Zhang *et al. BioData Mining* (2016) 9:19
DOI 10.1186/s13040-016-0098-0

BioData Mining

RESEARCH Open Access

Machine learning algorithms for mode-of-action classification in toxicity assessment

Yile Zhang[1], Yau Shu Wong[1], Jian Deng[1], Cristina Anton[2], Stephan Gabos[3], Weiping Zhang[4], Dorothy Yu Huang[5]* and Can Jin[6]

*Correspondence:
yhuang@ucalgary.ca
[5] Alberta Centre for Toxicology, University of Calgary, T2N 4N1, Calgary, Canada
Full list of author information is available at the end of the article

Abstract

Background: Real Time Cell Analysis (RTCA) technology is used to monitor cellular changes continuously over the entire exposure period. Combining with different testing concentrations, the profiles have potential in probing the mode of action (MOA) of the testing substances.

Results: In this paper, we present machine learning approaches for MOA assessment. Computational tools based on artificial neural network (ANN) and support vector machine (SVM) are developed to analyze the time-concentration response curves (TCRCs) of human cell lines responding to tested chemicals. The techniques are capable of learning data from given TCRCs with known MOA information and then making MOA classification for the unknown toxicity. A novel data processing step based on wavelet transform is introduced to extract important features from the original TCRC data. From the dose response curves, time interval leading to higher classification success rate can be selected as input to enhance the performance of the machine learning algorithm. This is particularly helpful when handling cases with limited and imbalanced data. The validation of the proposed method is demonstrated by the supervised learning algorithm applied to the exposure data of HepG2 cell line to 63 chemicals with 11 concentrations in each test case. Classification success rate in the range of 85 to 95 % are obtained using SVM for MOA classification with two clusters to cases up to four clusters.

Conclusions: Wavelet transform is capable of capturing important features of TCRCs for MOA classification. The proposed SVM scheme incorporated with wavelet transform has a great potential for large scale MOA classification and high-through output chemical screening.

Keywords: Time-concentrations response curve, Mode of action, Wavelet transform, Dose response curve, Machine learning, Support vector machine, Artificial neural network

Background

In recent years, considerable progress has been reported in the study of toxicity profiling using in vitro assays [1]. It is important to develop fast and effective methods capable of analyzing large amount of in vitro data set [2, 3]. By comparing the response profiles of chemicals with known mode of actions (MOAs), we are able to infer the MOA of tested chemicals [4, 5]. One such in vitro assay utilizes the real-time cell analysis system (RTCA)

[6–8]. The RTCA system integrates the micro-electrode on the bottom of the wells, such that the electronic impedance data reflect adherent cells status including cell number, cell morphology and adhesion strength. The impedance data at different time points are measured and converted to the cell index (CI) data for further analysis [9, 10]. The system allows multi-concentration assays, such that the Time Concentration Response Curves (TCRCs) can be generated and recorded. The TCRC profiles provide useful information to study the cell-chemical interaction mechanism.

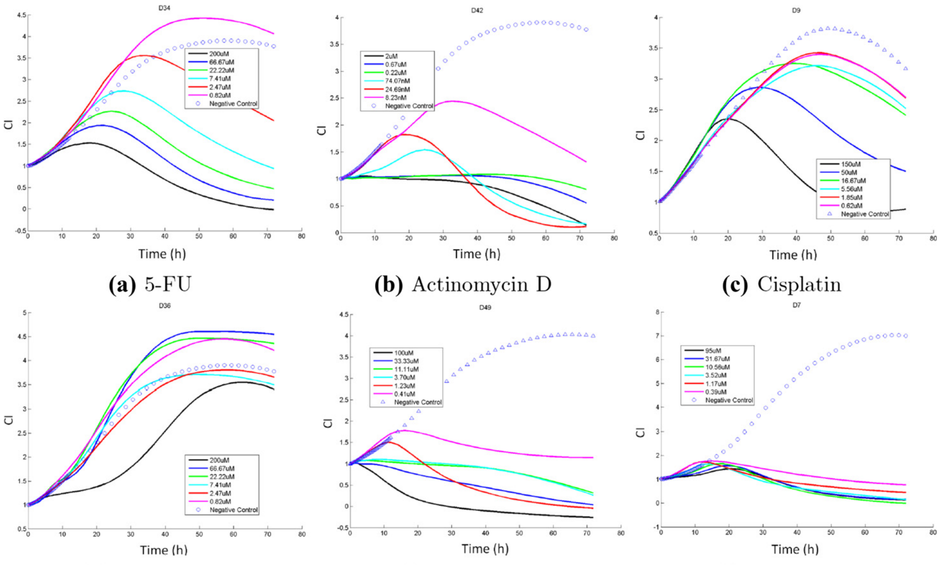
A few analysis methods have been developed to extract useful information from the TCRCs. For example, LC_{50} reflects the chemical concentration that leading to killing 50 % of tested cells [11], KC_{50} uses an exponential model to calculate the LC_{50} value [11–13], AUC_{50} represents the area under the normalized TCRCs, which can be employed to evaluate the toxicity [14]. Based on these indices, further classification or pattern recognition can be investigated. However, these indices only provide partial information of TCRCs and some significant features may not be uncovered. All these indices have the primary goal of detecting toxicity potency of the testing chemicals. The application into MOA classification is indirect, and not tested [15, 16].

MOA describes a cellular level functional change, which is a result of exposure of a living organism to a chemical. According to a pre-set criterion, the chemicals can be classified into different MOA clusters [17]. The machine learning approach has already been utilized in life science research including toxicity classifications [18, 19], analyzing high throughput screening data [20], and drug design [21]. Cheng et al. [22] investigated the toxicity pattern recognition for diverse industrial chemicals with substructure. Vanneschi et al. [23] compares different machine learning algorithms in classifying patients by using breast cancer dataset. Recently, Beck et al. [24] investigate the machine learning by random forests and logistic regression classifiers in bacterial vaginosis (BV) classification, Lareau et al. [25] apply machine learning to analyze functional effectors in microarray data, Lu et al. [26] compares four supervised learning methods in modeling the differentiation of CD4+ T cell.

In this study, we focus on MOA classification for the 63 chemical compounds screening data provided by the Alberta Centre for Toxicology. The list of the chemicals and their ten-cluster MOA classification are given in Additional file 1. The same chemicals were investigated by Pan et al. [13, 14] and Xi et al. [27] for toxicity assessment. Instead of using end-point results, the goal of this study is to develop a new machine learning methodology utilizing the entire TCRCs data recorded for the 63 chemical compounds to perform MOA clustering analysis. The results were validated with the known MOA classification.

It should be noted that it is not trivial to uncover the MOA correlation from the TCRC profiles. In Fig. 1, we display the TCRC profiles of six compounds in cluster 1: DNA/RNA-Nucleic Acid Targets, and the corresponding TCRCs are quite different. However, compounds with different clusters may have similar profiles as illustrated in Fig. 2 showing TCRCs from four different clusters resemble with each other. When the concentration is small, the TCRC profile is very close to the negative control curve. Therefore, to present a better illustration, only the six highest concentrations and the negative control are plotted in Figs. 1 and 2.

The main contributions presented in this work are twofold. First, a novel computational tool is developed based on machine learning for toxicity assessment which was validated for the effectiveness using the TCRCs of 63 chemicals with known MOAs as input. The

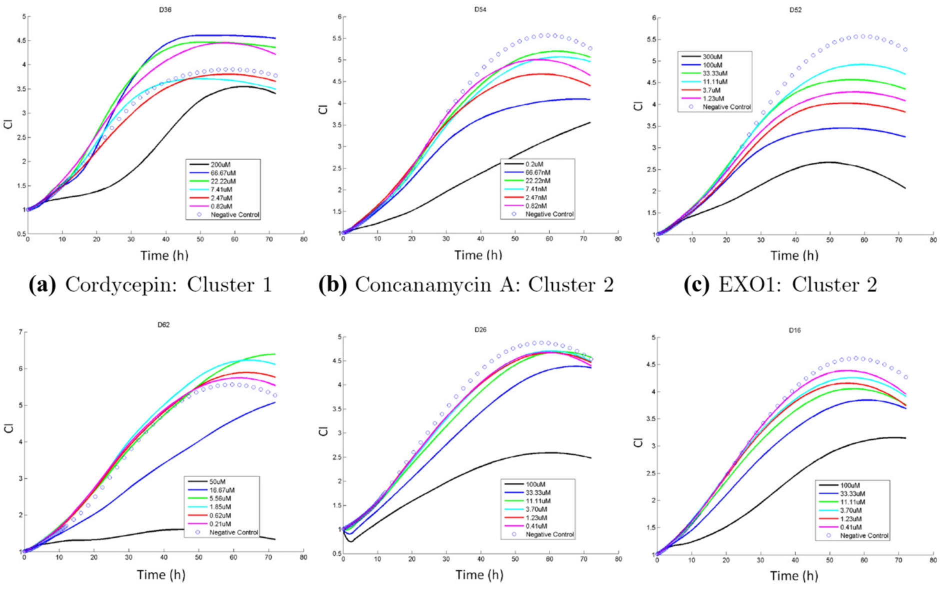


(a) 5-FU **(b)** Actinomycin D **(c)** Cisplatin

(c) Cordycepin **(e)** Doxorubicin **(f)** Topotecan

Fig. 1 TCRCs from Cluster 1: DNA/RNA-Nucleic acid target. The detail of chemicals in (**a**)-(**f**) are provided in the Additional file 1

machine learning methods are based on the artificial neural network (ANN) and support vector machine (SVM) with supervised learning algorithm. Second, wavelet transform is implemented to resolve the difficulty due to taking a large data set from the entire TCRCs. Therefore, instead of directly using the TCRCs as input data to the machine learning algorithms, the wavelet coefficients are selected as input. The application of wavelet preprocessing step not only significantly reduces in the input data, but extracts useful



(a) Cordycepin: Cluster 1 **(b)** Concanamycin A: Cluster 2 **(c)** EXO1: Cluster 2

(d) FK506: Cluster 8 **(e)** Blebbistatin: Cluster 10 **(f)** Monastrol: Cluster 10

Fig. 2 TCRCs from four different clusters. The detail of chemicals in (**a**)-(**f**) are provided in the Additional file 1

information and features of the original TCRCs. Consequently, success rate in clustering analysis is improved.

The remainder of this paper is organized as follows. The materials and data preprocessing of the present study is first previewed. The next section focuses on the methods based on machine learning approach using ANN and SVM, and the application of wavelet transform is discussed. To validate the developed computational tools, we present binary and multi-cluster classifications applied to the 63 compounds in the fourth section. The effectiveness of SVM is demonstrated by the excellent agreement resulted from the known clustering based on MOA applied to the tested chemical compounds. The use of DRCs is proposed, and the advantage of utilizing DRCs to enhance the performance of the machine learning algorithm for limited data set is reported. Finally, conclusion remark is presented.

Materials and data preprocessing

Cell line

Human hepato carcinoma cells line-HepG2 (ATCC, cat. no. HB-8065) were grown and tested in EMEM basal media supplemented with 10 % fetal bovine serum. All growth and assay were conducted in 37 °C tissue culture hood with 95 % humidity and 5 % CO2.

Chemicals

All testing chemicals were at least 95 % purity. They were obtained through commercial sources including Sigma-Aldrich, Cayman Chemicals, Tocris, and Santa cruz biotechnologies. Three solvents were used for powder solubilization: water, DMSO or ethanol. The solvent providing highest solubility when diluted in assay media were used for stock solution preparation. Stock solution were aliquoted for single usage and stored at −20 °C. The highest testing concentration is at most 1/500th of the stock concentration, so that solvent (DMSO or ethanol) concentration are no more than 0.2 %. Each chemical were tested with 11 concentrations, with 1:3 serial dilution.

RTCA HT assay

The xCELLigence RTCA HT system developed at ACEA Biosciences Inc. runs four 384x well E-Plates on four independent HT Stations. The continuous cell monitoring enabled both transient and long term effects being recorded. The system was integrated with the Biomek FXp System and the Cytomat hotels for fully automated liquid handling and plate shuffles. The HepG2 cells were seeded into the E-plate 384, and monitored once an hour in the first 24 h for initial attachment and growth. 11 concentrations of each chemical were applied into the wells by using automatic pipetting. The cellular responses were continuously monitored for at least 72 h.

Data preprocessing

The RTCA technology monitors the impedance signal generated by cells covering electrodes. The impedance signal R is converted to a parameter Cell Index (CI) with the following formulation [28, 29]:

$$CI = \max_{k=1,\cdots,K} \left[\frac{R_{cell}(f_k)}{R_b(f_k)} - 1 \right], \tag{1}$$

where $R_{cell}\left(f_k\right)$ and $R_b\left(f_k\right)$ are the electrode impedance with and without cell in the well, and k is the discrete time points.

To focus on cellular response to testing chemicals, *CI* differences from seeding and growth variation were minimized by using Normalized Cell Index (NCI), which is given by

$$NCI[\,k] = \frac{CI[\,k]}{CI[\,0]},\ k = 1, 2, \cdots, K. \tag{2}$$

Here, k refers to different time points after testing chemical addition, and k=0 refers to the time point right before treatment.

Because not much information can be extracted from the TCRCs before adding the compounds, we focus on the NCI data after chemical treatment. Moreover, for the irregular data set, the time grids for different compounds are not uniform. We apply a cubic spline to interpolate the non-uniform data into uniform grids, where the time interval is one hour for the interpolated data set. The uniform data set enables the use of wavelet transform, which is critical in data reduction and better extracting the features from the original TCRCs data set.

Methods

Two machine learning algorithms, namely artificial neural network (ANN) and support vector machine (SVM) were used in this study. The application of wavelet transform to enhance the performance and effectiveness of ANN and SVM will also be introduced.

Artificial neural network

Artificial neural network (ANN) is inspired by a biological neural network, and it can be considered as a computational information processing model simulating a "brain like" system of interconnected processing units. ANN has already been applied in toxicity study [30–32]. A typical feedforward multi-layer ANN [33] is shown in Fig. 3.

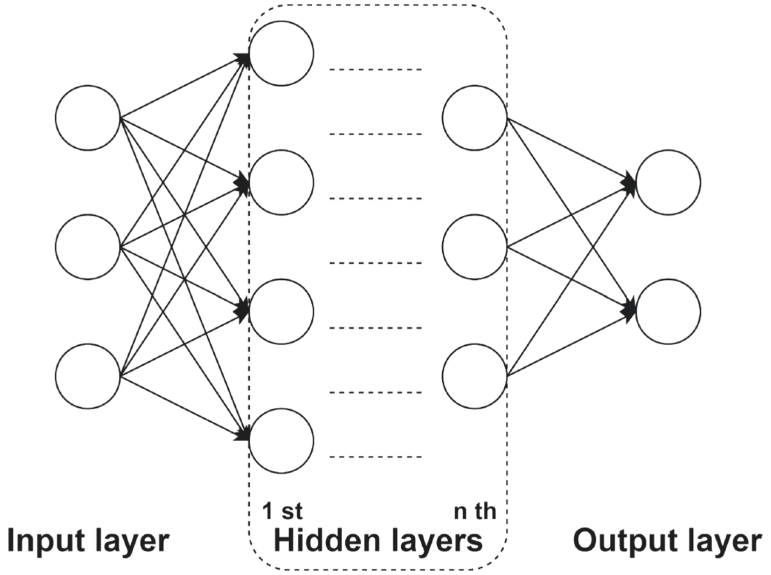


Fig. 3 Feedforward *n*-layer ANN

In the network, there are one input layer and one output layer. The number of input neurons equals to the number of attributes, and the number of output neurons depends on the particular application of the network. In the present study, input neurons are given by the time series of TCRCs, and the output neurons are determined by the number of clusters being classified. The layers between input and xoutput layers are the hidden layers. The network architecture (i.e., the size of hidden layers and the number of neurons in each hidden layer) is depended upon the complexity of a specified problem under investigation. Each neuron is interconnected with other neurone in the next layer, and the information passing though the neurons are determined by the weights. The weights are computed by a supervised algorithm by presenting the TCRCs data as input with known MOA clustering as desired output in the training phase. The weights will then be adjusted by minimizing a given objective function. The training process is conducted repeatedly until the network achieved a prescribed success rate for all training data and it will then be used to classify the future compounds with unknown MOA. Once the training is completed, the network is capable of performing a specified task rapidly with little computing time and it is particularly suitable for a real time application.

Mathematically speaking, training the ANN is to seek a function $f: X \to Y$ to fit a set of example pairs (x, y), x∈ X, y∈ Y. The network as a whole can be regarded as a multivariate function or multivariate vector function if there is multiple outputs. By minimizing $f(X) - Y$, we are able to find a function f to approximate the relationship between the attribute of sample set X and the corresponding cluster Y. By inputting the attribute of future sample $\hat{x}$ to the obtained function f, its classification information $\hat{y}$ can be inferred. Obviously, the information in the minimization process is unknown; the training process of ANN is actually a black box model. However, since there exists many local minimum in minimizing $f(X) - Y$, the same training set (X, Y) may produce totally different network parameters and lead to inconsistent classification results. This is particularly true for the MOA classification, where the data set is relatively small and imbalanced.

Support vector machine

In addition to ANN, another important machine learning algorithm is the support vector machine (SVM). The application of SVM in toxic predictions has been reported in [34, 35]. As one of the popular classifiers, the idea of SVM is quite different from that of ANN. To perform a classification for a given data set, SVM uses a hyperplane to separate the sample data points [36]. Assuming there is a set of data x_i along with their corresponding label y_i, and considering the data is composed of two clusters denoted by -1 and 1, then we have the data space

$$D = \{(x_i, y_i) | x_i \in R^P, y_i \in \{-1, 1\}\}_{i=1}^{n}.$$

Initially, we hope to find a hyperplane separating the sample data, in which each class of data belongs to one side. Let the plane be

$$w \cdot x - b = 0.$$

The problem of constructing such a hyperplane is to ensure its robustness. Supposing that there are two samples very closed to each other but on the different sides of the hyperplane, then it is not reasonable to classify them into different categories. To resolve the problem, we select two hyperplanes such that they separate the data with no point between them. The best robustness is achieved when the distance between them is maximized. The region bounded by the planes is called "margin", and the two hyperplanes can be rewritten as

$$w \cdot x - b = \pm 1,$$

therefore, the distance between them is defined by $\frac{2}{||w||}$. It is clear that to maximize the distance, we need to minimize the $||w||$. Consider the fact that if the sample x_i belongs to the first class, then $w \cdot x - b > 1$. Similarly, $w \cdot x - b < -1$ if it is in the second class. Thus, we can rewrite the classification problem as the following optimization problem

$$\min ||w|| \text{ subject to } y_i(w \cdot x_i - b) \leq 1 \ \text{ for } i = 1, \cdots, n.$$

The weight vector w and the parameter b are determined by a supervised learning algorithm similar to ANN. Now, the remaining problem is that for a large amount of data in the data space and due to the highly non-linearity in the sample data, it is not possible to divide them into multiple clusters by hyperplanes. This problem can be resolved by considering a mapping from a lower dimensional space to a high dimensional space using a suitable kernel, so that the data are expected to be separable in the high dimensional space. The selection of the kernel is critical to the success of SVM.

Recent studies indicates that the SVM is more accurate and robust than ANN in the chemical classification [37], and it is capable of handling data set with more complex structure. The SVM algorithm used in this study is based on the standard SVM classifier in MATLAB with a Gaussian kernel. Comparing with ANN, the most significant advantage of SVM is that it has global minima instead of local minima, so that the convergence speed is significantly faster than ANN. Therefore, in the multi-cluster classification, SVM is used as a main tool. Note that the classification of SVM is always binary, but the binary classification algorithm can be recursively applied for applications to multiple clusters. The details will be discussed in the next section.

Wavelet transform

The training process is a crucial component to ensure the success of a learning machine. To certain extent, large input data in the training will affect the structure of learning machine and also introduce more difficulty in the supervised learning. In the present study, the input data contains the time series of TCRCs, and it could have more than 850 points. For ANN, the size of the hidden layers and the number of neurons depends on the number of input neurons. Therefore, taking a large data set of input is not a trivial task for a learning machine, and this may be the reason why no reference has been reported on using ANN or SVM for toxicity assessment using TCRCs as input. We now propose a novel idea to deal with large input data by using wavelet transform. Different from the standard Fourier transform, which is only localized in frequency, wavelets are localized in both time and frequency. Wavelet transform has been successfully demonstrated to be a powerful tool for data compression and feature extraction in signal and image processing.

Let $\{e_i\}$ be an orthonormal and complete set in a Hilbert space H, and T be an arbitrary vector in H [38], then

$$T = \sum_i < T, e_i > e_i,$$

here T is the vector consisting of the data from TCRCs, e_i is the orthonormal basis, $<,>$ is the inner product and $< T, e_i >$ denotes the coefficients under the basis e_i. By selecting a set of orthonormal vectors e_i, we can use wavelet coefficients to represent the TCRCs toxicity data. An orthonormal basis $\psi_{s,\tau}(t)$ [39] having scale parameter s and translation parameter τ can be expressed in the following form:

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right).$$

Let $T(t)$ be the original TCRCs data, then the wavelet coefficients $X =< T, e_i >$ is a function of s and τ given by

$$X(s,\tau) = \int T(t)\psi^*_{s,\tau}(t)dt$$

where * denotes the complex conjugation, this equation shows how a $T(t)$ is decomposed into a set of wavelet basis function $\psi_{s,\tau}(t)$. Accordingly, $T(t)$ can be recovered by the inverse wavelet transform as

$$T(t) = \int\int X(s,\tau)\psi^*_{s,\tau}(t)dsd\tau,$$

where the wavelets are generated from one mother wavelet $\psi(t)$ by scaling and translation.

One of the advantages of wavelet transform lies in its ability to extract multiscale information from the input data. By recursively applying wavelet transforms, it leads to multi-level wavelet decomposition. The procedure for a three-level wavelet decomposition is illustrated in Fig. 4, where the raw TCRCs are represented by T. In the first level of wavelet transform, the original signal T is decomposed into two vectors CA_1 and CD_1 representing the approximate and detail coefficients, respectively. In the second level of decomposition, the wavelet transform is applied again to CA_1 resulting two decomposition CA_2 and CD_2. In a n-level wavelet decomposition, the wavelet

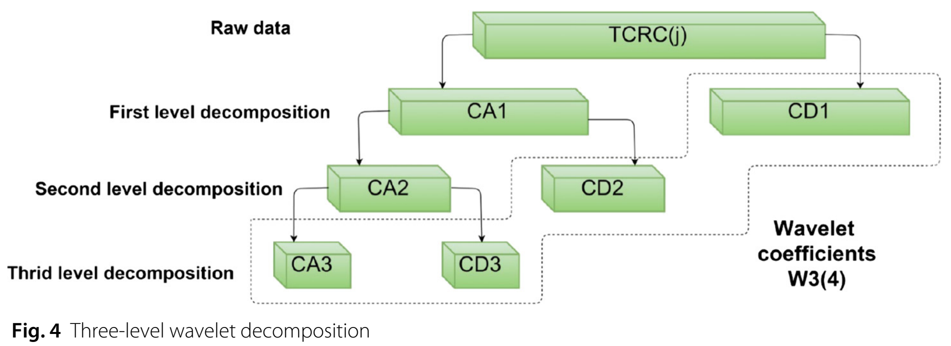
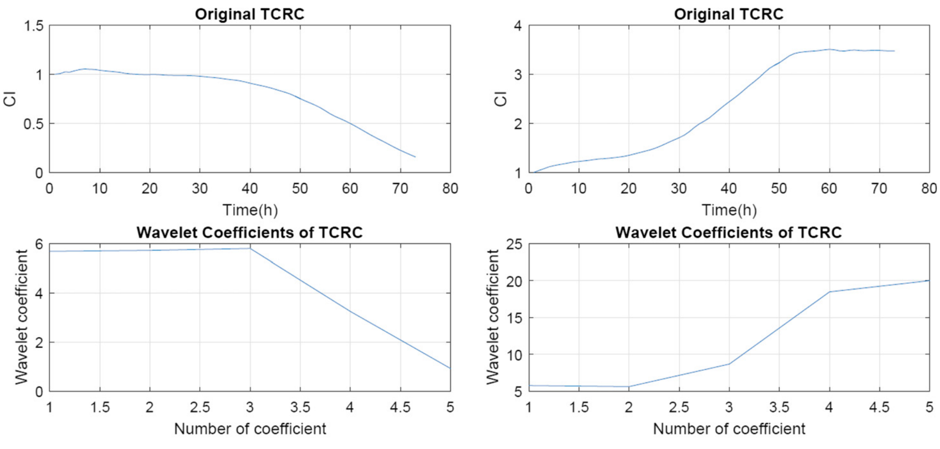

Raw data
TCRC(j)
First level decomposition
CA1
CD1
Second level decomposition
CA2
CD2
Thrid level decomposition
CA3
CD3
Wavelet coefficients W3(4)

Fig. 4 Three-level wavelet decomposition

transform is applied recursively to decompose the approximation coefficient CA_j at the jth level into the coefficients CA_{j+1} and CD_{j+1}. Therefore, applying an nth level wavelet decomposition, we have one approximation coefficient CA_n and detail coefficients CD_n, $\mathrm{CD}_{n-1}, \cdots, \mathrm{CD}_2, \mathrm{CD}_1$. We now denote all wavelet coefficients at the nth level decomposition as W_n. When particular coefficients are used instead of the entire wavelet coefficients, we denote the coefficients as $\mathrm{W}_n(m)$ where m is the number of coefficients. Generally speaking, the selection of wavelet coefficients starts from the approximation coefficient and highest level of detail coefficients, because the detail coefficients at lower level always contain small fluctuations including noise from the original information [40]. Consider a three-level decomposition (i.e., $n = 3$), $\mathrm{W}_3(4)$ means that four wavelet coefficients: $\mathrm{CA}_3+\mathrm{CD}_3+\mathrm{CD}_2+\mathrm{CD}_1$ are kept and $\mathrm{W}_3(2)$ implies taking two wavelet coefficients $\mathrm{CA}_3+\mathrm{CD}_3$.

To demonstrate the capability of extracting important feature of the original data using fewer wavelet coefficients, we apply wavelet transform to two compounds listed in cluster 1. Figure 5 displays the profiles of one concentration TCRC for two different compounds and the corresponding profiles using wavelets $\mathrm{W}_5(1)$. It is clear that the profiles are in good agreement, but a tremendous data reduction over 90 % is achieved using wavelet transform. Note that the original TCRC contains 72 data, while only five wavelet coefficients are in $\mathrm{W}_5(1)$.

We now illustrate how to construct input data for machine learning. A given set of TCRCs is arranged as shown in Fig. 6, where 1 denotes the TCRC with the highest concentration, 2 for the next highest concentration, and 11 for the lowest concentration. By concatenating the vectors according to the order 1, 2, $\cdots$, n, we form a new vector TCRC(n). Here, TCRC(1) contains data from the highest concentration, TCRC(2) contains the first two highest concentrations and TCRC(11) contains data from all 11 concentrations. It will be demonstrated later that including the negative control will enhance the performance of the developed machine learning tools. The new vector TCRC(n) can now be considered as input data to the machine learning algorithm. However, we also



(a) Actinomycin D **(b)** Cordycepin

Fig. 5 Plots of original TCRCs and wavelet coefficients $\mathrm{W}_5(1)$ for (**a**) Actinomycin D and (**b**) Cordycepin

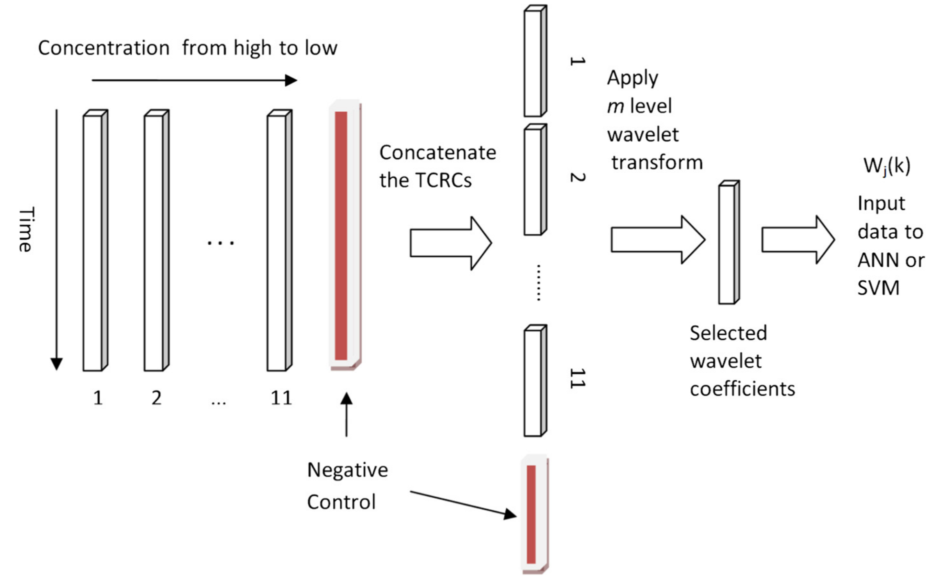


Fig. 6 Input data to machine learning using wavelet transform

consider using wavelets by applying wavelet transform to TCRC(n) and selecting specified multi-level wavelet coefficients as input to ANN or SVM. The advantages of using wavelets will be clearly demonstrated in the next section.

Results and discussion

To validate the developed machine learning tools based on ANN and SVM for MOA classification and to verify the effectiveness of using wavelets for input data preprocessing, we present the following computational simulation applied to the 63 compounds. As shown in Appendix, there are 10 clusters in the 63 compounds with imbalanced cluster distribution as illustrated in Fig. 7. Note that C1 and C10 contain 33 compounds, and they make up more than half of the 63 compounds. Here, we will not consider the three clusters C5, C7 and C9, since each cluster comprises only 3, 2 and 1 compounds, respectively.

For the ANN, a feedforward three-layer network with $24-12-6$ neurons in the hidden layers is used. The results are not sensitive even by doubling the hidden-layer neurons. In the training process, the network is accepted when the success rate of the targeted classification reaches 85 %. For problems with limited and imbalanced data, setting a higher success rate for training may lead to over-fitting and producing an inferior network performance.

Binary classification

We first consider the classification for the two largest clusters, namely C1 with target class DNA/RNA and C10 with target class protein. There are 20 compounds in C1 and 13 compounds in C10, therefore, using 70 % training data implies that 14 compounds in C1 and 9 compounds in C10 are available as training set. The remaining 30 % data, 6 compounds in C1 and 4 compounds in C10 will be considered as test set. All simulation reported in this work are based on 70 % training data and 30 % for the testing data.

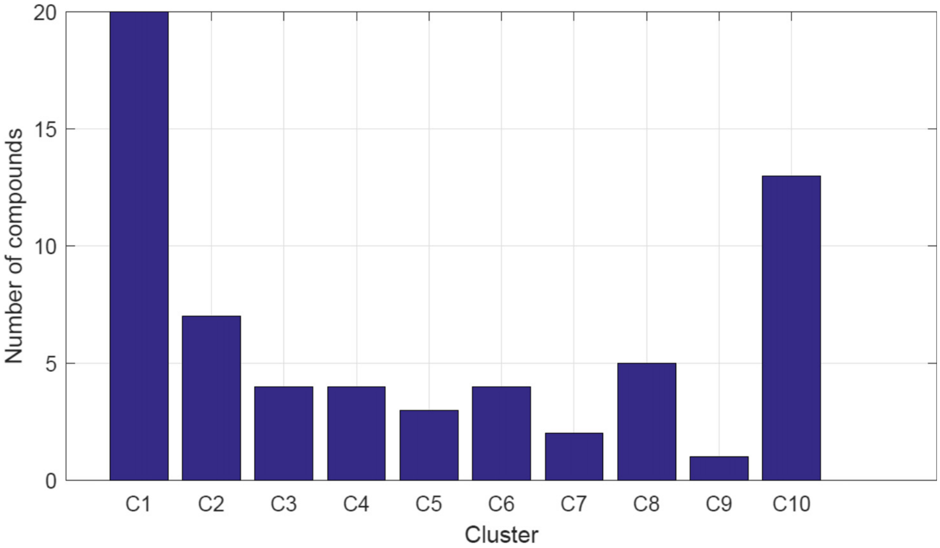


Fig. 7 Distribution of 63 compounds

We define the success rate (SR) for the classification as

$$SR = \frac{\text{Number of compounds classified into correct MOA}}{\text{Total number of compounds in datasets}}.$$

Once the number of compounds in the training set is determined, the developed machine learning tools can be used to perform the classification for C1 and C10. The effectiveness of ANN and SVM can then be evaluated by the computed success rate (SR). For example, in the case of 70 % training set, there are 10 compounds available for the test data. If 9 of them are classified into the correct clusters in C1 or C10, then the successful rate is 90 %. However, it is not reasonable to conclude about the performance of the classifier merely based on one result, especially because the current problem has limited test data for some clusters. To obtain a reliable conclusion for the machine learning tools, the classification process is conducted 100 times, and the training and test set are randomly selected for each simulation. Consequently, 100 SR will be computed from the 100 classifications using 100 different partitions of training and test set. The overall average of the 100 SR will be recorded as the final success rate. Different from the conventional cross validation, which is based on a fixed partition of the data set, the data set partition in the present study is in a more random fashion. This is due to the limited size of the data available in this study, so that a fixed partition can cause significant bias in the classification SR.

As mentioned before, the performance of the machine learning algorithms will be affected by the input data. Intuitively, one may expect that feeding more information to the input should improve the performance for the machine learning tools. In this study, the input is given by the TCRCs and a typical data set consists of 11 concentrations. Let TCRC(1) denote the data taking only the highest concentration, TCRC(2) for data with the first two highest concentrations and TCRC(11) for data including all 11 concentrations. In Tables 1 and 2, we report the SR for ANN and SVM using 70 % of the

Table 1 ANN SR with different concentrations

	Raw	W_1	W_2	W_3	W_4	W_5
TCRC(1)	0.550	0.705	0.738	0.731	0.742	0.701
TCRC(2)	0.711	0.782	0.760	0.774	0.750	0.795
TCRC(3)	0.741	0.782	0.779	0.774	0.798	0.787
TCRC(4)	0.739	0.788	0.811	0.796	0.817	0.820
TCRC(5)	0.750	0.803	0.802	0.829	0.822	0.811
TCRC(6)	0.767	0.819	0.836	0.817	0.826	0.827
TCRC(7)	0.770	0.831	0.825	0.850	0.843	0.817
TCRC(8)	0.838	0.856	0.836	0.861	0.836	0.832
TCRC(9)	0.864	0.852	0.845	0.873	0.829	0.827
TCRC(10)	0.859	0.871	0.849	0.830	0.834	0.838
TCRC(11)	0.855	0.879	0.865	0.863	0.861	0.855

observations as training data. TCRC(*j*) with $j = 1, 2, \cdots, 11$ denotes input using the raw data, and W_i for $i = 1, 2, .., 5$ indicates the corresponding wavelet coefficients from the *ith*-level wavelet decomposition is taken as input data. Using the raw data TCRC(*j*), the SR is poor and unacceptable when $j = 1$. As expected, the SR for ANN is improving when the value of j is increased. However, it is observed that the SR for SVM with TCRC(*j*), $j = 3, 4, \cdots, 7$ is even lower than the SR using the highest concentrations data TCRC(1)and the first two highest concentrations TCRC(2). The advantage of using wavelet coefficients W_i instead of the TCRC(*j*) raw data is clearly demonstrated from the results presented in in Tables 1 and 2. By first applying the wavelet transform to TCRC(*j*) data, consistent improvement in the SR results for both ANN and SVM is achieved as more data are taken as input. Using only the highest concentration TCRC(1), the ANN SR is improved by 35 % when the input data is using wavelet coefficients instead of the raw data. In addition to confirming that wavelet coefficients capture all features in the raw TCRCs data and yield better SR for ANN and SVM, another important enhancement can be achieved by selecting appropriate wavelet coefficients such that much less input data is needed for the machine learning tools. The details and the discussion will be presented shortly.

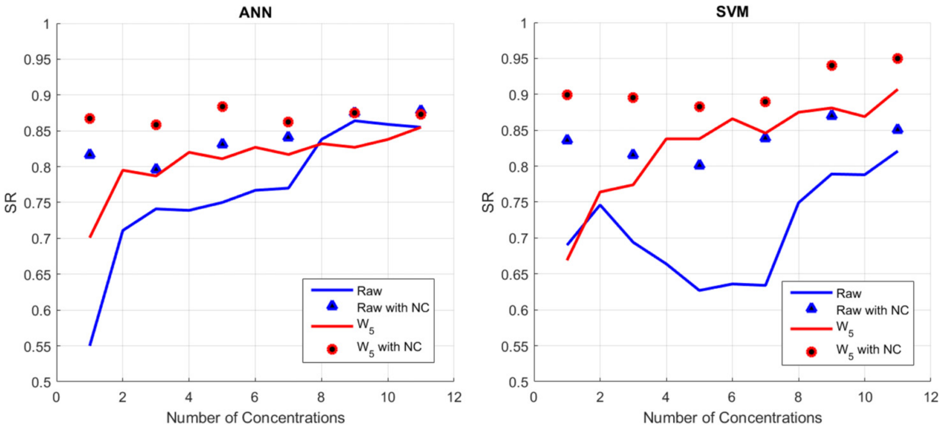
The computational results presented so far are based on input data taken from TCRC(*j*). However, the performance can be further enhanced by taking account information from

Table 2 SVM SR with different concentrations

	Raw	W_1	W_2	W_3	W_4	W_5
TCRC(1)	0.690	0.667	0.698	0.688	0.705	0.669
TCRC(2)	0.746	0.766	0.742	0.744	0.727	0.764
TCRC(3)	0.694	0.802	0.789	0.785	0.787	0.774
TCRC(4)	0.664	0.817	0.834	0.812	0.837	0.838
TCRC(5)	0.627	0.846	0.836	0.851	0.857	0.838
TCRC(6)	0.636	0.870	0.878	0.853	0.864	0.866
TCRC(7)	0.634	0.849	0.852	0.874	0.870	0.846
TCRC(8)	0.749	0.894	0.869	0.876	0.867	0.875
TCRC(9)	0.789	0.867	0.867	0.908	0.868	0.881
TCRC(10)	0.788	0.880	0.853	0.859	0.866	0.869
TCRC(11)	0.821	0.888	0.898	0.898	0.890	0.907

the negative control (NC). The improvement is due to the NC data containing information of the assays such as the cell plate condition, environment temperature, and so on. Figure 8 illustrates the classification SR for C1 and C10 using machine learning algorithms with input data given by TCRC(*j*) and W$_5$ with and without information of NC. There is no doubt that incorporating NC into the input data does play an important role of providing more information to the machine learning algorithms, and this leads to a significant enhancement for ANN and SVM. It is particularly noted that tremendous increase in SR is observed when the input is based on a few TCRC data. Using wavelet transform and the highest concentration TCRC(1) data without and with NC, the SR increases from 0.550 to 0.870 for ANN, and 0.691 to 0.907 for SVM. Among the two data mining tools, it is preferable to use SVM since it consistently produces a higher SR than that using ANN. Hence, the remaining results presented in this work will be based on SVM and with TCRC(*j*) including negative curve as input data.

By including the negative control (NC) with the TCRC(11), the input contains 876 data points, and almost the same amount of data will be required for wavelets W$_i$, $i = 1, 2, \cdots, 5$ if all coefficients in wavelet decompositions are kept. However, it is well known that wavelet transform is especially effective for data compression. Utilizing this attractive feature, we could achieve the same or better performance by appropriately pruning the wavelet coefficients. Consequently, much less data is needed as input for SVM. Now, consider a 5-level wavelet decomposition is applied to TCRC(11) with NC, and let W$_5(i)$ denote the corresponding wavelet coefficients, where $i = 1, 2, \cdots, 6$. Note that W$_5(6)$ corresponds to the case when all the wavelet coefficients are included, i.e., W$_5(6)$: CA$_5$ + CD$_5$ + CD$_4$ + CD$_3$+ CD$_2$ + CD$_1$ and only one set of coefficients is kept in W$_5(1)$, where W$_5(1)$ is the CA$_5$. The length of input data for TCRC(11) with NC and W$_5$(i) are listed in Table 3. Compared with the original TCRC(11) data with NC, savings of 74 % and 97 % are achieved when using W$_5(4)$ and W$_5(1)$ as input. Clearly, a tremendous data reduction is achieved by pruning the wavelet coefficients. Applying the same approach for the highest concentration data TCRC(1) with NC, the corresponding wavelet coefficients W$_5(i)$ are shown in Table 3.



(a) ANN **(b)** SVM

Fig. 8 Classification SR of (**a**) ANN and (**b**) SVM by using different number of concentrations

Table 3 Length of input data using raw data and wavelet coefficients

	Raw	$W_5(6)$	$W_5(5)$	$W_5(4)$	$W_5(3)$	$W_5(2)$	$W_5(1)$
TCRC(11) + NC	876	889	450	229	117	60	30
TCRC(1) + NC	146	157	83	45	25	14	7

Recall that by taking all wavelet coefficients from TCRC(1) and TCRC(11), the SVM SR shown in Fig. 8 is 90 % and 95 %, respectively. In Table 4, we evaluate the SVM performance for C1-C10 classification using coefficients based on various wavelet decomposition levels as listed in Table 3. Using the information from all TCRC(11) data, the SR is the range of 80.8 % to 96.4 %. It is remarkable to observe that even using $\mathrm{W}_5(2)$ and $\mathrm{W}_5(1)$ with 60 and 30 data points as input, over 80 % SR is achieved. Note that the original raw data contains 876 data, and using $\mathrm{W}_5(2)$ and $\mathrm{W}_5(1)$, the input data is being reduced by 93 % and 97 %. It is also worthwhile to note that over 90 % SVM SR is recorded when the input data is based on only the highest concentration TCRC(1) and with $\mathrm{W}_5(i)$ for $i > 3$. Without going through a detail study to optimize the wavelet decompositions, we now fucus on the data mining tools based on SVM(11) and SVM(1). Here, SVM(11) denotes SVM using input data from the $\mathrm{W}_5(4)$ based on entire TCRC(11), and SVM(1) corresponds to input using $\mathrm{W}_5(6)$ from the highest concentration TCRC(1). Thus, the input data in SVM(11) and SVM(1) are 229 and 157, and this produces a reduction of 74 % and 82 % compared to taking entire raw data TCRCs with 11 concentrations.

In Table 5, we present the two-cluster MOA classification results using SVM(11) and SVM(1). The two-cluster is defined by clustering C1 and Cj where $j \neq 1$. Let the error in each classification be (1-SR), and define the average error as E=$[\,(1 - SR(C1/C2)) + (1 - SR(C1/C3)) + \ldots + (1 - (SR(C1/C10))]\,/6$. The results presented in Table 5 reveal that the performance for SVM(11) and SVM(1) are comparable, and the average error in SR is 0.1592 and 0.1547 for SVM(11) and SVM(1), respectively. However, it is important to note that while SVM(11) produces low SR 74.2 % and 77.9 % for (C1/C2) and (C1/C6) classification, the corresponding SR using SVM(1) increases to 87.9 % and 83.3 %. Therefore, by examining the SR values resulting from SVM(11) and SVM(1), we can enhance the accuracy for the MOA classification. Let SVM denote by selecting the best SR from SVM (1) and SVM(11), and the SVM SR for the two-cluster classification is reported in Table 5. Note that, the average error E for SVM is now reduced to 0.1085. Although further improvement is possible by investigating other data from TCRC(k) where k $\neq$ 1 and 11 and by optimizing the wavelet coefficients, we will only carry out computation using SVM(1) and SVM(11) and the best value will be recorded as SVM in this study.

Multi-cluster classification

In many applications, a data set may contain more than two clusters. Therefore, it is necessary to expand machine learning algorithm from binary classification to multi-cluster classification. ANN can easily be adapted to deal with multi-cluster cases, and we only

Table 4 SVM SR for C1 and C10 classification

	Raw	$W_5(6)$	$W_5(5)$	$W_5(4)$	$W_5(3)$	$W_5(2)$	$W_5(1)$
TCRC(11) + NC	0.857	0.947	0.944	0.964	0.905	0.821	0.808
TCRC(1) + NC	0.845	0.909	0.904	0.930	0.795	0.779	0.770

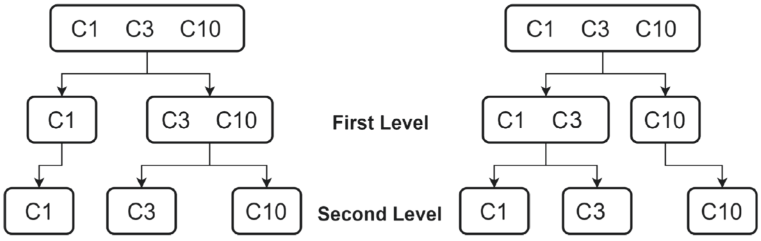
Table 5 SVM SR for two-cluster classification

	(C1/C2)	(C1/C3)	(C1/C4)	(C1/C6)	(C1/C8)	(C1/C10)
SVM(11)	0.742	0.995	0.845	0.779	0.720	0.964
SVM(1)	0.879	1.000	0.694	0.941	0.657	0.901
SVM	0.879	1.000	0.845	0.941	0.720	0.964

need to assign the number of output neurons equal to the number of clusters. Since the performance of ANN is not as effective as SVM, we will not present the results using ANN. To carry out multi-cluster classification for SVM, we utilize a tree structure strategy [41]. Due to the imbalanced data in the 63 chemical compounds, our study will focus on extending the SVM algorithm for classifications with three and four clusters.

First, consider an example of a three clusters C1, C3 and C10. Since C3 contains only four data sets which is much smaller than C1 and C10, a reasonable tree structure for classification is shown in Fig. 9, in which a binary classification is conducted at each level. Figure 9 illustrates extending a two-level tree structure methodology for three-cluster classification. For the left configuration, we first label both compounds in C3 and C10 as one class C, then a binary classification for C1 and C is carried out. In the second level, the cluster C is further classified into C3 and C10 by using binary classification again. Similarly, for the right configuration, C1 and C3 are first labelled as one class C in the first level, and then be classified in SVM algorithm. Although it is feasible to have a tree structure by first combing C1 and C10 into one class, this selection will not be recommended. It is known that SVM works well for balanced data set such that the training and test data in both groups are almost equal. For the structure given by [C3 and (C1+C10)] with 70 % training, we have a highly imbalanced data since there are only 3 training data in one group and 23 data in the other group. Using the same approach, we consider another three-cluster for C1, C2 and C10. The MOA classification results for the two test cases are shown in Table 6. Obviously, the clustering SR is sensitive to the specified tree structure. The overall SR for [(C1 + C3) and C10] is significantly higher than for [C1 and (C3 + C10)] as reported. However, for the second example, [C1 and (C2 + C10)] will be a better choice.

The methodology using a tree structure approach can be further extended to deal with four-cluster classification, and let consider MOA classification for C1, C3, C4 and C10 as shown in Fig. 10. Note than C3 and C4 contains less data than C1 and C10. By the same argument presented for a three-cluster classification, we propose two-level and



s

Fig. 9 Tree-structure for three-cluster classification

Table 6 SVM SR for three-cluster classification

C*	C1 and (C10 + C*)	(C1 + C*) and C10
C3	0.841	0.968
C2	0.880	0.810

three-level configurations for the four-cluster MOA classification. To evaluate the robustness of SVM for multi-cluster classification, we construct another test case by replacing C4 data sets by C2 data. The SVM SR for the four configurations are reported in Table 7. The best configuration is based on [(C1+C*)+C3] and [C10], for which 85.6 % and 84.7 % classification SR is achieved for $C^* = C4$ and $C^* = C2$, respectively.

Dose response curves

In order to deal with the limited data sets for some clusters considered in this study, we proposed to construct the Dose Response Curves (DRCs), and utilizing the information from DRCs as input to the developed learning algorithm.

Instead of using TCRCs as input data, we now utilize information from Dose Response Curves (DRCs) as input to SVM. The DRCs reveals the effect of the chemicals at different concentrations, and it can be computed from the difference between the time concentration response curves and the negative control curve at a particular time point. Let, denote

$$TE_t(k) = \frac{TCRC_t(k) - NC_t}{NC_t} * 100\% \tag{3}$$

where $TE_t(k)$ is the toxicity effect (TE) of the chemical with kth concentration at time t, NC_t is the cell index value of the negative control at time t. From this definition, it is clear that when $TE_t(k) = 0$, it implies that the chemical compound with concentration k has no toxicity effect to the cell growth at time t. Similarly, we can also define the TE by the area under the curve (AUC) as suggested in [14]:

$$TE_t(k) = \frac{AUC\{TCRC_t(k)\} - AUC\{NC_t\}}{AUC\{NC_t\}} * 100\% \tag{4}$$

where $AUC\{TCRC_t(k)\}$ denotes the area under the curve TCRC(k) between 0 to t hours, $AUC\{NC_t\}$ is the area under the negative control curve between 0 to t hours. Using

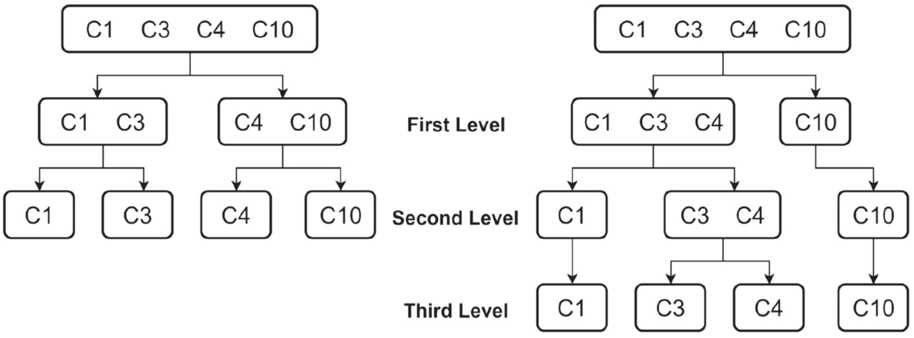


Fig. 10 Tree-structure for Four-cluster classification

Table 7 SVM SR for four-cluster classification

	Two-level approach		Three-level approach	
C*	[C1+C3]& &[C*+C10]	[C1+C*] &[C3+C10]	[(C1+C*)+C3] &[C10]	[(C1+C3)+C*] &[C10]
C4	0.807	0.750	0.856	0.838
C2	0.839	0.805	0.847	0.825

$TE_t(k)$, we can construct a sequence of TE at time t. For example, using the 11 concentrations TCRCs of the tested 63-chemicals data, the DRC can be computed at time T such that

$$DRC(T) = [TE_T(1) \quad TE_T(2) \quad \cdots \quad TE_T(11)]\,. \tag{5}$$

From (5), and using the TE_t defined by (3), we can define DRC at any given time t in our data set. Taking the compound 5-FU in cluster 1 as an example, we construct the DRC at 24 h, 48 h and 72 h as shown in Fig. 11.

According to the definition of DRCs in (5), DRC contains information regarding the reaction of the cell growth to the increment of the chemical concentrations. It is thus reasonable to assume that the compounds having different MOA may trigger different concentration-related reactions. Consequently, using DRCs data as training set may offer a way to improve the classification SR for those data that are not easy to be classified using TCRCs as input. Based on this approach, we carry out a SVM binary classification for C1 and C10 using DRC as input at a specified time point, and then linking the results at different time points together. The SVM results for clustering C1 and C10 are reported in Fig. 12. Different from using 803 data in TCRC(11), only 11 data are taken as input using DRC at a given time. The computational time is faster than that required based on TCRC as input data, but the overall SR is obviously not as good as those using TCRC.

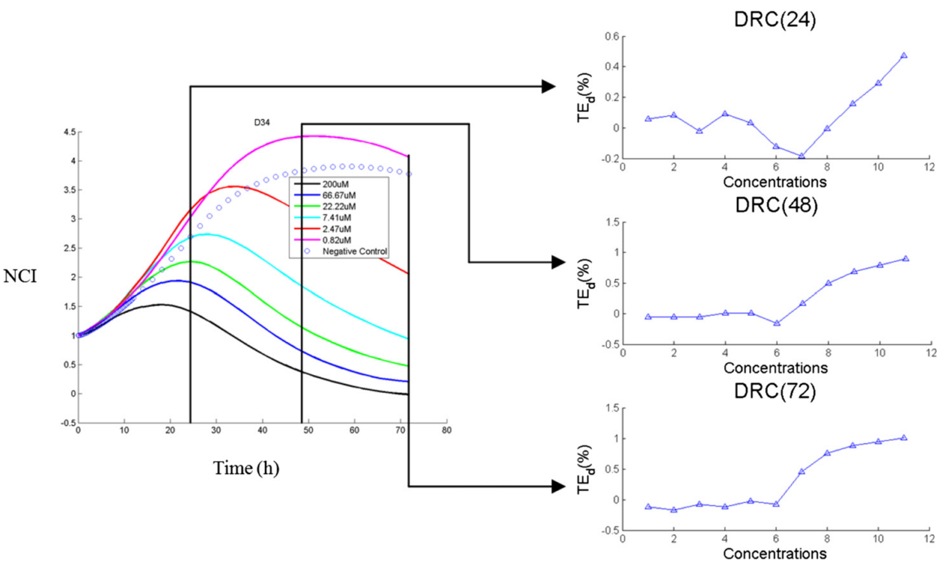


Fig. 11 DRC of 5-FU from TCRCs

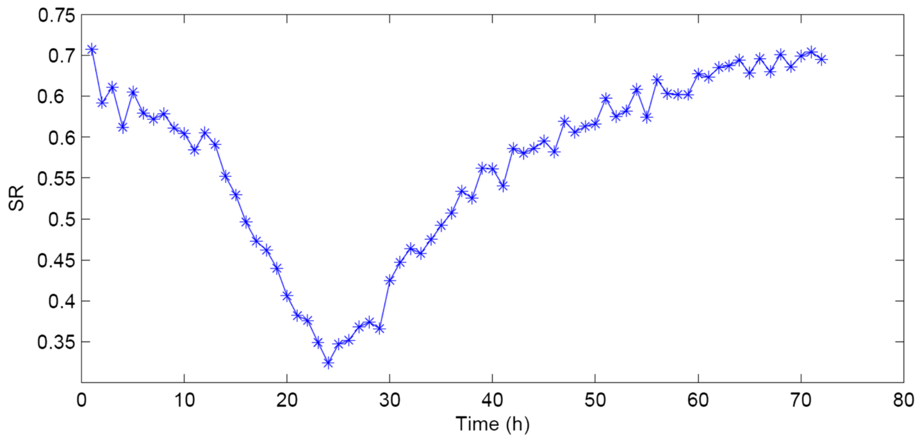
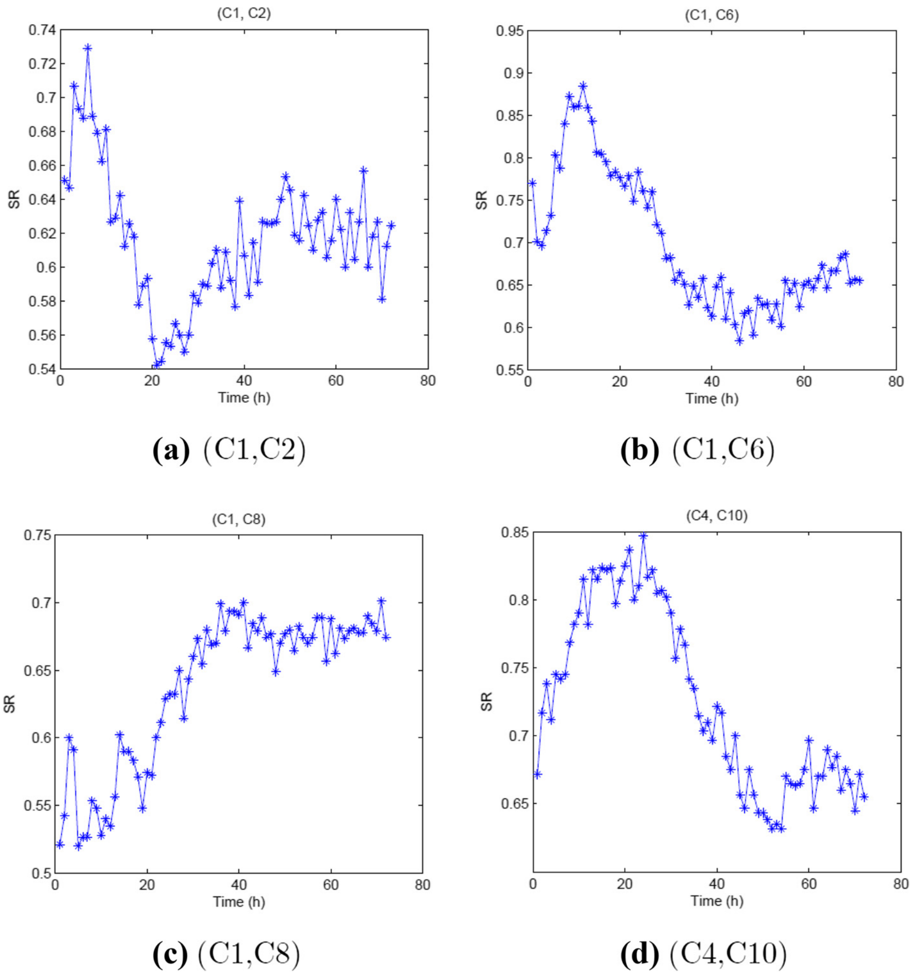


Fig. 12 SVM SR distribution for two-cluster C1 and C10 using DRC(*t*)



(a) (C1,C2)

(b) (C1,C6)

(c) (C1,C8)

(d) (C4,C10)

Fig. 13 SR distribution for (**a**) (C1, C2), (**b**) (C1, C6), (**c**) (C1, C8) and (**d**) (C4, C10)

Table 8 Selected time interval for TCRCs

Two-cluster	(C1,C2)	(C1,C6)	(C1,C8)	(C4,C10)
Selected interval	30–72 h	1–30 h	25–72 h	1–40 h

However, the plot in Fig. 12 reveals useful information, namely the time interval leading to a better SR can be determined. Thus, the methodology may offer a possible way to improve the performance of machine learning algorithm for imbalanced data set, since the time interval corresponding to low SR can be discarded in the input data.

Figure 13 displays the SR for binary clustering (C1, C2), (C1, C6), (C1, C8), (C4,C10). Recall that the four cases represent typical imbalanced data, and poor SR is observed using TCRC(11) as input as reported in Table 5. In Table 8, appropriate time intervals are selected by ignoring the time intervals corresponding to low SR. Using the TCRC(11) selected at the specified time intervals, the SR using SVM applied to RCRC(11) and $W_5(4)$ are reported in Table 9. Using the selected TCRC at certain time interval for cases with imbalanced data, the SVM SR is clearly improved for all cases as shown in Table 9. However, more work is needed to investigate the best way to utilize the information from DRC to further enhance the performance of SVM.

Conclusion

In this paper, we present an innovative approach using machine learning for toxicity assessment. The computational tools are developed based on ANN and SVM, which are capable of learning data from given TCRCs with known MOA clustering information and then making MOA classification for untested chemical compounds. There are two challenges and difficulties of this work. First the input data arising from the time-series TCRC data contains more than 850 data, and secondly, only limited data set are available for some clusters. A novel data processing technique using wavelet transform is introduced, so that not only a great reduction in input data is achieved but the MOA classification is more accurate due to wavelet coefficients have the ability to extract important features from the original TCRC data. Instead of using more than 850 data from the TCRCs with 11 concentrations, we only require 229 and 157 wavelet coefficients as input data to the developed data mining tools. In this study, it is also revealed that taking account the information from the negative control curve enhances the performance of the MOA classification. It has been illustrated that the machine learning algorithm can be improved by utilizing information from DRC, so that a time interval leading to higher classification success rate can be selected as input. From the computational simulations, SVM is more effective compared to ANN for MOA classification. The developed SVM classifier has been tested for multi-cluster MOA classification, and impressive SR in the range

Table 9 Improvement of SR by using TCRCs at selected time points from DRC distribution

	Time for 1–72 h		Selected time	
	TCRC	$W_5(4)$	TCRC	$W_5(4)$
(C1, C2)	0.734	0.742	0.736	0.797
(C1, C6)	0.809	0.779	0.830	0.857
(C1, C8)	0.699	0.720	0.803	0.766
(C4, C10)	0.695	0.795	0.716	0.832

of 85 to 95 % is obtained for m-cluster classification where $2 \leq m \leq 4$. The present work concludes that SVM is an effective and powerful machine learning tool for toxicity profiling.

It is noted that the proposed SVM is tested on the limited training and testing data, to perform a reliable validation of the proposed machine learning approach, it is desirable if more testing data are available. Even though the present study focuses on a MOA classification, the approach could be extended to other type of classifications such as a Globally Harmonized System (GHS) classification in toxicology investigation. Instead of a supervised learning approach, it is of great interest to consider an unsupervised methodology. Moreover, to better handle a multi-cluster classification and to enhance the robustness of a machine learning approach, it is useful to develop an expert system consisting of various classifiers, so that reliable classification results can be determined by incorporating a validation procedure.

Additional file

Additional file 1: 63 chemicals in 10 MOA clusters. This file includes the 63 compounds used for in the MOA classification. Details including the used solvent and concentrations are also provided. (PDF 19 kb)

Abbreviations

CI: Cell index; DRC: Dose response curve; MOA: Mode of action; ANN: Artificial neural network; SR: Successful rate; RTCA: Real time cell analysis; SVM: Support vector machine; TCRC: Time-concentration response curve.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YZ, YSW, JD, and CA carried out the study and drafted the manuscript. YZ and YSW performed the computation and simulations. YSW, WZ and SG initiated the project. SG, WZ, DYH and CJ provided background work, testing data and helpful input in preparing the article. All authors read and approved the final version of the manuscript.

Acknowledgements

The research is supported by Alberta Centre For Toxicology (ACFT), MITAC Accelerate Program and Natural Science and Engineering Research Council of Canada.

Author details

[1]Department of Mathematical and Statistical Science, University of Alberta, T6G 2G1, Edmonton, Canada. [2]Department of Mathematics and Statistics, Grant MacEwan University, T5P 2P7, Edmonton, Canada. [3]Department of Laboratory Medicine and Pathology, University of Alberta, T6G 2B7, Edmonton, Canada. [4]Alberta Health, T5J 1S6, Edmonton, Canada. [5]Alberta Centre for Toxicology, University of Calgary, T2N 4N1, Calgary, Canada. [6]AACEA Biosciences Inc, 92121, San Diego, USA.

Received: 13 October 2015 Revised: 4 April 2016
Published online: 13 May 2016

References

1. Xia M, Huang R, Witt KL, Southall N, Fostel J, Cho M, Jadhav A, Smith CS, Inglese J, Portier CJ, et al. Compound cytotoxicity profiling using quantitative high-throughput screening. Environ Health Perspect. 2008;116:284–91.
2. Cronin MT, Madden JC. In Silico Toxicology: Principles and Applications. Cambridge, UK; 2010.
3. Dearden JC. In silico prediction of drug toxicity. J Comput-Aided Mol Des. 2003;17:119–27.
4. Kavlock RJ, Ankley G, Blancato J, Breen M, Conolly R, Dix D, Houck K, Hubal E, Judson R, Rabinowitz J, et al. Computational toxicology - a state of the science mini review. Toxicol Sci. 2008;103:14–27.
5. Kavlock R, Dix D. Computational toxicology as implemented by the US EPA: providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk. J Tox Env Health. 2010;13:197–217.
6. Xing JZ, Zhu L, Jackson JA, Gabos S, Sun XJ, Wang XB, Xu X. Dynamic monitoring of cytotoxicity on microelectronic sensors. Chem Res Toxicol. 2005;18:154–61.
7. Xing JZ, Zhu L, Gabos S, Xie L. Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity. Toxicol Vitro. 2006;20:995–1004.
8. Boyd JM, Huang L, Xie L, Moe B, Gabos S, Li XF. A cell-microelectronic sensing technique for profiling cytotoxicity of chemicals. Anal Chim Acta. 2008;615:80–7.
9. Abassi Y. Label-free and dynamic monitoring of cell-based assays. Biochemica. 2008;3:8–11.
10. Slanina H, König A, Claus H, Frosch M, Schubert-Unkmeir A. Real-time impedance analysis of host cell response to meningococcal infection. J Microbiol Methods. 2011;84:101–8.

11. Zhang M, Aguilera D, Das C, Vasquez H, Zage P, Gopalakrishnan V, Wolff J. Measuring cytotoxicity: a new perspective on LC50. Anticancer Res. 2007;27:35–8.
12. Zhang M, Das C, Vasquez H, Aguilera D, Zage PE, Gopalakrishnan V, Wolff JE. Predicting tumor cell repopulation after response: mathematical modeling of cancer cell growth. Anticancer Res. 2006;26:2933–6.
13. Pan T, Khare S, Ackah F, Huang B, Zhang W, Gabos S, Jin C, Stampfl M. In vitro cytotoxicity assessment based on KC 50 with real-time cell analyzer (RTCA) assay. Comput Biol Chem. 2013;47:113–20.
14. Pan T, Huang B, Zhang W, Gabos S, Huang DY, Devendran V. Cytotoxicity assessment based on the AUC 50 using multi-concentration time-dependent cellular response curves. Anal Chim Acta. 2013;764:44–52.
15. Hawkins ED, Hommel M, Turner ML, Battye FL, Markham JF, Hodgkin PD. Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data. Nat Protoc. 2007;2:2057–67.
16. Opp D, Wafula B, Lim J, Huang E, Lo JC, Lo CM. Use of electric cell–substrate impedance sensing to assess in vitro cytotoxicity. Biosens Bioelectron. 2009;24:2625–9.
17. Cox JC, Coulter AR. Adjuvants - a classification and review of their modes of action. Vaccine. 1997;15:248–56.
18. Auer J, Bajorath J. Emerging chemical patterns: A new methodology for molecular classification and compound selection. J Chem Inf Model. 2006;46:2502–14.
19. Namasivayam V, Hu Y, Balfer J, Bajorath J. Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns. J Chem Inf Model. 2013;53:1272–81.
20. Judson R, Elloumi F, Setzer RW, Li Z, Shah I. A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model. BMC Bioinf. 2008;9:241.
21. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem. 2001;26:5–14.
22. Cheng F, Shen J, Yu Y, Li W, Liu G, Lee PW, Tang Y. In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. Chemosphere. 2011;82:1636–43.
23. Vanneschi L, Farinaccio A, Mauri G, Antoniotti M, Provero P, Giacobini M. A comparison of machine learning techniques for survival prediction in breast cancer. BioData Mining. 2011;4:12.
24. Beck D, Foster J. Machine learning classifiers provide insight into the relationship between microbial communities and bacterial vaginosis. BioData Mining. 2015;8:23.
25. Lareau CA, White BC, Oberg AL, McKinney BA. Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure. BioData Mining. 2015;8:5.
26. Lu P, Abedi V, Mei Y, Hontecillas R, Hoops S, Carbo A, Riera JB. Supervised learning methods in modeling of CD4+ T cell heterogeneity. BioData Mining. 2015;8:27.
27. Xi Z, Khare S, Cheung A, Huang B, Pan T, Zhang W, Ibrahim F, Jin C, Gabos S. Mode of action classification of chemicals using multi-concentration time-dependent cellular response profiles. Comput Biol Chem. 2014;49:23–35.
28. Ibrahim F, Huang B, Xing J, Gabos S. Early determination of toxicant concentration in water supply using MHE. Water Res. 2010;44:3252–60.
29. Pan T, Huang B, Xing J, Zhang W, Gabos S, Chen J. Recognition of chemical compounds in contaminated water using time-dependent multiple dose cellular responses. Anal Chim Acta. 2012;724:30–39.
30. Vracko M. Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies. Curr Comput-Aided Drug Des. 2005;1:73–78.
31. Anthony ML, Rose VS, Nicholson JK, Lindon JC. Classification of toxin-induced changes in 1 h NMR spectra of urine using an artificial neural network. J Pharm Biomed Anal. 1995;13:205–11.
32. Gini G, Lorenzini M, Benfenati E, Grasso P, Bruschi M. Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. J Chem Inf Comput Sci. 1999;39:1076–80.
33. Haykin S. Neural networks: a comprehensive foundation. Upper Saddle River, NJ, USA: Prentice Hall; 2008, p. 2.
34. Zhao C, Zhang H, Zhang X, Liu M, Hu Z, Fan B. Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology. 2006;217:105–19.
35. Yap C, Cai C, Xue Y, Chen Y. Prediction of torsade-causing potential of drugs by support vector machine approach. Toxicol Sci. 2004;79:170–7.
36. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16:906–14.
37. Balabin RM, Lomakina EI. Support vector machine regression (SVR/LS - SVM) - an alternative to neural networks (ANN) for analytical chemistry comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst. 2011;136:1703–1712.
38. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell. 1989;11:674–93.
39. Mallat S. A wavelet tour of signal processing. San Diego, California, USA: Academic press; 1999.
40. Krim H, Tucker D, Mallat S, Donoho D. On denoising and best signal representation. IEEE Trans Inf Theory. 1999;45: 2225–38.
41. De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology. 2000;81:3178–92.