MOA classification for cytotoxicity assessment by machine learning algorithms

Yile Zhang[a, *], Yau Shu Wong[a], Jian Deng[a], Cristina Anton[b], Stephan Gabos[c], Weiping Zhang[d], Dorothy Yu Huang[e], Can Jin[f]

[a]Department of Mathematical and statistical Sciences, University of Alberta, Edmonton, Alberta, T6G2G1, Canada
[b]Department of Mathematics and Statistics, Grant MacEwan University, Edmonton, Alberta, T5P2P7, Canada
[c]Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, Alberta, T6G2B7, Canada
[d]Alberta Health, Edmonton, Alberta, T5J1S6, Canada
[e]Alberta Centre for Toxicology, University of Calgary, Calgary, Alberta, T2N4N1, Canada
[f]ACEA Biosciences Inc, San Diego, California, 92121, USA

.

Abstract

Real Time Cell Analysis (RTCA) technology can monitor cellular changes continuously over the entire exposure period. Combining with different testing concentrations, the profiles have great potential in probing the MOA of testing substances, and thereafter, the prediction of toxicities. Although many indices have been developed to quantify the chemical toxicity, it is difficult to apply them in MOA classification. In this paper, we present an innovative approach using machine learning 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, and they are capable of learning data from given TCRCs with known MOA clustering information and then making MOA classification for the chemical of unknown toxicity. A novel data processing step using wavelet transform is introduced, so that not only a great reduction in input data is achieved, but the wavelet coefficients have the ability to extract important features from the original TCRC data. Moreover, to enhance the performance of the machine learning algorithm, we utilized information from the dose response curves so that time interval leading to higher classification success rate can be selected as input. 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 each tested with 11 concentrations. 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. The classification results show that the proposed machine learning method have a potential in large scale MOA identification and classification.

Keywords: Time-concentrations response curve, mode of action, machine learning, wavelet transform, dose response curve, support vector machine, artificial neural network.

***Corresponding Author.** Tel: +1 780 863 9630
Email Address: yile2@ualberta.ca (Y. Zhang), yauwong@ualberta.ca (Y.S. Wong), deng2@ualberta.ca (J. Deng), cristina.anton@hotmail.ca (C. Anton), sgabos@ualberta.ca (S. Gabos), W. Zhang (weiping.zhang@gov.ab.ca), yhuang@ucalary.ca (D. Y. Huang), cjin@aceabio.com (C. Jin)

1 Introduction

In recent years, considerable progress has been reported in the study of cytotoxicity profiling using *in vitro* assays data [1]. In order to deal with increasing number of chemical compounds, it is important to develop fast and effective cytotoxicity screening methods which are capable of analyzing the *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, 7, 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 achieved. The TCRC profiles can be used to study the cell-chemical interaction mechanism.

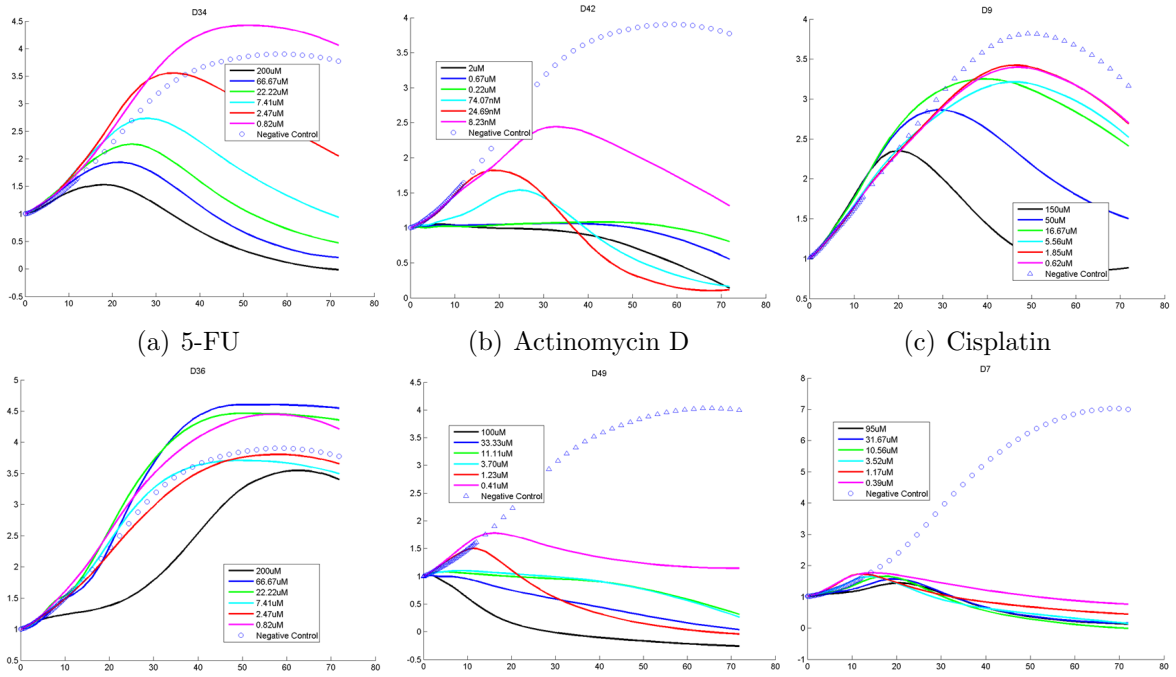
Some analysis methods have been developed to extract useful information from the TCRCs. For example, LC_{50} reflects the chemical concentration that leading to killing 50% of tested cells [11], KC_{50} uses an exponential model to calculate the LC_{50} value [12, 13, 15], AUC_{50} represents the area 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 be lost. The time-dependent indices can also be problematic if inappropriate time points or intervals are chosen. For example, the AUC for the time series between 1-72 hours can be different from that between 1-60 hour. In addition, all these indices have the primary goal of detecting toxicity potency of the testing chemicals. The application into MOA classification are indirect, and not tested [17, 18].

Recently, the machine learning method has been reported for toxicity classifications [21, 22]. Judson et al. [23] applied various machine learning algorithms to classify the high-throughput *in vitro* end point data. Burbidge et al. [24] employed the support vector machine for drug design using the pharmaceutical data, Cheng et al. [25] investigated the toxicity pattern recognition for diverse industrial chemicals with substructure. It is known that the mode of action (MOA) is closely related to the mechanism of chemicals [26]. According to pre-set criterion, the chemicals can be classified into different MOA clusters [27]. MOA classification provides a better understanding and enhance our knowledge of potential toxicity pathways.

In this study, we focus on MOA classification for the 63 chemical compounds provided by the Alberta Centre for Toxicology. The list of the chemicals and their ten-cluster MOA classification are given in Appendix. The same chemicals were investigated by Pan et al. [14, 15] and Xi et al. [16] for cytotoxicity assessment. Here, using the TCRCs from the 63 reference chemicals as input data, we develop a methodology based on machine learning to perform clustering analysis. The results were compared with known MOA for the 63 chemical compounds.

There are many difficulties in using the TCRCs from the 63-compounds data sets. In Figure 1, we display six TCRC profiles of the compounds in cluster 1: DNA/RNA target. It is observed that although the six compounds belong to one cluster, their corresponding TCRCs are quite different. On the other hand, compounds in different clusters may have similar TCRCs as illustrated in Figure 2 showing similar TCRCs profiles resulting from four different clusters. It should be noted each compound contains a set of TCRCs with

11 concentrations, and thus the entire TCRCs time series consists of more than 800 data points. Taking a large amount of data as input is not a trivial subject in machine learning, and we will discuss the difficulties in the next section. To the best of our knowledge, no study has been reported in using machine learning algorithm with entire TCRCs as input for toxicity assessment.



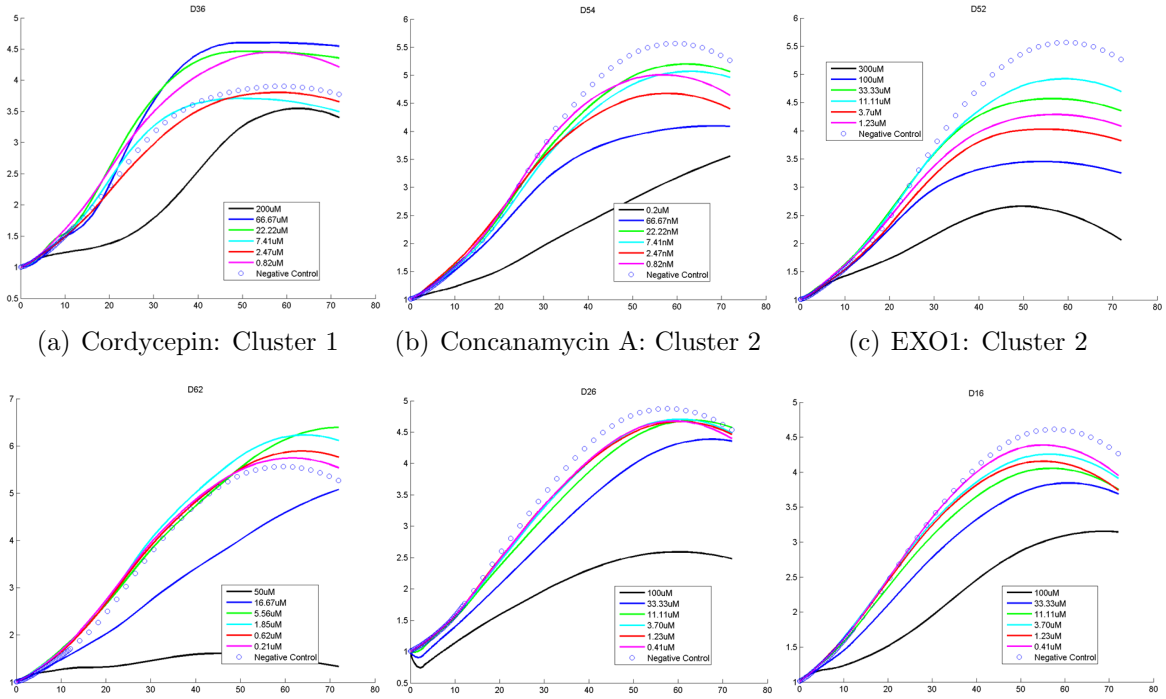
(a) 5-FU (b) Actinomycin D (c) Cisplatin

(d) Cordycepin (e) Doxorubicin(DOX) (f) Topotecan

Figure 1: TCRCs from Cluster 1: DNA/RNA targets

The main contributions presented in this work are twofold: to propose a novel computational tool based on machine learning for cytotoxicity assessment and to validate the effectiveness of the developed algorithms using the TCRCs as input by comparing the results with the known MOA clustering applied to the 63 chemical compounds. The machine learning methods are based on the artificial neural network (ANN) and support vector machine (SVM) with supervised learning algorithm. To resolve the difficulty due to taking a large data set from the entire TCRCs, an innovative idea of using wavelet transform is implemented. Consequently, instead of directly using the TCRCs as input data to the machine learning algorithms, the wavelet coefficients are selected as input. It will be verified that the application of wavelet preprocessing step not only leads to a significant reduction in the input data, but more importantly, it is capable of extracting useful information and features of the original TCRCs. Consequently, a great improvement in the success rate in clustering analysis is achieved. 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 utlizing the information from DRCs as input to the developed learning algorithm.

The remainder of this paper is organized as follows. In Section 2, the background and data preprocessing of the present study is presented. Section 3 focuses on the machine learning approach based on ANN and SVM, and the application of wavelet transform is also discussed. To validate the developed computational tools, we present binary and multi-cluster classifications using ANN and SVM applied to the 63 compounds in Section



(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

Figure 2: TCRCs from four different clusters

4. The effectiveness of SVM is clearly demonstrated by the excellent agreement resulted from the known clustering based on MOA applied to the tested chemical compounds. In section 5, the use of DRCs is proposed, and the advantage of utilizing DRCs to improve the performance of the machine learning algorithm for limited data set is reported. Finally, concluding remark are presented in Section 6.

2 Backgrounds and data preprocessing

2.1 Cell line

Human hepatocellular 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℃ tissue culture hood with 95% humidity and 5% CO2.

2.2 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℃. The highest testing concentration is at most 1/500th of the stock concentration, to keep the solvent (DMSO or ethanol) concentration under 0.2%. Each chemical were tested with 11 concentrations, with 1:3 serial dilution for the initial test, to cover full response range from no toxicity, to maximum toxicity.

2.3 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 hours 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 hours post exposure.

2.4 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 $R_{cell}(f_k)$ and $R_b(f_k)$ 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.

3 Machine learning algorithms

Machine learning is an active research topic, and it has been successfully applied to solve many complex problems in science and engineering. As a classifier, machine learning algorithm has the ability to learn and to make prediction on given data set. The goal of the present work is to develop learning algorithm to predict the mode of action of untested compounds by TCRCs resulting from the known MOA clustering information. We consider two learning algorithms, namely artificial neural network (ANN) and support vector machine (SVM). The use of wavelet transform to enhance the performance and effectiveness of ANN and SVM will also be presented.

3.1 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, 31, 32]. A typical feedforward multi-layer ANN [33] is shown in Figure 3.

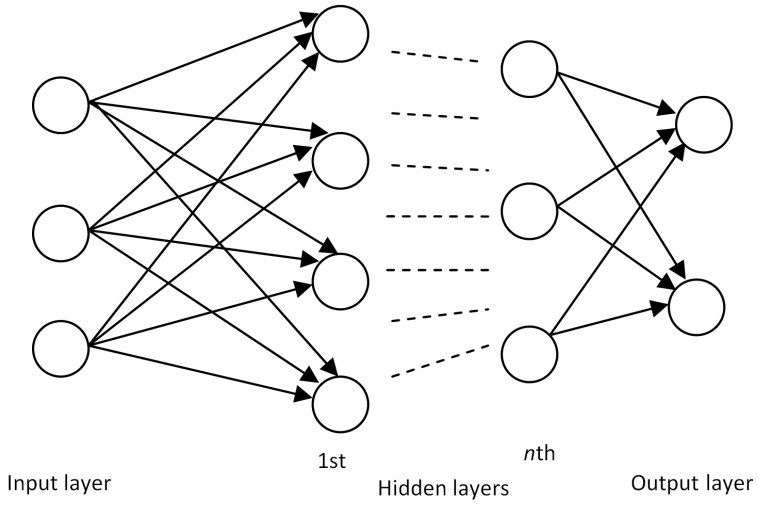


Figure 3: Feedforward n-layer ANN

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 output 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 connected by the weights represented by the arrows between the layers, and the information passing though the neurons are determined by the weights. The weights are determined 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 $f\colon X \to Y$ to fit a set of example pairs (x, y), x$\in$ X, y$\in$ 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.

3.2 Support vector machine

In addition to the 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 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.

3.3 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 800 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 cytotoxicity data. An orthonormal basis $\psi_{s,\tau}(t)$ [39] having scale parameter s and translation parameter τ can be expressed in the following form:

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

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 Figure 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 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, CD_{n-1}, $\cdots$, CD_2, 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 $W_n(m)$ 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., $n = 3$), $W_3(4)$ means that four wavelet coefficients: CA_3+CD_3+CD_2+CD_1 are kept and $W_3(2)$ implies taking two wavelet coefficients CA_3+CD_3.

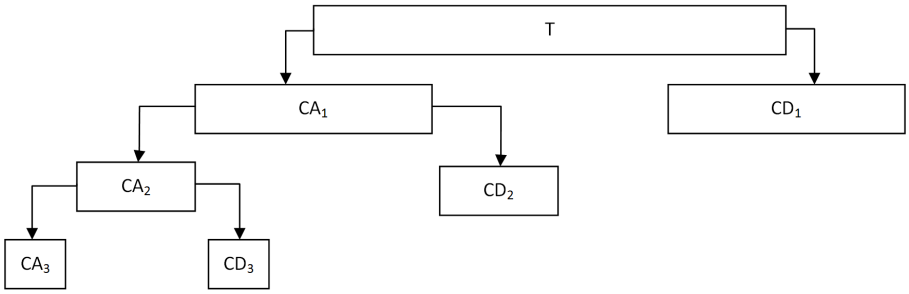


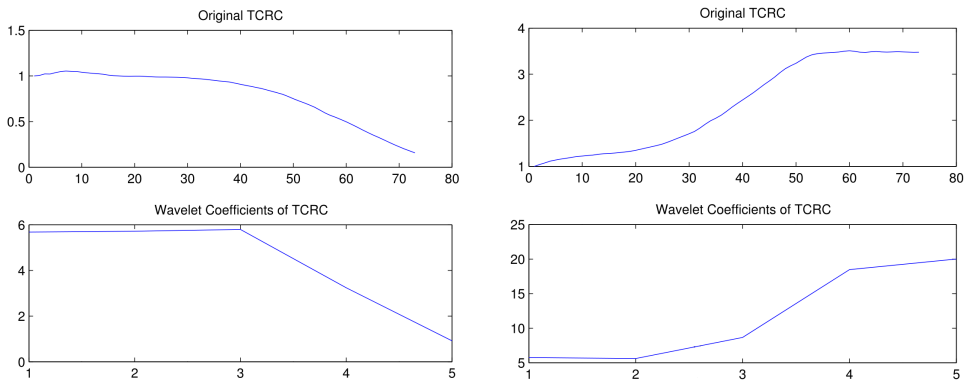
Figure 4: Three-level wavelet decomposition

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 $W_5(1)$. 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 $W_5(1)$.

We now illustrate how to construct input data for machine learning. A given set of TCRCs is arranged as shown in Figure 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 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

4 Computational simulations

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 Figure 7. Note that C1 and C10 contain 33 compounds,



(a) Actinomycin D (b) Cordycepin

Figure 5: Plots of original TCRCs and wavelet coefficients $W_5(1)$

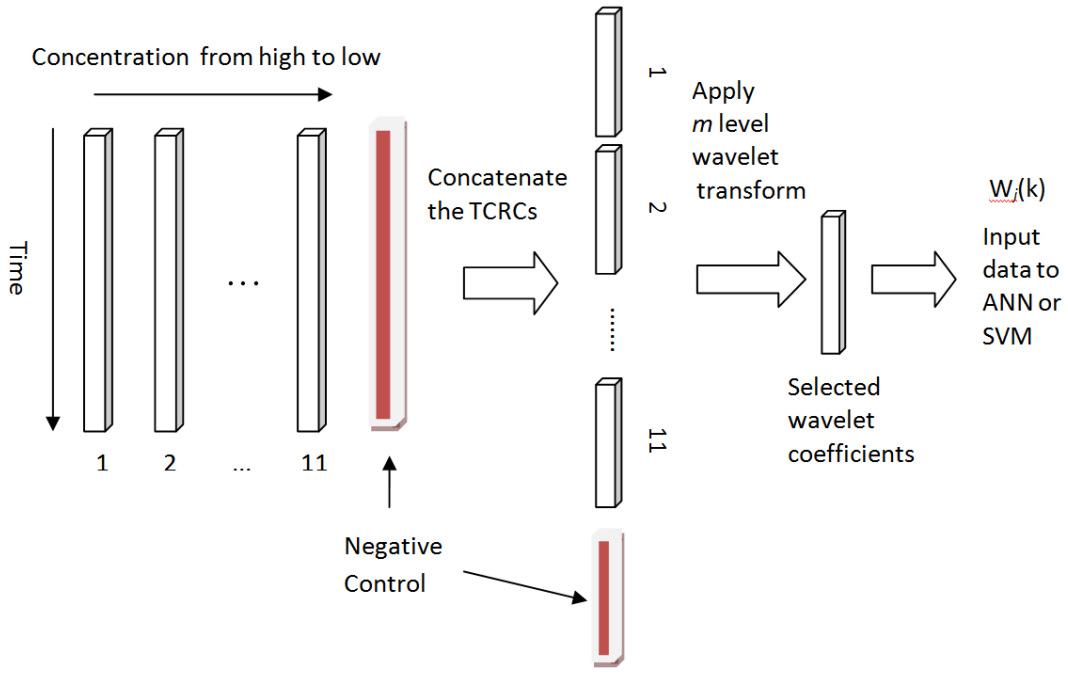


Figure 6: Input data to machine learning using wavelet transform

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 target 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.

4.1 Binary Classification

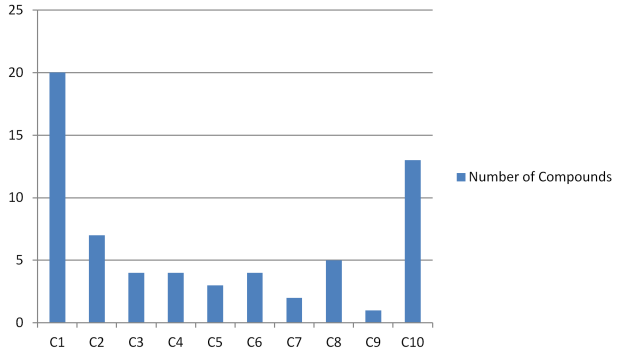


Figure 7: Distribution of 63 compounds

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. In Table 1, we list the number of compounds in training and test set with 70 % and 50% training data.

Table 1: Number of compounds in training and test set with 70 % and 50% training

	Training set			Test set		
	C1	C10	Total	C1	C10	Total
70%	14	9	23	6	4	10
50%	10	6	16	10	7	17

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 providing more information to the input should lead to better performance for the machine learning tools. In our application, the input is given by the TCRCs and a typical data set consists of 11 concentrations. Recall that TCRC(1) denotes the data taking only the highest concentration, TCRC(2) for data with the first two highest concentrations and TCRC(11) data including all 11 concentrations. In Table 2 and 3, we report the SR of ANN and SVM using 70% of the observations as training data. TCRC(j) with $j = 1, 2, \cdots, 11$ denotes the raw data, and W_i for $i = 1, 2, .., 5$ are the corresponding wavelets resulting from the ith-level wavelet decomposition. When using raw data TCRC(n), the SR are poor and unacceptable when $n = 1$. As expected, the SR improves when the value of n is increased.

The best SR for ANN and SVM are 86.4% and 82.1% when $n = 9$ and 11, respectively. The advantage of applying multi-level wavelet decomposition to TCRC(n) is clearly demonstrated in Tables 2 and 3. The best SR for ANN and SVM using wavelet coefficients as input are 87.9% and 90.8%. Over 80% SR using ANN and SVM can be achieved by employing wavelets applied to TCRC(4) and TCRC(3). It should be noticed that all wavelet coefficients are kept in the input W_i for $i = 1, 2, .., 5$. Additional enhancement by pruning the wavelet coefficients will be discussed later on.

The performance for the classification of C1 and C10 using machine learning algorithms with input data given by TCRC(n), W_3 and W_5 are shown in Figure 8. The computational results presented so far are based on input data taken from TCRC(n). However, a significant improvement in SR can be achieved by including information from the negative control (see Figure 6), and the results are reported in in Figure 8 and in Table 4. The improvement is due to the negative control data containing information of the assays such as the cell plate condition, environment temperature, and so on. Therefore, it does play a role of providing more information to the input data.

From the results presented in Figure 8, we observe that 96% SR is achieved using W_5 by taking 11 concentrations and 70% training. It is also important to notice that even with one concentration, the SR of W_5 is over 91%. The corresponding results using 50% training data are also reported in Table 4, and the robustness of learning algorithms is confirmed. From the computational results reported for the classification of C1 and C10, we conclude that it is preferable to include the negative control in the input. Hence the following results will include the negative control in the input data.

It has been demonstrated that input data with 11 concentrations will produce better performance for machine learning tools. By including the vector for the negative control, the input based on TCRC(11) contains 864 data points, and almost the same amount of data points will be required for wavelets W_i, $i = 1, 2, \cdots, 5$ by keeping all wavelet coefficients in wavelet decompositions. Since the wavelet transform is capable of data compression and feature extraction, we will demonstrate that appropriately pruning the wavelet coefficients, we could achieve the same or better performance, but using much

less data as input. Consider a 5-level wavelet decomposition is applied to TCRC(11), 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) and W_5(i) is listed in Table 5. Compared with the original TCRC(11) data, savings of 74% and 96% 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.

The performances for C1 and C10 classification using ANN and SVM with 70% and 50% training are reported in Table 6. It is noted that SVM generally performs better than ANN. When 803 raw data in TCRC(11) is used for 70% training data, the SR of SVM is 85.7% , while using the optimal pruning of the wavelet coefficients $W_5(4)$, the SR of SVM can be 96.4%. Not only over 10% improvement in SR is achieved, but the input data is also significantly reduced because $W_5(4)$ contains 212 data points, which is about 26% of the original raw data. The remaining computational results reported in this work will be based on SVM and using the input data $W_5(4)$.

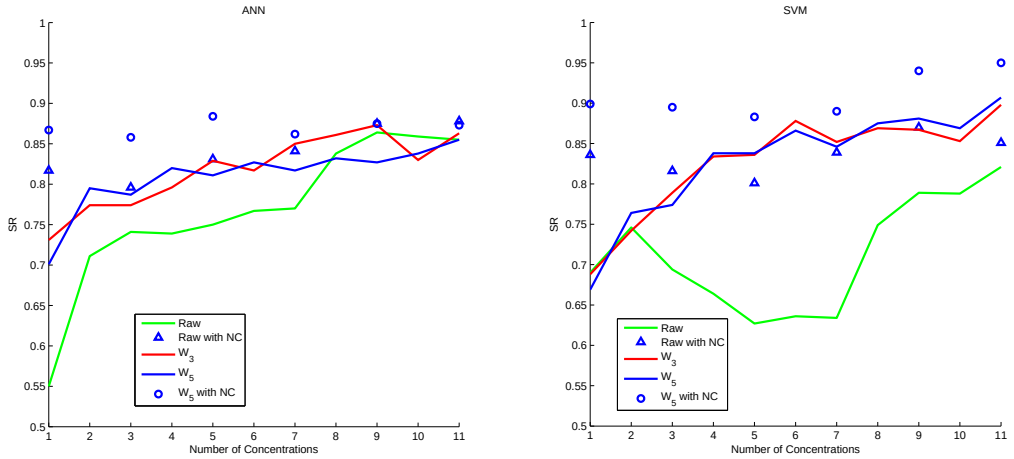
Table 2: ANN SR with 70% training with different concentrations

	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

Table 3: SVM SR with 70% training 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

In Table 7, we present the two-cluster classification results using SVM and $W_5(4)$ as input data with 70% and 50% training. The two-cluster is defined by clustering C1, Cj



(a) ANN (b) SVM

Figure 8: Classification SR using different number of concentrations

Table 4: SR with and without NC.

	Training	Raw	Raw + NC	W_5	W_5+NC
ANN	70%	0.855	0.872	0.855	0.873
SVM	70%	0.821	0.857	0.907	0.947
ANN	50%	0.827	0.835	0.838	0.848
SVM	50%	0.791	0.831	0.881	0.921

Table 5: Length of input data using raw TCRC(11) and $W_5(i)$

Raw	$W_5(6)$	$W_5(5)$	$W_5(4)$	$W_5(3)$	$W_5(2)$	$W_5(1)$
803	818	415	212	109	56	29

Table 6: SR using raw TCRC(11) and $W_5(i)$

	Training	Raw	$W_5(6)$	$W_5(5)$	$W_5(4)$	$W_5(3)$	$W_5(2)$	$W_5(1)$
ANN	70%	0.872	0.873	0.882	0.885	0.850	0.847	0.837
SVM	70%	0.857	0.947	0.944	0.964	0.905	0.821	0.808
ANN	50%	0.835	0.848	0.843	0.863	0.804	0.789	0.779
SVM	50%	0.831	0.921	0.905	0.910	0.859	0.768	0.778

where $j \neq 1$, and C10, Ck, where $k \neq 10$. The SR for some clusters is better than the others, and this may be affected by the size of clusters. In general, one would expect that for a two-cluster classification in which the number of compounds Cj or Ck is much smaller than that for C1 or C10, then the resulting SR may be relatively poor due to the imbalanced data set. For these difficult cases, the SVM performance may be improved by considering the input data utilizing information from the Dose Response Curve (DRC). This will be discussed in more details in section 5.

Table 7: SVM SR for two-cluster classification

Training		C1	C2	C3	C4	C6	C8	C10
70%	C1		0.742	0.995	0.845	0.779	0.720	0.964
	C10	0.964	0.807	0.948	0.795	0.824	0.756	
50%	C1		0.744	0.968	0.892	0.787	0.780	0.910
	C10	0.910	0.787	0.952	0.834	0.877	0.781	

4.2 Classification for multiple clusters

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 need to assign the number of output neurons equal to the number of clusters. However, the performance of ANN are not as effective as SVM, thus we will not present the results using ANN. For SVM, we could utilize a tree structure strategy [41]. Consider an example of classification for three clusters C1, C3 and C10. The tree structure methodology is shown in Figure 9, in which a binary classification is conducted at each level.

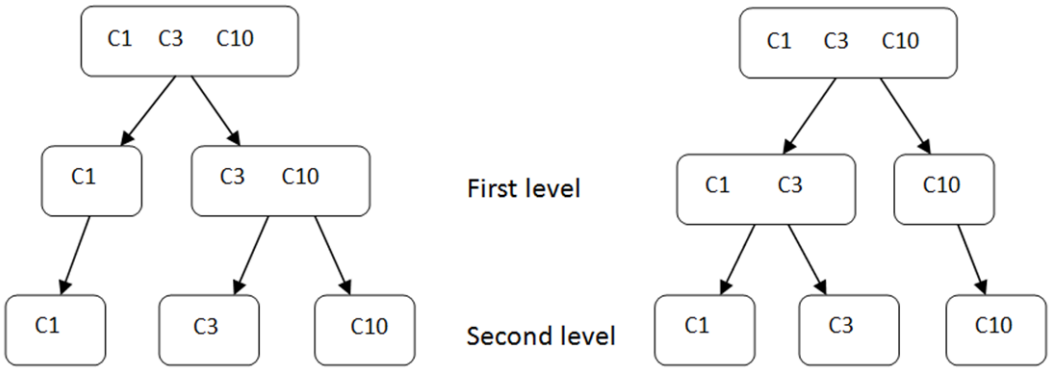


Figure 9: Tree-structure for three-cluster

Figure 9 illustrates two-level tree structure 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. Now, let denote the structure on the left and right as [C1 and (C10 + C3)] and [(C1 + C3) and C10] respectively, the classification results based on above tree configurations are shown in Table 8.

Table 8: SVM SR for three-cluster C1, C3 and C10

%	C1 and (C10 + C3)	(C1 + C3) and C10
70%	0.841	0.968
50%	0.836	0.915

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. Using 70% training set, the classification SR is more than 95% compared to 84% using the other tree configuration. Recall that in the binary classification, the pruning of the wavelet coefficients affects the SR. In Table 9, we present the SR for three clusters classification (C1+C3) and C10 using raw data TCRC(11) and 5-level wavelet transform $W_5(i)$. The results in the Table 9 reconfirm that the optimal pruning of wavelet

Table 9: SVM SR for three-cluster C1, C3 and C10

	TCRC(11)	$W_5(6)$	$W_5(5)$	$W_5(4)$	$W_5(3)$
Data length	803	818	415	212	109
70%	0.828	0.853	0.921	0.968	0.832
50%	0.812	0.847	0.899	0.915	0.813

coefficients is consistent for binary and multi-cluster classifications. More importantly, compared with the SR based on the TCRC(11) and $W_5(4)$, substantial improvement in classification SR are obtained from 82.8% to 96.8% using 70% training, and from 81.2% to 91.5% using 50% training. The advantage of using wavelet transform for input data is clearly illustrated.

The methodology using tree structure can be further extended to deal with classification for four clusters C1, C3, C4 and C10 as shown in Figure 10. However, more variations to construct tree structure are admitted as the number of clusters increases. In Figure 10, we present two-level and three-level configurations for the three-cluster classification. The classification SR for four configurations are reported in Table 10, and the results are based on SVM with 70% training applied to TCRC(11) and $W_5(4)$. The best configuration is based on [(C1+C4)+C3] and [C10] for which 85.6% classification SR using wavelet $W_5(4)$ as input is achieved.

Table 10: SVM SR for four-cluster C1, C3, C4 and C10 with 70% training

TCRC(11)	[C1+C3] & [C4+C10]	[C1+C4] & [C3+C10]	[(C1+ C4) + C3] & [C10]	[(C1+ C3) + C4] & [C10]
Raw	0.625	0.665	0.671	0.693
$W_5(4)$	0.807	0.750	0.856	0.838

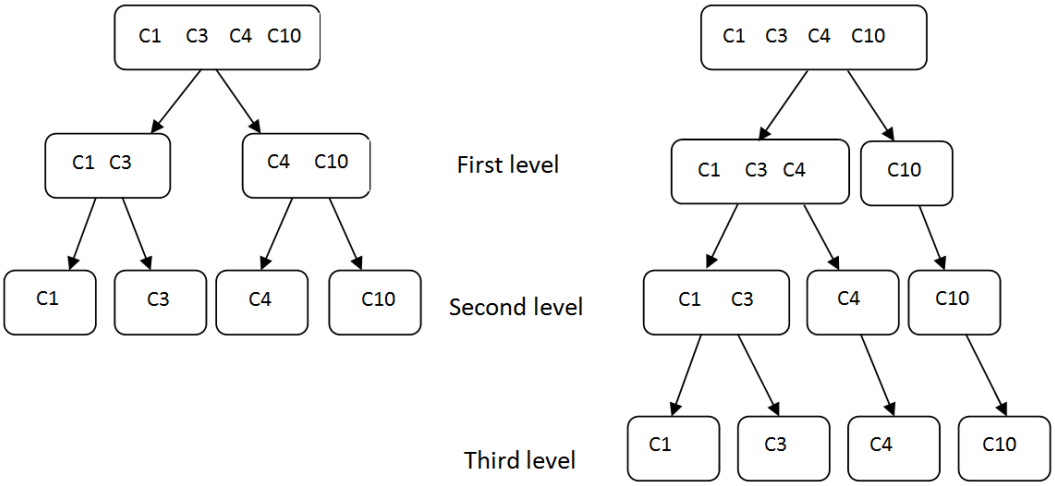


Figure 10: Tree-structure for four-cluster

5 Dose Response Curves

Instead of using TCRCs as input data, we now consider utilizing 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 curves 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 $TE_t(k)$, 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 24h, 48h and 72h as shown in Figure 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

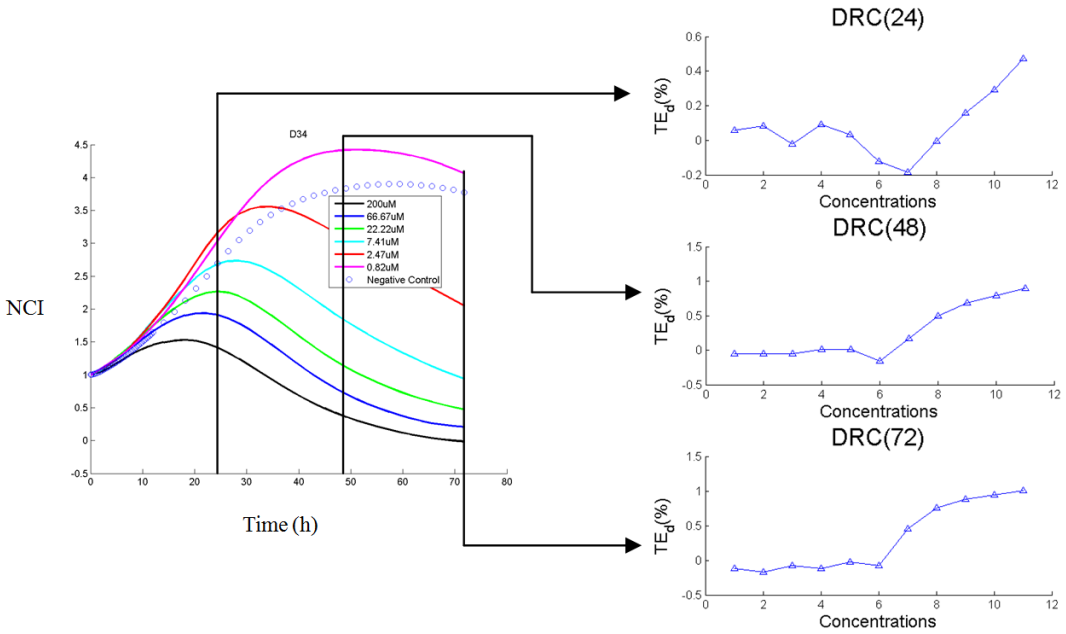


Figure 11: DRC of 5-FU from TCRCs.

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 Figure 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. However, the plot in Figure 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.

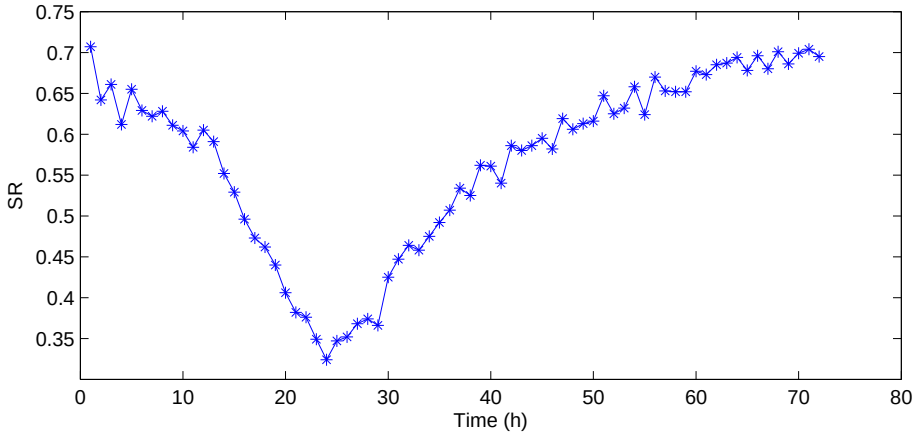
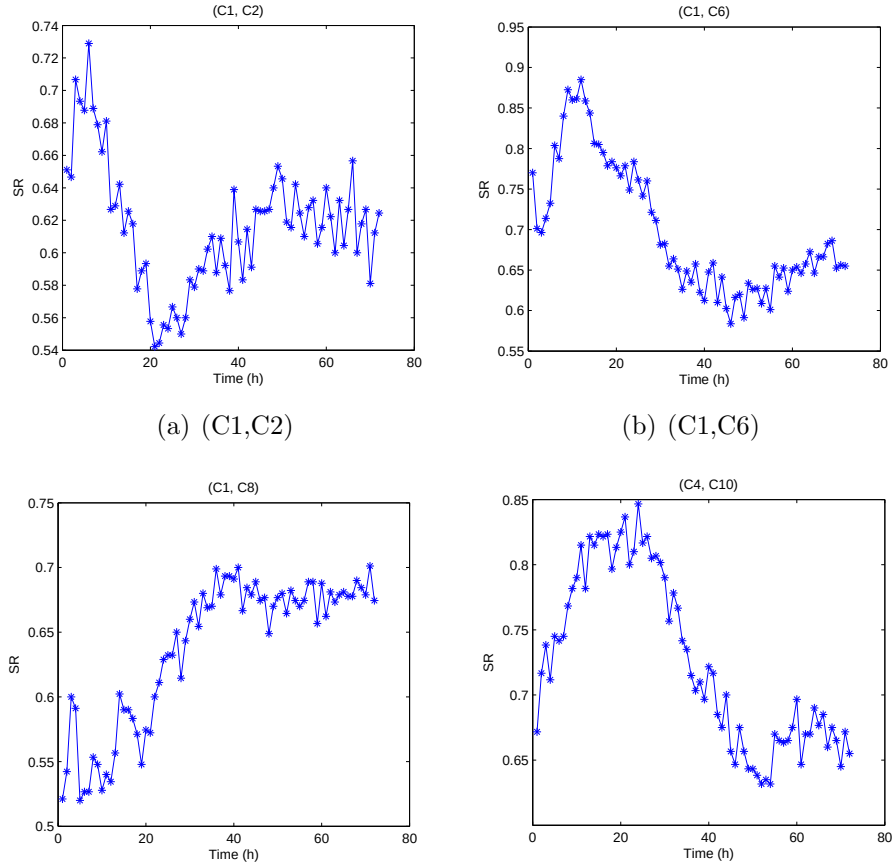


Figure 12: SVM SR distribution for two-cluster C1 and C10 using DRC(t)

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 7. In Table 11, 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 12. 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 12. However, more work is needed to investigate the best way to utilize the information from DRC to further enhance the performance of SVM.



(a) (C1,C2) (b) (C1,C6)

(c) (C1,C8) (d) (C4,C10)

Figure 13: SR distribution for (C1, C2), (C1, C6), (C1, C8), (C4,C10)

Table 11: Selected time interval for TCRCs

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

6 Conclusion

In this paper, we present an innovative approach using machine learning for cytotoxicity 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 800 data, and secondly, only limited data set are available for some clusters. A novel data processing technique using wavelet transform is introduced,

Table 12: Improvement of SR by using TCRCs at selected time points from DRC distribution

	Time for 1-72h		Selected time shown in Table 11	
	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

so that not only a great reduction in input data is achieved but the wavelet coefficients have the ability to extract important features from the original TCRC data. It has been illustrated that the performance of 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. Impressive SR in the range of 85% to 95% is obtained using SVM for m-cluster MOA classification where $2 \leq m \leq 4$. The present work concludes that SVM is an effective and powerful machine learning tool for the study of toxicity profiling.

It is noted that the proposed SVM is tested on 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 GHS classification in toxicology investigation. To better handle a multi-cluster classification and to enhance the performance and robustness of a machine learning approach, we are currently pursuing on developing an expert classification system in which it consists of three or more classifiers. The important feature of the expert system is that it will be capable of producing the best classification by incorporating a validation procedure. The details and results will be reported in another paper.

Acknowledgement

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

References

[1] M. Xia, R. Huang, K. L. Witt, N. Southall, J. Fostel, M. Cho, A. Jadhav, C. S. Smith, J. Inglese, C. J. Portier, et al, Compound cytotoxicity profiling using quantitative high-throughput screening, Environ. Health Perspect. 116 (2008) 284-291.

[2] M. T. Cronin and J. C. Madden, In Silico Toxicology: Principles and Applications. No. 7.. Royal Society of Chemistry, 2010.

[3] J. C. Dearden, In silico prediction of drug toxicity, J. Comput.-Aided Mol. Des. 17 (2003) 119-127.

[4] R. J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, et al, Computational toxicologya state of the science mini review, Toxicol. Sci. 103 (2008) 14-27.

[5] R. Kavlock and D. Dix, 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 13 (2010) 197-217.

[6] J. Z. Xing, L. Zhu, J. A. Jackson, S. Gabos, X. J. Sun, X. B. Wang, and X. Xu, Dynamic monitoring of cytotoxicity on microelectronic sensors, Chem. Res. Toxicol. 18 (2005) 154-161.

[7] J. Z. Xing, L. Zhu, S. Gabos, and L. Xie, Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity, Toxicol. in Vitro 20 (2006) 995-1004,.

[8] J. M. Boyd, L. Huang, L. Xie, B. Moe, S. Gabos, and X. F. Li, A cell-microelectronic sensing technique for profiling cytotoxicity of chemicals, Anal. Chim. Acta 615 (2008) 80-87.

[9] Y. Abassi, Label-free and dynamic monitoring of cell-based assays, Cell 4 (2008).

[10] H. Slanina, A. König, H. Claus, M. Frosch, and A. Schubert-Unkmeir, Real-time impedance analysis of host cell response to meningococcal infection, J. Microbiol. Methods 84 (2011) 101-108.

[11] K. Kothawad, A. Pathan, and M. Logad, Evaluation of in vitro anti-cancer activity of fruit lagenaria siceraria against MCF7, HOP62 and DU145 cell line, Int. J. Pharm. & Technol, 4 (2012) 3909-4392.

[12] M. Zhang, C. Das, H. Vasquez, D. Aguilera, P. E. Zage, V. Gopalakrishnan, and J. E. Wolff, Predicting tumor cell repopulation after response: mathematical modeling of cancer cell growth, Anticancer Res, 26 (2006) 2933–2936.

[13] M. Zhang, D. Aguilera, C. Das, H. Vasquez, P. Zage, V. Gopalakrishnan, and J. Wolff, Measuring cytotoxicity: a new perspective on LC50, Anticancer Res. 27 (2007) 35-38.

[14] T. Pan, B. Huang, W. Zhang, S. Gabos, D. Y. Huang, and V. Devendran, Cytotoxicity assessment based on the AUC 50 using multi-concentration time-dependent cellular response curves, Anal. Chim. Acta 764 (2013) 44-52.

[15] T. Pan, S. Khare, F. Ackah, B. Huang, W. Zhang, S. Gabos, C. Jin, M. Stampfl, In vitro cytotoxicity assessment based on KC 50 with real-time cell analyzer (RTCA) assay, Comput. Biol. Chem. 47 (2013) 113-120.

[16] Z. Xi, S. Khare, A. Cheung, B. Huang, T. Pan, W. Zhang, F. Ibrahim, C. Jin, S. Gabos, Mode of action classification of chemicals using multi-concentration time-dependent cellular response profiles, Comput. Biol. Chem. 49 (2014) 23-35.

[17] E. D. Hawkins, M. Hommel, M. L. Turner, F. L. Battye, J. F. Markham, and P. D. Hodgkin, Measuring lymphocyte proliferation, survival and differentiation using CFSE time-series data, Nat. Protoc. 2 (2007) 2057-2067.

[18] D. Opp, B. Wafula, J. Lim, E. Huang, J. C. Lo, and C. M. Lo, Use of electric cell–substrate impedance sensing to assess in vitro cytotoxicity, Biosens. Bioelectron. 24 (2009) 2625-2629.

[19] C. A. Marchant, Computational toxicology: a tool for all industries, WIREs Comput. Mol. Sci. 2 (2012) 424-434.

[20] M. P. Smithing and F. Darvas, Hazardexpert: an expert system for predicting chemical toxicity, In ACS Symposium series American Chemical Society, 1992.

[21] J. Auer and J. Bajorath, Emerging chemical patterns: A new methodology for molecular classification and compound selection, J. Chem. Inf. Model. 46 (2006) 2502-2514.

[22] V. Namasivayam, Y. Hu, J. Balfer, and J. Bajorath, Classification of compounds with distinct or overlapping multi-target activities and diverse molecular mechanisms using emerging chemical patterns, J. Chem. Inf. Model. 53 (2013) 1272-1281.

[23] R. Judson, F. Elloumi, R. W. Setzer, Z. Li, and I. Shah, A comparison of machine learning algorithms for chemical toxicity classification using a simulated multi-scale data model, BMC Bioinf. 9 (2008) 241.

[24] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Comput. Chem. 26 (2001) 5-14.

[25] F. Cheng, J. Shen, Y. Yu, W. Li, G. Liu, P. W. Lee, and Y. Tang, In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods, Chemosphere 82 (2011) 1636-1643.

[26] K. H. Kuck and U. Gisi, FRAC mode of action classification and resistance risk of fungicides, Mod. crop Prot. Compd., (2007) 415-432.

[27] J. C. Cox and A. R. Coulter, Adjuvantsa classification and review of their modes of action, Vaccine 15 (1997) 248-256.

[28] F. Ibrahim, B. Huang, J. Xing, and S. Gabos, Early determination of toxicant concentration in water supply using MHE, Water Res. 44 (2010) 3252-3260.

[29] T. Pan, B. Huang, J. Xing, W. Zhang, S. Gabos, and J. Chen, Recognition of chemical compounds in contaminated water using time-dependent multiple dose cellular responses, Anal. Chim. Acta 724 (2012) 30-39.

[30] M. Vracko, Kohonen artificial neural network and counter propagation neural network in molecular structure-toxicity studies, Curr. Comput.-Aided Drug Des. 1 (2005) 73-78.

[31] M. L. Anthony, V. S. Rose, J. K. Nicholson, and J. C. Lindon, Classification of toxin-induced changes in 1 h NMR spectra of urine using an artificial neural network, J. Pharm. Biomed. Anal. 13 (1995) 205-211.

[32] G. Gini, M. Lorenzini, E. Benfenati, P. Grasso, and M. Bruschi, 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. 39 (1999) 1076-1080.

[33] S. Haykin and N. Network, A comprehensive foundation, Neural Networks, 2 (2004).

[34] C. Zhao, H. Zhang, X. Zhang, M. Liu, Z. Hu, and B. Fan, Application of support vector machine (SVM) for prediction toxic activity of different data sets, Toxicology 217 (2006) 105-119.

[35] C. Yap, C. Cai, Y. Xue, and Y. Chen, Prediction of torsade-causing potential of drugs by support vector machine approach, Toxicol. Sci. 79 (2004) 170-177.

[36] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (2000) 906-914,

[37] R. M. Balabin and E. I. Lomakina, 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 136 (2011) 1703-1712, 2011.

[38] S. G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674-693.

[39] S. Mallat, *A wavelet tour of signal processing*, Academic press, 1999.

[40] H. Krim, D. Tucker, S. Mallat, and D. Donoho, On denoising and best signal representation, IEEE Trans. Inf. Theory 45 (1999) 2225-2238.

[41] G. De'ath and K. E. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology 81 (2000) 3178–3192.

Appendix: 63 compounds in 10 clusters using MOA classification

Cluster I: DNA/RNA-Nucleic Acid Target
Fluoropyrimidine 5-fluorouracil (5-FU)
2'-Deoxy-2',2'-difluorocytidine (Gemcitabine HCl/dFdc)
Etoposide phosphate
Doxorubicin (DOX)
Merbarone
Clofarabine (CLOF)
Hydroxyurea (HU)
SN 38
Topotecan
Irinotecan (CPT-11)
Cytosine β-D-arabinofuranoside (Cytarabine)
ABT-888 (veliparib)
Mitoxantrone
CRT0044876 (7-nitro-1H-indole-2-carboxylic acid)
NU 7026
Mitomycin
Cordycepin
Actinomycin D
Cisplatin
Ochratoxin A
Cluster II: Transport Protein-Primary Activetransporter Targets
Brefeldin A (BEF)
Exo 1 (2-(4-Fluorobenzoylamino)-benzoic acid methyl ester)
Leptomycin B (LMB)
Concanamycin A (CMA)
Thapsigargin
BHQ
Cluster III: Protein-Actin,Targets
Bafilomycin A1
Cytochalasin D
Cytochalasin B
Latrunculin A
Latrunculin B
Cluster IV: Protein-Tublin,Targets
Docetaxel
Paclitaxel
Vincristine Sulfate
Vinblastine sulfate

Cluster V: Ribosome-50S Subunit Targets
Emetine
Puromycin
Anisomycin
Cluster VI: Transport Proteins -Electrochemical Potential-driven Transporters
Oligomycin
Antimycin A
Rotenone
CCCP
Cluster VII: Ion Channel Targets
Valproic acid
BAPT-am
Cluster VIII: Enzyme Targets
Cyclosporin A
FK-506 (tacrolimus)
(S)-HDAC-42
SAHA (Vorinostat/ Suberoylanilide Hydroamic Acid Zolinza)
W7 HCl
Cluster IX
benzo[a]pyrene
Cluster X: Protein- Motor Targets
Monastrol
Strityl-cysteine
Dimethylenastron
Y-27632
HA 1100 hydrochloride
Ro32-3555
Batimastat
MLCKInhibPep18
Blebbistatin
ML 7 hydrochloride
FAKInhibitor14
PF573228
PF 431396