Recognition of chemical compounds in contaminated water using time-dependent multiple dose cellular responses

T.H. Pan\textsuperscript{a,b}, B. Huang\textsuperscript{b,*}, J.Z. Xing\textsuperscript{c}, W.P. Zhang\textsuperscript{d}, S. Gabos\textsuperscript{d}, J. Chen\textsuperscript{e}

\textsuperscript{a} School of Electrical and Information Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China
\textsuperscript{b} Department of Chemical and Material Engineering, University of Alberta, Edmonton, Alberta T6G 2G6, Canada
\textsuperscript{c} Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta T6G 2Z5, Canada
\textsuperscript{d} Alberta Health and Wellness, Edmonton, Alberta T5J 1S6, Canada
\textsuperscript{e} Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 2Z5, Canada

\textbf{A R T I C L E   I N F O}

\textbf{Article history:}
Received 5 January 2012
Received in revised form 2 March 2012
Accepted 2 March 2012
Available online 8 March 2012

\textbf{Keywords:}
Cytotoxicity
Real-time cell electronic sensor system
Water protection
Support vector machine
Multi-class classification
Cellular state variable identification

\textbf{A B S T R A C T}

An early determination of toxicant compounds of water contaminations can gain critical time to protect citizens’ health and save substantial amounts of medical costs. To determine toxins in real time, a multi-dose classification algorithm using cellular state variable identification (CSVID) is developed in this paper. First, the dynamic cytotoxicity response profiles of living cells are measured using a real-time cell electronic sensing (RT-CES) system. Changes in cell number expressed as cell index (CI) are recorded on-line as time series. Then CSVID, which reflects the cell killing, cell lysis and certain cellular pathological changes, is extracted from those dynamic cellular responses. Finally, a support vector machine (SVM) algorithm based on CSVID is employed to classify chemical compounds and determine their analogous cellular response pathway. In order to increase the classification accuracy, a majority vote of the class labels is also proposed. Several validation studies demonstrate that CSVID-based classification algorithm has great potential in distinguishing the cytotoxicity response of the cells in the presence of toxins.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Early detection of water contamination has gained a lot of attention in the past decades. Water contamination has two major effects on human cells, namely, toxicity effects (cell killing by apoptosis and/or necrosis) and cancer effects (uncontrolled cell proliferation caused by cancer contaminants stimulations) [1]. Therefore, how to build an early warning system is a critical issue for detection of any sudden deterioration in the quality of water supply. Many methods have been proposed to deal with this issue. One approach is to use the analytical chemical method to detect the special chemical compound or a range of compounds having similar properties [2]. Another approach is to use the biological approach to detect the effect of events and give an early bio-alarm [3]. Other approaches include mathematical modelling to assess the water quality and predict the cytotoxicity [1,4].

In practice, there is still a lot of room to build the early warning system in order to detect hazardous events in water supplies more effectively. An efficient early warning system must include the ability of an early determination of the toxins class when the water is contaminated. The timely determination of the toxins class is challenging and yet important for two reasons. First, the cytotoxicity which refers to the effects of toxins on the human cells, has similar characteristics when the living cell is exposed to a low concentration of toxic compounds and it is important to be able to classify them. Second, conduction of biological experiments is costly and time consuming. Thus, our main objective in this paper is to use several pattern and clustering strategies including feature extraction algorithm, so-called CSVID, multi-class classification and a majority vote of the class labels, to determine the class of a chemical compound in water supply in real time.

There is an abundant literature relating to the impacts of different chemical compounds in water contamination using clustering analysis. Hussain et al. [5] proposed the cluster analysis in both Q-mode and R-mode to classify 23 water quality parameters and recognize changes in water quality in different seasons. This algorithm was used to assess the environmental impacts of agricultural practices and also to give the level of water quality using ions and water quality parameters cluster. Yidana [6] also used Q-mode and
R-mode to spatially classify groundwater samples, determine the probable sources of variation in groundwater salinity, and determine the suitability of groundwater for irrigation activities. Preis and Ostfeld [7] presented hybrid model trees – linear programming algorithm – to identify the contamination source in water distribution system. The approach can estimate the time, the location, and the concentration of the contamination injection sources.

Unlike the direct estimation of water contamination, there are significant efforts to identify physicochemical characteristics of pollution source that causes the toxic activity. Liu et al. [8] presented a logistic regression model based on three descriptors to identify and label toxic versus nontoxic events compared with an unexposed control cell population. Sårbu and Pop [9] provided a fuzzy cross-validation algorithm to give the qualitative and quantitative identification of the characteristics (chemical compounds) which revealed the similarities and dissimilarities in the in vitro test systems. This algorithm was utilized to estimate the possible harmful effects of chemicals in vivo test system. Based on a data-driven method, Molnár et al. [10] provided an artificial neural network using atomic fragmental descriptors to recognize cytotoxic and non-cytotoxic compounds according to their evaluation of in vitro human cytotoxicity.

Although the mentioned methods can work well in many cases, they only use the static state of cellular response or endpoint detection. It is well-known that the human cell proliferation/death is a dynamic process when they are exposed to the toxin. Moreover, different dynamic cellular responses are caused by a toxin at different dose levels. The dynamic responses with multi-dose is similar to the human’s fingerprint, which may imply different cytotoxicity mechanisms. We can utilize the family of dynamic response curves to recognize the toxic category. Since time-dependent cellular response data are high dimensional, it is difficult to analyze the problem using traditional classification methods due to the “curse of dimensionality”. The contributions of this paper are that a CSVID method is proposed to mine the cell state of a dynamic cellular response and several approaches are presented to segment the dynamic response. Then the cellular dynamic response can be approximated by two or three linear models. Only parameters of the linear models are selected as features and used for further analysis, which can greatly simplify the classification problem. What’s more, a majority vote of the class labels is cast into this classification framework to further improve clustering performance. As a result, the recognition algorithm based on dynamic responses with multi-dose can improve the classification accuracy and overcome similarity problem of cellular response when they are contaminated by toxins with low concentrations.

To demonstrate the proposed recognition algorithm, five chemical compounds labeled as: 1-Naphthol [Naph (I)], Naphthenic acids [NAS (III)], Pyrene [Py (V)], Methylmercury [Mn (IV)] and Anthracene [Anth (I)] are investigated. These chemical compounds are well-known pollutants which accumulate in groundwater and soil through release from industrial practices and even natural sources [11], and are harmful for human health. To develop a sensitive and reliable real-time classification algorithm, a real-time cell electronic sensing (RT-CES, ACEA Biosciences, CA, USA) system was used in this work, which has been demonstrated to provide sensitive monitoring of cellular responses in a real-time continuous manner [12,13]. We are interested in multiple cell profiles drawn by online monitoring data collected from RT-CES, which allows us to distinguish between different chemical compounds.

The organization of the paper is as follows. The experiment of cytotoxicity assessment is presented in Section 2. Section 3 describes the proposed method including CSVID extracted from dynamic cellular response curves, the procedure of multi-class classification and the majority vote of the class labels. Results and concluding remarks are given in Sections 4 and 5, respectively.

2. Experiment

Cytotoxicity testing is performed using the RT-CES 16 × system (ACEA Biosciences, CA, USA). This system includes: an electronic sensor analyzer, a device station, and a 16-well E-plate. Cells are grown on the surfaces of microelectronic sensors. The sensor devices with cultured cells are mounted on a device station placed inside a CO2 incubator. Electrical cables connect the device station to the sensor analyzer. Sensor analyzer can automatically select wells for measurement and continuously save the measured impedance data. The RT-CES can dynamically monitor cell viabilities because cells grow in the space or well between two electrodes, and any cell changes will affect the electrical conductivity in terms of cell index (CI). For instance, increases of CI indicate that cells are growing. So, the CI provides quantitative information about the biological status of the cells such as the cell number, cell viability (cell fusion, cell apoptosis) and cell morphological change. More details on the RT-CES system can be found in [12–14].

The parameter CI is calculated as [15]

\[
CI = \max_{i=1,...,n} \left[ \frac{R_{cell}(f_i)}{R_{cell}(f_0)} - 1 \right] 
\]

where \( R_{cell}(f_i) \) and \( R_{cell}(f_0) \) are the frequency-dependent electrode impedance (resistance) without or with cells present in the wells, respectively, and \( n \) is the number of the frequency points at which the impedance is measured.

For cytotoxicity assessment, NIH 3T3 cells are selected because these cells can easily be transfected and have already been used in many reports as model cells [16,17]. So, the results obtained from the cell-based in vitro assays can be evaluated objectively. To perform such studies, the NIH 3T3 cell-lines are treated with five potential water contaminations: 1-Naphthol [Naph (II)], Naphthenic acids [NAS (III)], Pyrene [Py (V)], Methylmercury [Mn (IV)] and Anthracene [Anth (I)]. The starting cell number was 10,000 cells per sensor well. When the CI values reached a range between 1.0 and 1.2, the cells were then exposed to one of the toxicant compounds with a certain concentration. Then, the cell responses were continuously monitored and recorded every half an hour by the RT-CES system. Fig. 1 shows seven dynamic cell toxic responses to five toxins of different doses (concentrations), respectively. Taking Fig. 1(a) for example, the NTH 3T3 cells are exposed to different doses of [Anth (I)]: 0.05644, 0.23, 0.9, 3.61, 14.45, 57.79 and 231.16 μM, respectively. In this figure, dynamic cellular responses increase at various rates. Higher dose toxicant yields faster-decaying CI as expected. Compared with Fig. 1(a), Fig. 1(b) shows different pattern of dynamic cellular responses when cells are exposed to various doses of [Naph (II)]. Those dynamic cytotoxic responses reflect the complex cytotoxicity mechanism, such as cell fusion, cell apoptosis, and cell necrosis [18]. The increase, decrease, or transient change of CI profile reflect the proliferation, apoptosis/necrosis, and morphology change in the cells, respectively.

It should be noted that during the experiment each response was repeated minimum in triplicate in order to increase the reliability of biological experiment readings. The median value of each triplicate is calculated for the following feature extraction.

Here, we are not concerned with the specific cytotoxicity mechanism of individual chemicals, but focus on how to distinguish those toxicant compounds by using time-dependent cellular responses to multi-dose toxins.
Fig. 1. Dynamic cytotoxic response of NIH 3T3 cells to different doses of five toxicants: (a) Anth (I), (b) Naph (II), (c) NAs (III), (d) Mm (IV), (e) Py (V). Cells were allowed to grow for 20 h prior to the introduction of the toxins to the culture. CI was recorded every half an hour. Each trace at each concentration was an average of three replicates.

3. Method

As mentioned before, the cells are the living component of the cell-electronic sensors and their responses provide dynamic information on chemical toxicity. A range of doses can be used to develop clustering technique in order to classify toxins, because different doses have different responses to chemicals.
3.1. Cellular state variable identification (CSVID)

Notably, the dynamic CI patterns of the NIH 3T3 cells in response to five toxins at different concentrations are distinct (Fig. 1). Taking Anth (I) for example, the CI has a significant increase at the given dose range (from 0.05644 μM to 231.46 μM) during the first 24 h (Fig. 1(a)) after introducing the toxins. This phenomenon has been explained as cell fusion [13,18]. Unlike treatment with Anth (I), cells treated with Naph (II) show a significant increase in the CI at low concentrations (from 0.06938 μM to 0.28 μM), and a quick declining in the CI at higher concentrations (2.2 μM, 4.4 μM) (Fig. 1(b)) due to quick apoptosis and necrosis mediated by reactive oxygen species with increasing membrane permeability [13].

There are also another phenomena, in which the CI has a significant but transient increase during the first several hours after the treatment, followed by gradually declining (Fig. 1(d), cells treated with Mm (IV) at a concentration 8.8 μM).

The time-dependent cellular response is clearly shown so that the shape of curves can reflect cytotoxicity mechanism of chemicals. In light of three kinds of responses at different time ranges, we can segment those cellular responses by run time range and extract the features within each range, as discussed below.

3.1.1. Significant increase in the CI

As mentioned above, the dynamic cellular response has significant increase after living cells are treated with toxicant at a certain concentration (shown in Fig. 2). Although the cell population increases, the responses have various shapes due to the treatment by different chemicals. That means the increase of CI is not uniform during the whole 24 h. In Fig. 2, the response has two obvious corners (seen points (B) and (C)). So, we can extract different features of the sections segmented by the corners to differentiate the responses.

In Fig. 2(a), the curve can be segmented into three parts separated by the corner points, and the curvature $\kappa[i]$ is selected as segmentation criterion, which is calculated according to Eq. (2).

$$\kappa[i] = \frac{2[CI[i+1] - 2CI[i] + CI[i-1]]*\Delta T[i] - 4[CI[i+1] - 2CI[i] + CI[i-1]]*\Delta CI[i]}{\Delta T[i]^2 + (\Delta CI[i])^2}$$

In Eq. (2), $\Delta T[i] = (T[i+1] - T[i-1])/2$ and $\Delta CI[i] = (CI[i+1] - CI[i-1])/2$. $T[i]$ is time sequence and CI[i] is value of CI corresponding to T[i], where $i$ is the sampling index.

The points with maximum and second maximum value of $\kappa[i]$ are selected as corner points. Then, the time-dependent cellular responses can be approximated by three linear models, and parameters (slope $B_i$ and range $\Delta T_j, j = 1, 2, 3$) of each linear model are selected as CSVIDs. This means that the slope (B$_1$) and time range ($\Delta T_1$) between (A) and (B), the slope (B$_2$) and time range ($\Delta T_2$) between (B) and (C), and the slope (B$_3$) and time range ($\Delta T_3$) between (C) and (D) are calculated and selected as CSVIDs. Taking $B_1$ and $\Delta T_1$ for example,

$$B_1 = \frac{CI[b] - CI[a]}{T[b] - T[a]}$$

$$\Delta T_1 = T[b] - T[a]$$

Then, the feature of this kind of dynamic cellular response can be extracted and formulated as shown in Eq. (4):

$$X = [x_1, x_2, x_3, x_4, x_5, x_6] = [\Delta T_1, B_1, \Delta T_2, B_2, \Delta T_3, B_3]$$

Remark 1. With a linear model to approximate each segmented part of the time-dependent cellular response, only the slope $B_i (i = 1, 2, 3)$ of the linear model is selected as CSVID and the other parameter of the linear model (offset) is discarded. The reason is that the repeatability of biological experiment is not sufficiently good particularly in terms of the offset due to many uncontrolled factors, such as cells initial state of each experiment, manual error, and different operating environment. Different experiments may
have different starting CI, as expected. In contrast, the slope of the cellular response is more robust characteristic for classification.

**Remark 2.** Parameters $B_1$ and $\Delta T_1$ directly reflect the transient increase in CI during the first $\Delta T_1$ after the treatment. The larger $B_1$ and smaller $\Delta T_1$ are, the faster increment of living cells is. The parameters in other sections have the similar interpretation.

**Remark 3.** Using the SCVID approach, the massive raw data of the cellular response can be projected into six parameters. In addition, the high-dimension (various doses, various time points, and various CIs) of cellular response is now reduced to six-dimensional. The six feature parameters can now be used effectively for statistical analysis.

**Remark 4.** The time-dependent cellular responses have different shapes caused by different cytotoxicity mechanism. In the experiment of the five chemical compounds of interest, using three linear models has reasonably captured the dynamics. The proposed segmentation and approximation method can be easily extended to other dynamic responses in presence of other toxins.

**Remark 5.** To obtain a more robust feature of the cellular responses, the smoothing algorithm can be used to smooth the response curve because the real data often contain noises. The moving average filter is defined as:

$$C(i) = \frac{C[i-2] + C[i-1] + C[i] + C[i+1] + C[i+2]}{5}$$

(5)

**Remark 6.** Although the filter algorithm is used to alleviate the influence of various noises, some selected corner points may be unreasonable. Each selected corner point should be double checked visually. If the selected corner point is unreasonable, other remedied strategies should be taken. For example, some abnormal curvatures should be set as 0, then the corner point selected procedure is implemented again.

In the cell proliferation case, the time-dependent cellular response has another type of shapes (as shown in Fig. 3). It can be seen that the curve has no obvious corner, but has an arched curve. So, it can still be approximated by two linear models. Being short of the third part, [0, 0] should be added to complete the six-dimensional feature vector $X$ for this curve to keep the consistent dimension of CSVIDs:

$$X = [x_1, x_2, x_3, x_4, x_5, x_6] = [\Delta T_1, B_1, \Delta T_2, B_2, 0, 0]$$

(6)

3.1.2. A quick decrease in the CI

As shown in Fig. 4, a quick decrease in the CI with strictly dose-dependence can occur at the first several hours of the treatment, due to quick apoptosis and necrosis mediated by reactive oxygen species with increasing membrane permeability [13].

Using the same approximation approach as in the increasing CI case, the curve in Fig. 4(a) can be easily approximated by three linear models and the CSVIDs of each segmented part can be calculated and composed as mentioned before (Eqs. (3) and (4)). It should be noted that there is another shape (see Fig. 4(b)), that is approximated only by two linear models, similar to the increasing CI case shown in Fig. 3.

3.1.3. A transient increase followed by gradual decrease in the CI

As mentioned before, there are some time-dependent cellular responses with a significant but transient increase in the CI during the first several hours, followed by gradual decrease in the CI (shown in Fig. 5). In this case, the point with maximum value of CI can be selected as the first corner point, and the one with maximum curvature to the right (Fig. 5(b)) can be selected as another corner.
point. Then the curve can be approximated as three linear models. If the curve is only approximated by using two linear models (see Fig. 5(a)), [0, 0] is added to complete the feature vector as explained in Eq. (6).

Fig. 6 shows another special shape. It can be seen that the cell population goes through three states: increase in the first several hours, followed by gradually decrease lasting for several hours, and then increase again. This phenomenon reflects the complex cytotoxicity mechanism. However, this cellular response can be easily approximated by three linear models, in which the points with the maximum and minimum values of CI are selected as corner points.

3.2. Least square support vector machine (LS-SVM)

Based on collected CSVIDs, a classified algorithm should be taken to distinguish the toxins. Commonly used computer-based classifiers include nearest-neighbor classifier, k-nearest-neighbor, neural networks, a support vector machine, etc. As shown in Fig. 1, the cellular responses have the similar trend when the cells are exposed to the toxins at the low concentration. That means the overlap between the classes is high. So, the multi-class classification based on support vector machine is used in this work.

Support vector machine (SVM) was originally proposed by Vapnik [21] for binary classification. SVM is gaining popularity due to many attractive features and promising empirical estimation performance. The least squares support vector machine (LS-SVM) is another modified SVM which uses a least squares loss function and equality to substitute inequality constraints [22]. Then, a quadratic programming problem can be simplified as solving a set of linear equations, which significantly reduces the complexity and computational effort. The detail is described as follows.

Given a training data set \( \{ X_i, y_i \}_{i=1}^{n} \) with each input \( X_i \in \mathbb{R}^{d} \) and corresponding binary class labels \( y_i \in \{-1, +1\} \), the LS-SVM classifier is constructed by minimizing:

\[
J(\omega, b) = \frac{1}{2} \omega^T \omega + \frac{1}{y} \sum_{i=1}^{n} e_i^2
\]

s.t. \( y_i [\omega^T \Phi(X_i) + b] = 1 - e_i, \quad i = 1, \ldots, n \)

with \( \Phi : \mathbb{R}^{d} \rightarrow \mathbb{R}^{df} \) a mapping from the input space into a high-dimensional feature space of dimensions \( df \), where weight vector

![Fig. 6. A transient increase, then declining and followed by gradually increase of cells to Py(V) with 400 μM.](image)

![Fig. 7. Classified results for Anth(I), Naph(II), NAs(IV) and Py(V) with a single dose (date: 2011-02-01).](image)
\[ \omega_c \in \mathbb{R}^d \] is in primal weight space, \( \gamma \) is a positive regulariza-
tion constant, \( b_i \) is a bias term, error variables \( e_i \in \mathbb{R} \), and \[ e_i = y_i - (\omega_i^T \Phi(X_i) + b_i). \]

Since the LS-SVM classifier formulation involves the equality
constraints only, the solution is obtained by solving a system of
linear equations:
\[
\begin{bmatrix}
0 \\
1_v
\end{bmatrix}
\begin{bmatrix}
\Omega + \gamma^{-1}I_n \\
\Omega
\end{bmatrix}
\begin{bmatrix}
b \\
\alpha
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
y
\end{bmatrix}
\] (8)

where \( Y = [y_1, y_2, \ldots, y_v]^T \), \( 1_v = [1, 1, \ldots, 1]^T \), \( \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_v] \), \( \Omega \in \mathbb{R}^{n \times n} \) with \( \Omega_{ij} = \Phi(X_i)^T \Phi(X_j) = k(X_i, X_j) \) is a kernel function,
which allows us to work in the feature space without explicitly
constructing it. In practice, popular kernels are the linear kernel
to build classifiers with a linear decision boundary in the input
space, and the nonlinear radial basis function (RBF) kernel. The RBF
kernel has parameter \( \sigma \) to denote the kernel width.

Parameters \( \alpha \) and \( b \) can be obtained using the conjugate gradient
method.

3.3. Multi-class framework

Multi-class classification algorithms aim at assigning a class
label for a single dose. Given a training data set of the form \( \{X_i, y_i\}_{i=1}^n \)
with each CSVIDs \( X_i \in \mathbb{R}^d \) (here \( d = 6 \)) and corresponding \( y_i \in \{1, 2, \ldots, K\} \) with \( K \geq 3 \) (here \( K = 5 \)), we aim at finding a learning model \( H \)
such that \( H(X_i) = y_i \) for new unseen CSVIDs.

The learning model \( H \) is trained by the LS-SVMlab toolbox
kuleuven.ac.be/sista/lssvmlab/

An algorithm using LS-SVM for multi-classification is presented
below:

**Algorithm 1.**

**Step 1:** Divide the entire dataset into two subsets: the training set \( (Z^T) \) and the validation data set \( (Z^V) \);

**Step 2:** Set the class label value \( K_i \) in the training set;

**Step 3:** Determine the optimal values of hyper-parameters from
the training set;

**Step 4:** Select a query CSVIDs from the validation set, and estimate
the label \( f \) using the training classifier.

**Remark 7.** To obtain a good performance of classifier, the proper
kernel function (classifier) and the optimal LS-SVM parameters are
critical in the training phase.
1. Compared with other feasible kernel functions, the radial basis function (RBF) as a non-linear function is a more compacted supported kernel and is able to reduce the computational complexity of the training procedure and gives a good performance under general smoothness assumptions [23]. The RBF function can also handle the non-linear relationships between the spectra and target attributes. Thus, RBF kernel is recommended as the kernel function of LS-SVM in this work.

2. The regularization parameter γ determines the tradeoff between minimizing the training error and minimizing model complexity. The parameter σ² of the RBF kernel function is the bandwidth and implicitly defines the non-linear mapping from input space to some high-dimensional feature space. In this work, we use L-fold cross-validation (here: L = 8) to obtain the optimal parameter values.

3. The Matlab code for training phase is:

   >> model = initlssvm(Xtrain, Ytrain, 'c', [], [], 'RBF', 'kernel');
   >> model = tuneLSSVM(model, 'simplex', 'crossvalidateLSSVM', 8, 'miclass', 'code „OneVsOne”');
   >> model = trainLSSVM(model);

3.4. A majority vote of the class labels

As shown in Fig. 1, small doses of different toxins often produce similar cellular responses. Using the response to a single small dose to determine the category of chemical compounds can obviously lead to a high false classification results. As the final step of the proposed classification method, a majority vote of the class labels is taken to determine the final class assignment.

Given an unlabeled working data set consisting of the chemical compound with multiple doses (m different doses, where m > 1), the feature of the cellular response with each single dose is extracted and used for the developed multi-class classification (Algorithm 1). Then the chemical compound with multiple doses is assigned a membership degree to the suggested class indicating its percentage association with that class. A new chemical compound will be assigned to one of the five classes as detailed in Algorithm 2.

Algorithm 2.

Step 1: Construct a classifier based on Algorithm 1 using the training dataset;
Step 2: Aiming to the unlabeled chemical with m(m > 2) doses, the ith dose response is processed as follows (for i = 1 to m):
   − extract a feature from the ith time-dependent cellular response;
   − use the training classifier to predict and record the class label for the ith response according to the extracted features;
Step 3: Use a majority vote according to the class labels from 1 to m to obtain the final classification decision.

4. Results and discussion

4.1. Classification of time-dependent cellular response

In this section, data of six experiments executed on 2011-01-12, 2011-01-18, 2011-02-01, 2011-02-24, 2011-03-02 and 2011-03-09 are selected to validate the proposed method (data are shown in Table 1). First, we use the proposed CSVID algorithm to determine the features from each time-dependent cellular response among those experiments. Then, cellular responses collected from:
Table 2
Classified result for different chemical compounds with a single dose.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Classified rate for a single dose (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2011-02-01</td>
</tr>
<tr>
<td>Anth (I)</td>
<td>77.78</td>
</tr>
<tr>
<td>Naph (II)</td>
<td>85.71</td>
</tr>
<tr>
<td>NAs (III)</td>
<td>85.71</td>
</tr>
<tr>
<td>Min (IV)</td>
<td>-</td>
</tr>
<tr>
<td>Py (V)</td>
<td>100</td>
</tr>
<tr>
<td>Average</td>
<td>87.65</td>
</tr>
</tbody>
</table>

2011-01-12, 2011-01-18, 2011-02-24 are selected as the training set, and the remaining data are used as the validation dataset. Fig. 7 shows the classification results for four chemicals with a single dose (2011-02-01). It can be seen that the proposed method achieves a reasonable classification result (the worst classification rate is 66.67%). The details are summarized in Table 2.

It appears that the successful classification rate is lower for some chemicals based on the single dose responses. For instance, the classification performance for Anth (I) on 2011-03-02 is only 66.67%. This means that the dynamic CI pattern of the NIH 3T3 cells in response to Anth (I) is similar to some other chemicals. So, we need an additional step to improve the classified ratio. A majority vote of the class labels for these m doses is taken in this study. Taking Anth (I) with seven doses collected on 2011-03-02 for example, four of them are classified into Anth (I), the others are classified into Naph (II), NAs (III) and Py (V), respectively (shown in Fig. 8(a)). Then, the class label of this chemical is determined as Anth (I) through a majority vote of the class labels (4/7). Using the majority vote, the re-classified results for data sampled on 2011-02-01, 2011-03-02 and 2011-03-09 are shown in Fig. 8(b)–(d), respectively. One can observe that the classifier with multi-dose and majority vote has clearly better performance than that of a single dose. The classification rates achieve 100% in this case.

4.2. Discussion

The proposed classification algorithm is based on the concept that the CSVID of the time-dependent cellular response can reflect the cell killing, cell lysis and certain cellular pathological changes, such as cellular morphological and adhesion change, induced by toxic agents. The sign and value of slope $B_i$ ($i = 1, 2, 3$) suggest the increase/decrease of cells, and the proliferated/death rate of cells, respectively.

The proposed strategy also includes a majority vote of the class labels. This is made possible since multiple doses of chemicals are commonly applied to different wells of RT-CES simultaneously during each RT-CES experiment. If a majority does not exist in the clustering process, or the majority is too narrow to be confident, then the chemical is assigned to the ‘Uncertain’ class.

The core of our proposed classification strategy is the CSVID of dynamic cellular responses. The success of this strategy also depends on the quality of experiment. As mentioned in Section 3, the repeatability performance of biological experiment is not sufficiently satisfactory, which can be seen from Fig. 9. Among five experiments, although the trends of dynamic cellular responses are same, there exist considerable differences particularly in the absolute location. So, improving repeatability of the cell experiment will certainly lead to better classification.

The ultimate objective of this study is to develop early detection systems to warn citizens as early as possible when the water is contaminated by toxins. Timely determination of the category of chemicals is a central element of a early warning system which will integrate other related devices, several software elements and other prediction features presented in our previous work [1,4,13,15,18].

5. Conclusion

In this paper, we have considered classification of several different toxins induced by water contaminants. The CI data from a real-time cell electronic sensing (RT-CES) system was used for feature extraction. Least square support vector machine was applied to develop a classifier for the extracted features. A majority vote of the class labels was used to obtain the final classification decision. The developed classifier was verified using data that had not participated in the training. We also examined the reproducibility of CSVIDs-based classifier on several experiments. It was concluded that the proposed classifier had great potential to differentiate toxins.

Acknowledgements

The authors thank Mrs. Min Huang (Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, AB T6G 252, Canada) for conducting all experiments, Dr. Iraj for helpful discussions to improve the presentation quality of the paper, and gratefully acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Alberta Health and Wellness, the Jiangsu Government Scholarship for Overseas Studies and the National Nature Science Foundation of China under the grant 60904053.

References

null