@ufl.edu (Panos M. Pardalos)

Results The proposed algorithmic scheme can classify A549 lung cancer cells from three different classes (apoptotic death, necrotic death and control cells) with $97.78\% \pm 0.047$ accuracy. Further evidence of the validity of the technique is obtained with the hyperthermia example.
Keywords: Raman spectroscopy, spectral clustering, dimensionality reduction, generalized eigenvalue classification

1. Introduction

Cell classification has recently emerged as one of the most critical problems in modern biology. It has a wide range of applications: cell line purity[1], stem cells differentiation[2], cancer diagnosis and cell death[3] to mention a few. In case of cell death, the assessment of the cell pathway is critical when it comes to in-vitro trials of chemotherapeutic drugs[4,5] and substance toxicity[6]. To this day, the most accurate way of discriminating among various cells is through genetic profiling (gene expression)[7] or identification with biomarkers, which can take long time, requires expensive equipment and trained personnel[8]. In cases where the cell type or cell itself needs to be recognized immediately, researchers rely on the morphological features of the cell. The latter strategy, however, is highly subjective and prone to errors.

One of the recently emerging techniques that has potential for cell discrimination is Raman spectroscopy[9]. The advantage of this method over Fourier Transform Infrared (FTIR) spectroscopy, or other vibrational techniques, is the very little interference of water, which is the basic constituent of the cells. In addition, it does not need any marker, or chemical. Furthermore, it can be done in-vitro or in-vivo and it is non-invasive. Finally, it is

relatively rapid (10-30s per acquisition) and it can be used outside laboratories[10]. Raman spectroscopy is based on the scattering of a laser light in the interaction with a sample. In the case of cells, the obtained spectrum is a set of frequencies changed by the interactions with DNA/RNA, proteins, lipids and amino acids[11].

Recently, there have been attempts to integrate statistical techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)[12,9]. Moving to more advanced methodologies, Widjaja et al.[13] classified colonic tissue and Pyrgiotakis et al.[14] classified cell death by integrating Support Vector Machines (SVM) into Raman analysis. In the latter work, even if SVM was used only on binary classification problems, comparing pairs of classes, it gave more insight to data analysis.

In the present work we examine the problem of cell death discrimination among necrotic, apoptotic and control cells as a three class supervised classification problem. Raman spectra are collected for A549 cell line (human lung cancer epithelia cells) that has been exposed to toxins to simulate the necrotic and apoptotic death.

We reduce the dimensionality of the problem, obtaining a smaller subset of the original wavelengths (features) that still allows to discriminate the samples into three classes. To this extend, the dimensionality reduction technique we introduce is based on spectral clustering, a well studied clustering technique. Here, the name *spectral* refers to the use of an eigenvector technique and it has not to be confused with the Raman spectra of cells, which are the type of data we analyze in the present work. The data are then classified with the multi-class Regularized Generalized Eigenvalue Classifica-

tion (multiReGEC)[15], an algorithm based on generalized eigenvalues[16]. The advantages of the obtained algorithm relies in its accuracy in detecting cell death, and in the capability to handle multi-class problems.

A word about notation. Matrices are indicated with capital letters and vectors with small letters. Vectors are always column vectors. The transpose of a vector x is x^T, the transpose of a matrix A is A^T. With C_i we denote the i-th row of C, while a column vector of ones of arbitrary dimension is denoted with 1.

The rest of the work is organized as follows: In section 2 we describe the dimensionality reduction technique and the classification algorithm used. In section 3 we present results and we provide a biological validation of the method. Last section consists of concluding remarks.

2. Materials and methods

2.1. Cell line and protocols

For this set of experiments the A549 lung epithelia cells are used (from ATCC; cell line number CCL-185)[17,18]. They have been selected due to the high resilience to harsh conditions.

The growth media is made of 89% RPMI-1640 with L-glutamine (from Cellgro; Cat #: 25-053-CI), 10% Fetal Bovine Serum (four times filtered through 0.1 μm filter, from Hyclone; Cat. #: SH30070.03) and 1% antibiotic-antimycotic solution (from Cellgro; Cat. #: 30-004-CL). The cells were cultured and plated according to the ATCC protocols[18]. The cell count is done with the ViCell from Beckmann-Coulter (Fullerton, CA). For these experiments the cells were seeded on a 5×5 mm MgF_2 substrate at approximately

$\sim 5 \times 10^3$ cells. The MgF_2 crystal is used to reduce the background radiation from the petri-dish. The seeded MgF_2 substrates are placed in a 6 well plate (9.6 cm^2 per well) and were let in the incubator for 45 min, sufficient for the cells to attach on the MgF_2. Following, 3 ml of growth media is added and the seeded cells are incubated at 37℃ and 5% CO_2 for a minimum of 24 hours before the toxin dosing.

2.2. Toxic agents standards and dosing

The triggers for the two different cellular deaths are, etoposide (apoptosis)[19] and Triton X-100 (necrosis)[20]. Etoposide is insoluble in water, so a stock solution is prepared with 100mM of etoposide in di-Methyl-sulfo-oxide (DMSO).

The concentration of the toxins was based on the previous work, that suggests these values will impact the cells, but not catastrophically. For the experiments, the agents concentration is 100 μM for Triton-X[21] and 80 μM[22,19,9] for the etoposide. These concentrations are expected to induce damage in the cells without completely lysing the cells in the first 24 hours of the experiment. The solution is prepared immediately prior to dosing.

2.3. Toxic agent dosing

After reaching 80% confluency (on the MgF_2 plate) the growth media is removed and the cells are rinsed twice with Hanks' Balanced Salt Solution (HBSS) to remove traces of proteins. Then, 2ml of media containing the toxic agent are added and the cells are moved in the incubator. The absence of proteins does not have any effect on the cells for the time period the experiments last (approximately 1 hour)[23].

2.4. Raman data acquisition

The Raman microscope used is the InVia system by Renishaw, consists of a Leica microscope connected to a Renishaw 2000 spectrometer. The high power diode laser (250 mW) produces laser light of 785 nm and does not cause any damage to the cells even after 40 minutes of exposure time. The MgF_2 plate after rinsing with the HBSS is moved on a Delta T Culture Dish (from Biotechs; Cat #: 04200415C), and 2 ml of RPMI 1640 is added. The dish is placed onto a heating stage (Delta T4 Culture Dish Controller, Biotechs, Butler, PA, USA) to regulate the temperature. The exact procedure and data acquisition has been discussed extensively elsewhere[14,3,24]. In this study, heat treatment at 45℃ over 30 minutes is used to test the predictive strength of the model by using a different cell death trigger that would induce a form of programmed cell death. The heating stage Delta T4 Culture Dish Controller is used and the ramping rate is approximately 0.5℃/min. In Fig. 1 we present the acquired spectra, which consists of 1301 wavelengths for each of the 84 samples. Each sample corresponds to a single cell. The acquisition of Raman spectra has been conducted at the Particle Engineering Research Center, University of Florida.

Figure 1: Spectrum of data: control cells are displayed in blue, apoptotic in red, necrotic in green and heat cells in black. Data have been normalized and shifted for clarity.

2.5. Method description

In this section we detail the proposed dimensionality reduction technique and the supervised classification method.

2.5.1. Spectral based dimensionality reduction

The main idea of dimensionality reduction is to condense the initial features into a sufficiently small subset that retains all information needed to solve the problem.

Spectral clustering has been extensively used for image[25] and video[26] segmentation, biomedical imaging[27], social network clustering[28] and biological network clustering[29]. Spectral clustering is usually used to cluster samples. In this work, we use it to cluster features, in such a way one of the clusters contains the most representative features for the problem. Here the word spectral refers to the use of eigenvalues and eigenvectors theory rather than to Raman spectra of data.

Spectral clustering is based on the notion of *normalized cut* partitioning of a graph. Let $G(N, E)$ be a weighted graph, where N is the set of vertices and E the set of edges. Then, the adjacency matrix $W = (w_{ij})_{n\times n}$, for an undirected graph, is given by:

$$w_{ij} = \begin{cases} c_{ij} & i \neq j \\ 0 & i = j \end{cases}, \tag{1}$$

where c_{ij} is the weight associated to the edge between vertices i and j. Given a partition $p_1, \ldots, p_k$ of N, we can define:

$$E(p, q) = \sum_{i\in p, j\in q} w_{i,j},$$

and

$$cut(p_i, \bar{p_i}) = \frac{1}{2}\sum_{i=1}^{k} E(p_i, \bar{p_i}).$$

where $\bar{p_i}$ is the complement of p_i.

Figure 2: Heatmap of the pairwise correlation among all features. The two axes labels correspond to features and the color of the heatmap corresponds to the value of the pairwise correlation between features. The majority of the feature pairs are highly correlated and therefore the initial dataset possesses highly redundant information.

To partition the graph in k parts, it is possible to choose the partition $p_1, \ldots, p_k$ which minimizes the *normalized cut*:

$$\min_{p_1,\ldots,p_k} \sum_{i=1}^{k} \frac{cut(p_i, \bar{p}_i)}{vol(p_i)} \tag{2}$$

where $vol(p_i)$ is the sum over the weights of all edges attached to vertices in p_i.

It can be shown that finding the minimum of normalized cut (2) of a network is equivalent to solve a generalized eigenvalue problem subject to binary constraints, which is an NP hard problem[25]. Relaxing the binary constraints of the problem, it becomes an eigenvalue problem. The clustering of the first k eigenvectors of this problem provides a clustering of the original dataset, with the properties of the minimum normalized cut.

In the case of Raman data, the number of features (wavelengths) describing the problem is much larger than the number of samples (cells). Therefore, the information contained in the features is redundant. Furthermore, the analysis of the heatmap of variables pairwise correlation, as shown in Fig. 2, provides an insight in the way they can be analyzed. Since all features have a pairwise correlation greater than 0.9, the idea is to identify a cluster of features that are as much as possible uncorrelated with the remaining ones, and therefore have maximum descriptive capability with minimum redundancy.

In this sense, every feature corresponds to a node in the graph and the

correlation between two features is represented by c_{ij} in the similarity matrix. To our knowledge, we are the first to propose the use of correlation in the spectral clustering of features.

We can now describe the algorithm. First, we construct the normalized adjacency matrix $\hat{W}$ of the data:

$$\hat{W} = D^{-1/2} W D^{-1/2}. \tag{3}$$

being D the diagonal matrix with elements $d_j = \sum_{i=1}^{n} c_{ij}$. Next, we form the matrix $\hat{V}$ from the eigenvectors related to the m largest eigenvalues of $\hat{W}$. Every row of this matrix can be seen as one point in the $m-$dimensional space. In this study, we apply *k-means* clustering to the first m eigenvectors of the normalized similarity matrix, to obtain the clustering of the initial features. For example, in Fig. 3a we can see the features embedded in the eigenspace spanned by the three most important eigenvectors ($m = 3$) of the normalized adjacency matrix. The color of each point represents the volume d_j of the related feature j. As stated before, the idea of this dimensionality reduction algorithm is to identify clusters of features with the lowest correlation with features outside the cluster. In Fig. 3b we can see the k-means clustering results (for $k = 5$). We note that cluster of features have similar values of the average correlation d_j.

We proceed by selecting the cluster containing the features that have lowest inter-correlation. In other words, for every cluster p_i in the partitioning $\mathcal{P} = \{p_1, p_2, \ldots, p_k\}$, we compute the metric:

$$s_i = \sum_{j \in p_i} d_j \tag{4}$$

(a)

(b)

Figure 3: Features embedded in the eigenspace spanned by the three most important eigenvectors of the normalized adjacency matrix. In (a) points are colored with the average correlation (per point) whereas in (b) we present the clusters found by k-means algorithm ($k = 5$).

and we select the cluster $p_{i_{select}}$ where $i_{select} = \arg\min_{1 \leq i \leq k} s_i$. We repeat the process for different values of k in a given interval $[low_k, high_k]$, and select the features that belong to the cluster with the lowest score (4). The pseudocode is given in Algorithm 1.

2.5.2. MultiReGEC

State of the art classification methods are based on the idea that every discrimination problem can be described as a separation problem[30]. This class of algorithms is usually referred as *Support Vector Machines* (SVM). The strength of these methods relies in their background theory, based on optimization[31] and statistical learning theory[32]. SVM have been successfully applied to many biomedical problems[33,34] and many software implementations are freely available for different problem environments and data mining suites like Matlab, R and Weka.

Given two classes of linearly separable points, represented by the rows of $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times n}$, SVM find two parallel hyperplanes $x^T\omega - \gamma = \pm 1$, with $\omega \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$, leaving all points of the two classes A and B on different sides, and whose distance is maximum. This can be formulated as a convex quadratic optimization problem[35]. In case of nonlinearly separable

Algorithm 1 Spectral Clustering Dimensionality Reduction Algorithm

Compute W from A as in Eq. 1
Compute $\hat{W}$ as in Eq. 3
Compute V by eigendecomposing $\hat{W}$
$S_{old} \leftarrow +\infty$
Form $\hat{V}$ from the first m columns of V
for k= low_k to high_k **do**
    $\mathcal{P}$=k-means($\hat{V}$,k)
    for every $p_i \in \mathcal{P}$ **do**
        Compute s_i from Eq. 4
    end for
    $i_{select} \leftarrow \arg\min_{1 \leq i \leq k} s_i$
    if $s_{i_{select}} < S_{old}$ **then**
        $S_{old} \leftarrow s_{i_{select}}$
        $i_{opt} \leftarrow i_{select}$
    end if
end for$F_{selected} = p_{i_{opt}}$
return $F_{selected}$

classes, it is possible to use a nonlinear transformation to embed points in a higher dimensional space, and search for a linear discriminant in that space. These nonlinear transformations are usually referred as *kernel functions* and they have been successfully applied to solve many problems[36]. Among kernel functions, there are *Gaussian* kernels, which provide a dissimilarity measure between any two points x_1 and x_2, given by:

$$K(x_1, x_2) = e^{-\frac{\|x_1 - x_2\|^2}{\sigma}}, \tag{5}$$

where σ is the parameter used to adjust the interval in which the function is non-zero.

More recently, Mangasarian et al.[37] state the classification problem so that the hypothesis of parallel hyperplanes is no longer needed. In that new formulation, the solution hyperplane $x^T\omega - \gamma = 0$ for class A is the closest to that class, and the furthest from B. This can be formulated as:

$$\min_{\omega,\gamma\neq 0} \frac{\|A\omega - 1\gamma\|^2}{\|B\omega - 1\gamma\|^2}, \tag{6}$$

Now, let:

$$G = [A \quad -1]^T[A \quad -1], \quad H = [B \quad -1]^T[B \quad -1], \tag{7}$$

where $[A \quad -1]$ and $[B \quad -1]$ are the matrices obtained concatenating the vector -1 to A and B. The problem (6) becomes:

$$\min_{z\neq 0} \frac{z^T G z}{z^T H z}, \tag{8}$$

with $z = \begin{bmatrix}\omega^T & \gamma\end{bmatrix}^T$. This is the Raleigh quotient of the generalized eigenvalue problem $Gz = \lambda Hz$. Since both G and H are symmetric, a real nontrivial

solution exists and it is unique if both G and H have full rank. Since G and H have been built using equation (7) their rank is at most $\max(m, p)$, while their dimension is $n+1$, with $n \gg \max(m, p)$. To overcome this problem, we applied a regularization technique firstly proposed in[16], obtaining a method called *Regularized Generalized Eigenvalue Classifier* (ReGEC).

The ReGEC algorithm can be used in conjunction with kernel methods. If $C \in \mathbb{R}^{(p+m)\times n}$ is the matrix obtained concatenating the rows of B to A, and $x \in \mathbb{R}^n$, then each component i of the transformed point x in the nonlinear space is:

$$y_i = K(x^T, C) = e^{-\frac{\|x-C_i\|^2}{\sigma}},$$

where C_i is the i-th row of C. The kernel version of the algorithm now requires to compute the proximal surface $K(x^T, C)r - \eta = 0$, with $r \in \mathbb{R}^{m+p}$ and $\eta \in \mathbb{R}$ as solution of:

$$\min_{r,\eta\neq 0} \frac{\|K(A, C)r - 1\eta\|^2 + \epsilon\|K_B r - 1\eta\|^2}{\|K(B, C)r - 1\eta\|^2 + \epsilon\|K_A r - 1\eta\|^2}. \tag{9}$$

In the previous equation, K_A and K_B are the main diagonals of $K(A, C)$ and $K(B, C)$, and ϵ is the regularization parameter. To decide the class of a new test point, its distance from the two hyperplanes is computed and it is assigned to the class related to the closest hyperplane.

To generalize a binary classifier to c classes, it is possible to solve $c(c-1)/2$ binary tasks, in which $c-1$ classifiers are built for each class against the remaining ones. Then, the $c-1$ models are merged to obtain one model for each class. This strategy, usually called *one-against-one*, well suits our needs, because once ReGEC has produced the $c-1$ planes, for any given class, we can merge them to obtain an *average* plane. The normal vector

of that average plane can be computed averaging the normal vectors of the $c-1$ planes. Since each plane has a single normal direction, the problem has unique solution. The averaging of normal vectors is produced with a Singular Value Decomposition (SVD) applied to the $c-1$ normal vectors computed for each class. The singular vector related to the maximum singular value represents the direction that minimizes the variance of the normal vectors. Algorithm 1 depicts this one-against-one procedure to obtain the average planes in case of c classes.

Algorithm 2 Regularized Generalized Eigenvalue Classification

1: **for** each class A_i, $i = 1, \ldots, c$ **do**

2: Compute the $c-1$ discriminating hyperplanes $K(x^T, C)r_{i,l} - \eta_{i,l} = 0$, $l = 1, \ldots, i-1, i+1, \ldots, c$ and let $Z_i = [z_{i,1}, \ldots, z_{i,i-1}, z_{i,i+1}, \ldots, z_{i,c}]$, with $z_{i,l} = [r_{i,l}^T \quad \eta_{i,l}]^T$

3: Compute the normalized vectors $\bar{R}_i = [\bar{r}_{i,1}, \ldots, \bar{r}_{i,i-1}, \bar{r}_{i,i+1}, \ldots, \bar{r}_{i,k}]$, with $\bar{r}_{i,l} = r_{i,l}/\|r_{i,l}\|$.

4: Evaluate the SVD decomposition of $\bar{R}_i$: $\bar{R}_i = USV^T$.

5: Compute the coefficients of mean hyperplane $[\tilde{r}_i \quad \tilde{\gamma}_i]^T = Z_i U_1$.

6: **end for**

In the SVD, S is the diagonal matrix containing the singular values in decreasing order, U is a basis formed by the eigenvectors of the data space, and V spans the class space. The column vector U_1, where U_1 is the first column of U, represents the direction along which the angles with the normal vectors $\bar{R}_i$ of the planes are minimum. The linear combination of the plane coefficients Z_i with constants U_1 represents the hyperplane with maximum correlation with the Z_i planes. The detailed description of the algorithm and

the results for the linear case are given in[15].

3. Results and Discussion

Performance results are carried out on an Intel Pentium 4 3.40 GHz, 2 GB of RAM running Windows XP, with Matlab 7.5. Spectral data represent 84 cells with 1301 features, with 28 samples in each of the three classes: control, apoptotic and necrotic.

For cross-validation, we use repeated random sub-sampling, which has the advantage over the classical k-fold cross-validation that the number of repetitions are independent of the number of folds[38]. The number of repetitions q_1 has been fixed to 100 for all cross-validation tests and $q_2 = 1000$ iterations for the computation of p-values. The training set is composed of 90% of the available points. The remaining non overlapping 10% of samples is used to test the classifier. A Gaussian kernel is used for multiReGEC classifier with a fixed value for the regularization parameter $\epsilon = 10^{-9}$. The σ value has been obtained with a grid search of the best kernel parameter $\sigma = 10^i, i \in \{1, 2, 3, 4\}$ on a validation set composed of 10% of the training set. Features have been normalized during cross-validation so that each feature in the training set has zero mean and unitary standard deviation (normalization was performed with *zscore()* function of Matlab).

3.1. Significance of classification accuracy

Given the high dimensionality of Raman data, the first test investigates how the classification accuracy is influenced by chance. This problem has been already addressed in the case of microarray data, where the size of the

dataset is usually much smaller than the dimension of the space in which the data are embedded[39,40].

To assess the statistical significance of the classification accuracy rate, the statistical method of hypothesis testing is applied. Let the null hypothesis $H0$ be that the points in the dataset and their class label are instances of independent random variables. To evaluate the p-value corresponding to the classification accuracy acc, we need to evaluate the probability density function of acc under the null hypothesis. Since this is unknown, a method is needed for estimating the empirical probability density function under $H0$, from the available data. A nonparametric permutation test[41] is used, which consists in permuting randomly the labels of the training set, training the classifier on this randomly labelled training set and testing the accuracy of the classifier on a test set having examples with correct labels. The construction of training sets by label permutations is justified by the fact that, under the null hypothesis that data and labels are independent, all pairs of points and labels have the same probability to occur.

The permutation test determines the percentage of classification models, obtained on random labels, having an accuracy larger than acc, when classifying the test data with true labels. The following procedure is carried out: for all q_1 random hold outs in the cross validation, perform q_2 random permutations of the labels of examples belonging to the training set. For each permutation, train the classifier and test its accuracy on the test set formed by the remaining data with their correct labels.

Let $acc_{n_{ij}}$ be the accuracy of the classifier trained on n examples with random labels, in the i-th hold out of the cross validation and in the j-th

random permutation. Then, the empirical probability density function of the error rate under the null hypothesis is:

$$p_n(acc) = \frac{1}{q_1 q_2} \sum_{i=1}^{q_1} \sum_{j=1}^{q_2} \delta(acc < acc_{n_{ij}}),$$

which is composed of a sum of Kronecker δ-functions. The statistical significance (p-value) of the accuracy acc is therefore given by the percentage of random accuracies $acc_{n_{ij}}$ greater than acc. The classification accuracy of the method is $92.22\% \pm 0.095$ (p=0.04).

3.2. Significance of selected features

A second test assesses the significance of classification accuracy results obtained on features selected by the spectral clustering. In this case, the null hypothesis is that the features selected by the spectral method and the classification accuracy obtained using only those features are instances of independent random variables. We therefore followed a similar procedure: choose a set of random features, train the classifier using those randomly selected features and test the classifier on a test set with only selected features. For all q_1 hold outs, we compute the number s of features selected by the spectral clustering. Then, we use s random features to train the classifier and count the number of times the accuracy on random features is larger than on selected features. We repeat the comparison q_2 times for each hold out and we compute the p-value as the fraction of times the random features produce a larger accuracy than selected ones.The accuracy is $97.78\% \pm 0.047$ (p=0.05) using only the features selected by the spectral clustering.

The results of these two tests show that the Raman data can be used to classify the cellular death, and that the frequencies selected by the spectral

clustering significantly discriminate among apoptotic, necrotic and control cells.

3.3. Misalignments in samples

To visualize the effect of misalignments among samples in our dataset, we perform a test on the number of times a feature is selected during cross-validation. As it can be seen in Fig. 4, a cumulative histogram describes the percentage of times the features are selected in the total repetitions of cross-validation. All features are always selected in the same part of the spectrum, corresponding to 1700-1800 cm^{-1}, that represent the $> C = O$ bond from the lipids and the lipid bilayer. During necrotic death, the cell is lysing and the lipid bilayers are destroyed, while during the apoptotic death the lipids just change conformation. In both cases, these changes are significant and it is logical why this contribution becomes the most dominant feature of the layers and therefore the most important changes are in that part of the spectrum.

Figure 4: Count of features selected during cross validation (percentage)

3.4. Robustness of dimensionality reduction

We test the sensitivity of the proposed algorithm to different sizes of the training set. We train the algorithm with a percentage of available data in the range $20\% - 90\%$ and we test on the remaining samples.

The results of these tests are shown in Fig. 5. We apply the dimensionality reduction algorithm obtaining a subset of features. Then, we select a subset of random features of the same size. e compare the accuracy of the

algorithm with all features, with those selected by dimensionality reduction technique and random features. The classification accuracy of the spectral algorithm is always higher with respect to features randomly selected and no dimensionality reduction. We observe that the classification accuracy never drops below 84% even when the training is based on only 20% of the samples available in the dataset.

Figure 5: Classification accuracy versus size of training set. Horizontal axis denotes the percentage of the total dataset that was used for training. The vertical axis is the mean classification accuracy using repeated random sub-sampling validation with 100 repetitions.

The absence of sharp transitions on the solid blue line in Fig. 5 indicates the robustness of the proposed algorithm with respect to small changes of the training dataset size.

3.5. Biological validation

Finally, the developed algorithm has been used to assess the effect of heating on the A549 cells. The new data were obtained at different time points, when the temperature was elevated at 40°C. The original dataset was used to train the algorithm and these new data were used as a test set. Therefore, the new data points were assigned to the class whose representing hyperplane was closest. The algorithm classified all test data as apoptotic, as expected on previous work reported in literature[42,43,44].

It has been established that elevated temperatures alone cause cell death in a predictable manner that is linearly dependent on exposure time and is non-linearly dependent on temperature[42,43]. A variety of cell lines, including

A549, have been reported to undergo apoptosis[44,45] during mild heat treatment and necrosis during prolonged or intensified exposure[46,47]. This result confirms the validity of the methodology and shows the flexibility of Raman spectroscopy in conjunction with the proposed algorithm.

4. Conclusion

In this paper we introduce a data mining approach for the multiclass cancer cell death type discrimination problem based on Raman spectra. This approach firstly reduces the dimensionality of the Raman spectra with a novel technique and then classifies the data with ReGEC, a generalized eigenvalue classifier. This is the first application of multi-classification algorithms and spectral clustering for the analysis of cell death discrimination based on Raman spectroscopy data. The proposed algorithm achieves $97.78\% \pm 0.047$ classification accuracy in cross validation, it provides statistically significative features for discrimination, and the results obtained are biologically valid.

Acknowledgment

Mario R. Guarracino was partially funded by the Italian National Research Council. Vera Tomaino was supported by the Department of Experimental Medicine and Clinic, University Magna Græcia of Catanzaro, Italy. The experimental work was funded by the National Science Foundation (NSF grant EEC-94-02989, NSF-NIRT Grant EEC-0506560 and the National Institutes of Health (NIH Grants 1-P20-RR020654-01, RO1HL75258, R01HL78670). This work was also supported by Airforce grants.

References

[1] P. Hughes, D. Marshall, Y. Reid, H. Parkes, C. Gelber, The costs of using unauthenticated, over-passaged cell lines: how much more data do we need?, BioTechniques 43 (5) (2007) 575.

[2] T. W. Plaia, R. Josephson, Y. Liu, X. Zeng, C. Ording, A. Toumadje, S. N. Brimble, E. S. Sherrer, E. W. Uhl, W. J. Freed, et al., Characterization of a new NIH-registered variant human embryonic stem cell line, BG01V: a tool for human embryonic stem cell research, Stem cells 24 (3) (2006) 531.

[3] G. Pyrgiotakis, T. K. Bhowmick, K. Finton, A. K. Suresh, S. G. Kane, J. R. Bellare, B. M. Moudgil, Cell (A549)-particle (Jasada Bhasma) interactions using Raman spectroscopy, Biopolymers 89 (6) (2008) 555–64.

[4] D. A. Karnofsky, J. H. Burchenal, The clinical evaluation of chemotherapeutic agents, Columbia University Press, 1949.

[5] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, P. P. Carbone, Toxicity and response criteria of the eastern cooperative oncology group, American Journal of Clinical Oncology 5 (6) (1982) 649.

[6] R. Baselt, Disposition of toxic drugs and chemicals in man, Biomedical Publications, Foster City, CA, 2008.

[7] J. Phuchareon, Y. Ohta, J. M. Woo, D. W. Eisele, O. Tetsu, Genetic profiling reveals cross-contamination and misidentification of 6 adenoid

cystic carcinoma cell lines: ACC2, ACC3, ACCM, ACCNS, ACCS and CAC2, PLoS One 4 (6) (2009) e6040.

[8] S. Azari, N. Ahmadi, M. J. Tehrani, F. Shokri, Profiling and authentication of human cell lines using short tandem repeat (STR) loci: Report from the national cell bank of iran, Biologicals 35 (3) (2007) 195–202.

[9] C. A. Owen, J. Selvakumaran, I. Notingher, G. Jell, L. L. Hench, M. M. Stevens, In vitro toxicology evaluation of pharmaceuticals using Raman micro-spectroscopy, J. Cell. Biochem. 99 (1) (2006) 178–186.

[10] E. Carter, H. Edwards, Infrared and Raman spectroscopy of biological materials, Marcel Dekker, New York, NY, 2000.

[11] I. Notingher, Raman spectroscopy cell-based biosensors, Sensors 7 (8) (2007) 1343–1358.

[12] I. Notingher, C. Green, C. Dyer, E. Perkins, N. Hopkins, C. Lindsay, L. L. Hench, Discrimination between ricin and sulphur mustard toxicity in vitro using Raman spectroscopy, Journal of The Royal Society Interface 1 (1) (2004) 79–90.

[13] E. Widjaja, W. Zheng, Z. Huang, Classification of colonic tissues using near-infrared Raman spectroscopy and Support Vector Machines., Int. J. Oncol. 32 (3) (2008) 653–662.

[14] G. Pyrgiotakis, O. E. Kundakcioglu, K. Finton, P. M. Pardalos, K. Powers, B. M. Moudgil, Cell death discrimination with Raman spectroscopy and Support Vector Machines, Annals of Biomedical Engineering 37 (7) (2009) 1464–1473.

[15] M. Guarracino, A. Irpino, R. Verde, Multiclass generalized eigenvalue proximal Support Vector Machines, in: Proceedings of International Conference on Complex, Intelligent and Software Intensive Systems - CISIS, IEEE CS, 2010, pp. 25–32.

[16] M. R. Guarracino, C. Cifarelli, O. Seref, P. M. Pardalos, A classification method based on generalized eigenvalue problems, Optimization Methods and Software 22 (1) (2007) 73–81. doi:http://dx.doi.org/10.1080/10556780600883874.

[17] http://www.atcc.org/ (November 2009).
URL http://www.atcc.org/ATCCAdvancedCatalogSearch/ProductDetails/tabid/452/Default.aspx?ATCCNum=CCL-185&Template=cellBiology

[18] D. J. Giard, S. A. Aaronson, G. J. Todaro, P. Arnstein, J. H. Kersey, H. Dosik, W. P. Parks, In vitro cultivation of human tumors: establishment of cell lines derived from a series of solid tumors, J. Natl. Cancer Inst. 51 (5) (1973) 1417.

[19] N. O. Karpinich, M. Tafani, R. J. Rothman, M. A. Russo, J. L. Farber, The course of etoposide-induced apoptosis from damage to DNA and p53 activation to mitochondrial release of cytochrome c, J. Biol. Chem. 277 (19) (2002) 16547–16552.

[20] D. Boesewetter, J. Collier, A. Kim, M. Riley, Alterations of A549 lung cell gene expression in response to biochemical toxins., Cell Biology and Toxicology 22 (2) (2006) 101–108.

[21] I. Notingher, S. Verrier, S. Haque, J. M. Polak, L. L. Hench, Spectroscopic study of human lung epithelial cells (A549) in culture: living cells versus dead cells, Biopolymers 72 (4) (2003) 230–240.

[22] Y. Huang, A. M.-L. Chan, Y. Liu, X. Wang, N. J. Holbrook, Serum withdrawal and etoposide induce apoptosis in human lung carcinoma cell line A549 via distinct pathways, Apoptosis 2 (2) (1997) 199–206.

[23] G. Yogalingam, A. M. Pendergast, Abl kinases regulate autophagy by promoting the trafficking and function of lysosomal components, Journal of Biological Chemistry 283 (51) (2008) 35941–53.

[24] T. K. Bhowmick, G. Pyrgiotakis, K. Finton, A. K. Suresh, S. Kane, J. Bellare, B. Moudgil, Raman spectroscopy study of the effect of JB particles on Saccharomyces Cerevisiae (yeast) cells by Raman spectroscopy, J. Raman Spectrosc. 39 (12) (2008) 1859 – 1868.

[25] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on pattern analysis and machine intelligence 22 (8) (2000) 888–905.

[26] J. Shi, S. Belongie, T. Leung, J. Malik, Image and video segmentation: The normalized cut framework, in: Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on, Vol. 1, IEEE Comput. Soc., 1998, pp. 943–947.

[27] A. Brun, H. Knutsson, H. J. Park, M. E. Shenton, C. Westin, Clustering fiber traces using normalized cuts, Lecture notes in computer science (2004) 368–375.

[28] M. Kurucz, A. A. Benczúr, K. Csalogány, L. Lukács, Advances in Web Mining and Web Usage Analysis, Vol. Volume 5439/2009 of Lecture Notes in Computer Science, Springer, 2009, Ch. Spectral Clustering in Social Networks, pp. 1–20.

[29] Y. Kluger, R. Basri, J. T. Chang, M. Gerstein, Spectral biclustering of microarray cancer data: coclustering genes and conditions, Genome Research 13 (4) (2003) 703.

[30] O. L. Mangasarian, Linear and nonlinear separation of patterns by linear programming, Operations Research 13 (3) (1964) 444–452.

[31] P. Pardalos, H. Romeijn (Eds.), Handbook of Optimization in Medicine, Springer US, 2009.

[32] V. N. Vapnik, The Nature of Statistical Learning Theory (Information Science and Statistics), Springer-Verlag, Berlin, Germany, 1999.

[33] W. S. Noble, What is a Support Vector Machine?, Nature Biotechnology 24 (12) (2006) 1565–1567. doi:http://dx.doi.org/10.1109/TPAMI.2006.17.

[34] M. R. Guarracino, S. Cuciniello, D. Feminiano, G. Toraldo, P. M. Pardalos, Data mining and mathematical programming, Vol. 45, Centre de Recherches Mathématiques CRM Proceedings & Lecture Notes of the American Mathematical Society, 2008, Ch. Current Classification Algorithms for Biomedical Applications, pp. 109–126.

[35] P. Pardalos, M. Resende (Eds.), Handbook of Applied Optimization, Oxford University Press, New York, 2002.

[36] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, U.K., 2004.

[37] O. L. Mangasarian, E. W. Wild, Multisurface proximal Support Vector Machine classification via generalized eigenvalues, IEEE Trans. Pattern Anal. Mach. Intell. 28 (1) (2006) 69–74. doi:http://dx.doi.org/10.1109/TPAMI.2006.17.

[38] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, Vol. 14, Citeseer, 1995, pp. 1137–1145.

[39] N. Ancona, R. Maglietta, A. Piepoli, A. D'Addabbo, R. Cotugno, M. Savino, S. Liuni, M. Carella, G. Pesole, F. Perri, On the statistical assessment of classifiers using dna microarray data, BMC Bioinformatics 7 (387).

[40] M. Guarracino, S. Cuciniello, P. Pardalos, Classification and characterization of gene expression data with generalized eigenvalues, Journal of Optimization Theory and Applications 141 (3) (2009) 533–545.

[41] P. Good, Permutation tests: a practical guide to resampling methods for testing hypothesis, Springer New York, 1994.

[42] S. A. Sapareto, W. C. Dewey, Thermal dose determination in cancer therapy, Int. J. Radiat. Oncol. Biol. Phys. 10 (6) (1984) 787–800.

[43] M. W. Dewhirst, D. A. Sim, S. Sapareto, W. G. Connor, Importance of minimum tumor temperature in determining early and long-term re-

sponses of spontaneous canine and feline tumors to heat and radiation, Cancer Res. 44 (1) (1984) 43–50.

[44] S. Hayashi, M. Hatashita, H. Matsumoto, Z. H. Jin, H. Shioura, E. Kano, Modification of thermosensitivity by amrubicin or amrubicinol in human lung adenocarcinoma A549 cells and the kinetics of apoptosis and necrosis induction., International Journal of Molecular Medicine 16 (3) (2005) 381–387.

[45] E. P. Armour, D. McEachern, Z. Wang, P. M. Corry, A. Martinez, Sensitivity of human cells to mild hyperthermia., Cancer Research 53 (12) (1993) 2740–2744.

[46] T. Komata, T. Kanzawa, N. Takeo, A. Hiroshi, S. Endo, M. Nameta, T. Hideaki, Y. Tadashi, K. Seiji, T. Ryuichi, Mild heat shock induces autophagic growth arrest, but not apoptosis in U251-MG and U87-MG human malignant glioma cells, Journal of Neuro-Oncology 68 (2) (2004) 101–111.

[47] K. V. Prasad, A. Taiyab, D. Jyothi, U. K. Srinivas, A. S. Sreedhar, Heat shock transcription factors regulate heat induced cell death in a rat histiocytoma, Journal of Biosciences 32 (3) (2007) 585–593.