Journal of Molecular Graphics and Modelling 28 (2009) 236–244

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Prediction of novel and selective TNF-alpha converting enzyme (TACE) inhibitors and characterization of correlative molecular descriptors by machine learning approaches

Yong Cong [a], Xue-gang Yang [a], Wei Lv [a], Ying Xue [a,b,*]

[a] Key Lab of Green Chemistry and Technology in Ministry of Education, College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China
[b] State Key Laboratory of Biotherapy, Sichuan University, Chengdu 610064, People's Republic of China

ARTICLE INFO

Article history:
Received 19 May 2009
Received in revised form 17 July 2009
Accepted 3 August 2009
Available online 8 August 2009

Keywords:
TNF-α converting enzyme (TACE)
Inhibitor
Non-inhibitor
Machine learning (ML)
The support vector machine (SVM)

ABSTRACT

The inhibition of TNF-α converting enzyme (TACE) has been explored as a feasible therapy for the treatment of rheumatoid arthritis (RA) and Crohn's disease (CD). Recently, large numbers of novel and selective TACE inhibitors have been reported. It is desirable to develop machine learning (ML) models for identifying the inhibitors of TACE in the early drug design phase and test the prediction capabilities of these ML models. This work evaluated four ML methods, support vector machine (SVM), *k*-nearest neighbor (*k*-NN), back-propagation neural network (BPNN) and C4.5 decision tree (C4.5 DT), which were trained and tested by using a diverse set of 443 TACE inhibitors and 759 non-inhibitors. A well-established feature selection method, the recursive feature elimination (RFE) method, was used to select the most appropriate descriptors for classification from a large pool of descriptors, and two evaluation methods, 5-fold cross-validation and independent evaluation, were used to assess the performances of these developed models. In this study, all these ML models have already achieved promising prediction accuracies. By using the RFE method, the prediction accuracies are further improved. In *k*-NN, the model gives the best prediction for TACE inhibitors (98.32%), and the SVM bears the best prediction for non-inhibitors (99.51%). Both the *k*-NN and SVM model give the best overall prediction accuracy (98.45%). To the best of our knowledge, the SVM model developed in this work is the first one for the classification prediction of TACE inhibitors with a broad applicability domain. Our study suggests that ML methods, particularly SVM, are potentially useful for facilitating the discovery of TACE inhibitors and for exhibiting the molecular descriptors associated with TACE inhibitors.

1. Introduction

Tumor necrosis factor-α (TNF-α) is a pro-inflammatory cytokine, produced primarily by activated monocytes and macrophages, and plays an important role in immunity and inflammation [1,2]. TNF-α is initially synthesized as a membrane-anchored 26 kDa precursor. The proteolysis of the peptide bond between Ala76 and Val77 in TNF-α leads to the mature cytokine being shed from the cell as a homotrimer of the 17 kDa C-terminal fragment [1–3]. This processing step is performed by a 85 kDa membrane-anchored zinc metalloprotease, a TNF-α converting enzyme (TACE) [4,5]. Involvement of this cytokine has been validated in many disease states such as rheumatoid arthritis (RA) [6] and Crohn's disease (CD) [7] and implicated in diverse neuroimmunological pathologies such as multiple sclerosis [8], Alzheimer's [9], and stroke [10]. At present, data from human clinical trials have supported the therapeutic utility of two anti-TNF biologics, etanercept and infliximab in RA and CD, which are the soluble TNF-α receptor-Fc dimmer (Immunex) and anti-TNFα antibody (Centocor), respectively [11]. However, these two biologics are orally inactive and are inevitably administered parenterally. Furthermore, biologics commonly have a long half-life in the body and the above two biologics are not the exception. This may untimely cause infection incidents during the anti-TNF-α therapy, since TNF-α also plays an important role as a self-defense factor against pathogen. Another concern is allergic reactions to the biologics and the neutralization antibody productions [12,13]. Based on these points of view, inhibition of TACE by small molecular weight orally bioavailable drugs would be more desirable than these biological agents in blocking downstream cytokine production.

TACE is structurally categorized into the ADAM (a disintegrin and metalloproteinase domain) family that belongs to the superfamily of metzincin, but its catalytic site is quite similar to that of

* Corresponding author at: Key Lab of Green Chemistry and Technology in Ministry of Education, College of Chemistry, Sichuan University, Chengdu 610064, People's Republic of China. Tel.: +86 28 85418330.
E-mail addresses: yingxue_99@yahoo.com, yxue@scu.edu.cn (Y. Xue).

1093-3263/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2009.08.001

matrix metalloproteinases (MMPs) [14]. Not surprisingly, many of MMP inhibitors reported over the years in clinical trials were also found to be good inhibitors of TACE. The discovery of TACE inhibitors logically started from the better studied MMP inhibitors. For instance, some pipecolic hydroxamic acid-based TACE inhibitors were modified and optimized during a broad search for MMP inhibitors [15]. However, the structural similarity between the active sites of various MMPs and TACE also offers a big challenge for the design of specific TACE inhibitors with highly selectivity over MMPs. One major approach reported to address this issue was to exploit the differences in the primary and secondary structures of the active site loops of MMPs and TACE that form the S1′ pocket. Molecular modeling studies of homology models [16] and X-ray crystal structure of TACE complexes [14] revealed that the S1′ and S3′ pocket of TACE are connected, providing a subsite with a unique shape that sets it apart from the S1′ subsites of the MMPs. The acetylenic P1′ group could fit in the tunnel connecting the S1′ and S3′ pockets of TACE [17]. So some novel, selective hydroxamic TACE inhibitors bearing a butynyloxy P1′ group were designed and synthesized [18–21]. It is also found that certain quinolines when properly placed could fit into the curvature of the TACE S1′ pocket while clashing with the linear MMP S1′ pocket [16,22]. Incorporating these advantageous quinoline P1′ group onto different hydroxamic acid cores led to the discovery of a new series of TACE inhibitors [23–27]. A large variety of intrinsic potent hydroxamic TACE inhibitors have been developed based on this way, but hydroxamic acids often exhibited poor oral absorption in vivo and significant metabolic liabilities (rapid hydrolysis and glucoronidation) [28]. There is considerable interest to discover alternative groups to replace the hydroxamic acid as another major approach. Several cases have been reported about highly selective non-hydroxamate TACE inhibitors. For instance, Kamei et al. have created a series of reverse hydroxamate (RH) derivatives with excellent TACE inhibitory activities [29]. With the combination of these two approaches, the explosion of novel non-hydroxamate TACE inhibitors was suddenly brought about [30–34]. In particular, based on molecular modeling study, researchers at Bristol–Myers Squibb Pharmaceutical Research Institute (BMSPRI) developed a highly plausible pharmacophore model to rationalize the observed TACE activity of hydroxamate and non-hydroxamate (hydantoins and barbiturate) TACE inhibitors [35], and then used this pharmacophore model to guide the design of some fresh heterocyclic non-hydroxamate TACE inhibitors such as triazolones, imidazolones and pyrimidine-2,4,6-trione inhibitors of TACE [32,33]. Later, it is argued that compounds that are active against TACE as well as MMPs may be more efficacious than a selective TACE inhibitor only, since MMPs are also overexpressed in RA. So some dual TACE and MMP-13 inhibitors which do not inhibit MMP-1 were developed [36]. In addition to these TACE inhibitors, there remain lots of new TACE inhibitors designed by other strategies [37–41].

These TACE inhibitors from different chemical classes have different interactions with TACE. Sheppeck et al. (at BMSPRI) have also reported the interaction study of various classes of non-hydroxamate and hydroxamate TACE inhibitors [35]. In their report, it is suggested that hydroxamic acid forms two hydrogen bonds with Glu406 and Gly349 of TACE apart from a bidentate ligation of the hydroxyl and carbonyl oxygen to the zinc atom. The interaction of non-hydroxamate zinc binding group with the TACE active site differs from those of the hydroxamate ones because they ligate Zn in monodentate fashion and hydrogen bond to the same residues but in a fundamentally different way. Even so, the binding modes of more TACE inhibitors are still unknown. There is an urgent need for developing common pharmacophore for TACE inhibitors so that the design of various chemical classes of TACE inhibitors could be rationalized.

Although the structure-based and mechanism-based drug design methods reported have obtained great success in developing large number of novel and selective TACE inhibitors, the application of these methods may be hampered by the limited availability of target 3D co-crystal structures, and the difficult tasks of studying a highly diverse range of TACE inhibitors and analyzing multiple interaction modes, some of which are still ambiguous. Therefore, structure- and mechanism-based drug design alone may be insufficient for the revelation of universal TACE inhibitor pharmacophore. Moreover, selecting a priori possible candidate from an existing or virtual chemical database is considered as one of the most significant aspects in drug design. Hence, it is highly desirable to explore machine learning (ML) methods that make the identification of TACE inhibitors in the early drug design phase and extract common structural and physicochemical features from data mining of a diverse range of TACE inhibitors to facilitate the discovery of common pharmacophore.

Quantitative structure–activity relationship (QSAR) has been successfully applied in the prediction of TACE inhibitors. Initial leads of TACE inhibitors were derived from broad spectrum MMP inhibitors because of high sequence similarity in the active site regions of TACE and MMPs, and among the various categories of MMP and TACE inhibitors, the hydroxamic-acid based inhibitors have been most widely studied. Thus, most of the available two-dimensional QSARs on MMP and TACE inhibitors are related to only this class of inhibitors [42]. But very few two-dimensional QSAR studies so far have been directed toward other specific class of TACE inhibitors. In order to gain further insight into the structure requirements of small molecules possessing TACE inhibitory activities, three-dimensional quantitative structure–activity relationship (3D-QSAR) studies have also been performed. For instance, Murumkar et al. [43] selected a set of 29 benzothiadiazepine hydroxamates possessing TACE inhibitory activities, which were further divided into a training set (23 compounds) and a test set (6 compounds) to derive 3D-QSAR, comparative molecular field analysis (CoMFA), and comparative molecular similarity indices (CoMSIA) models for the atom-based, centroid/atom-based, data-based, and docked conformer-based alignments. The data-based alignment provided the optimal predictive CoMFA model for the training set with cross-validated $r^2 = 0.510$, non-cross-validated $r^2 = 0.972$, standard error of estimates = 0.098, and $F = 215.44$ and the optimal CoMSIA model with cross-validated $r^2 = 0.556$, non-cross-validated $r^2 = 0.946$, standard error of estimates = 0.163, and $F = 99.785$. The test set predictions of these two models for six compounds with predictive r^2 values were 0.460 and 0.535, respectively. Likewise, the investigation of 3D-QSAR CoMFA and CoMSIA for beta-amino hydroxamic acid-derived TACE inhibitors has also been studied by Murumkar et al. [44]. These present 3D-QSAR studies have successfully investigated the indispensable structural features of different chemical class of molecules which can be exploited for the structural modifications of these lead molecules to achieve improved TACE inhibitory activity. However, most of the prediction models have been developed and tested by using no more than ~60 compounds that are significantly less than the 443 known TACE inhibitors in our study, which included not only different chemical classes of hydroxamic-acid based TACE inhibitors but also highly selective non-hydroxamate TACE inhibitors. Moreover, to the best of our knowledge, our work is the first attempt for the classification prediction of TACE inhibitors on more diverse data set (1202 compounds).

On the other hand, support vector machine (SVM) and other machine learning approaches (MLA) have been consistently shown to have excellent performance for predicting various pharmacodynamic, pharmacokinetic and toxicological properties of compounds of diverse structures [45–48]. Therefore, in this study, it is of interest to test the usefulness and performance of SVM and other

MLA as potential tools for the prediction of TACE inhibitors. We trained and tested a SVM model by using the known TACE inhibitors from the literatures and non-inhibitors from those well-studied clinical drugs in MDL drug data report (MDDR) database. The prediction accuracies of this model were evaluated by two well-established testing methods, 5-fold cross-validation and independent evaluation set. Moreover, we compared the prediction accuracies of SVM model with those derived from three other ML methods, *k*-nearest neighbor (*k*-NN), C4.5 decision tree and back-propagation neural network (BPNN) by using the same dataset, molecular descriptors, and external evaluation set, in order to objectively examine to what extent SVM is effective for the prediction of TACE inhibitors. A feature selection method, recursive feature elimination (RFE), was used in this study to select a set of the most relevant and appropriate descriptors from a pool of 189 descriptors for the prediction of TACE inhibitors.

2. Materials and methods

2.1. Data sets

A total of 443 TACE inhibitors used in this work were collected from recently published papers [15,18–21,23–27,29–41]. The majority of the tested inhibitors are efficient TACE inhibiting agents showing IC_{50} value from 0.000066 μM to 100 μM. Since there are few non-inhibitors reported in the literature, we collected non-inhibitors from some well-studied clinical drugs in MDL drug data report (MDDR) database, in order to sufficiently represent the vastness of TACE non-inhibitors, 759 putative non-inhibitors were extracted from MDDR database by means of *k*-means clustering [49]. We divided more than 150,000 compounds from MDDR (removing those entries that have invalid structures or molecular descriptors) which have not been reported to have TACE binding activity into 759 clusters based on 189 calculated descriptors (vide infra). The scale of generated compound clusters in this work is consistent with that reported in other studies [50,51]. For each cluster, the compound closest to the centroid of the corresponding cluster was selected. The 2D structure of each compound was drawn by ChemDraw [52] and subsequently converted into 3D structure by Corina [53], and then followed by optimization using AM1 method. The resulted 3D structure of each compound was manually checked to ensure that the chirality of the chiral agent is properly generated and no structure of compounds was duplicated. These 1202 compounds have been assigned as either having or not having TACE inhibiting activity without the specific quantitative activity. We further separated them into the training set (198 inhibitors and 333 non-inhibitors), testing set (126 inhibitors and 222 non-inhibitors), and external validation set (119 inhibitors and 204 non-inhibitors) based on their similarity and distribution in the chemical space. The chemical space is defined by the commonly used structural and chemical descriptors [54], and each compound was distributed in a special position of the chemical space.

2.2. Validation methods

Two methods, 5-fold cross-validation and independent validation methods were used for the evaluation of the prediction accuracies of our ML models. In the 5-fold cross-validation method, we aimed at selecting the most important subset of descriptors for the classification by recursive feature elimination (RFE) method and testing the prediction capability of support vector machine (SVM) model, which was constructed by using this most important descriptor subset. The number of 879 compounds in the training and testing sets above were randomly divided into five subsets of approximately equal size. Four of the subsets were used to train the SVM model, and the remaining subset was used for the prediction. This process was repeated five times in order that each subset can be used for the prediction once. The external validation set did not involve in the selection of descriptors of 5-fold cross-validation, this made the selected important descriptor subset independent from the external validation set. In the independent validation method, the training set was used for training the ML models, the testing set was used for optimizing the parameters of these ML models, the external validation set was used for evaluating the prediction accuracies of the resulting ML models.

2.3. Molecular descriptors

Molecular descriptors have been routinely used for quantitative description of structural and physicochemical features of molecules in QSAR and MLAs [45–48,55]. In this work, a total of 189 molecular descriptors were used, this set of descriptors was manually selected from more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or irrelevant to the prediction of pharmaceutical agents [56]. These descriptors, given in Table 1, include 18 descriptors in the class of simple molecular properties (such as molecular weight and number of rotatable bonds), 27 descriptors in the class of molecular connectivity and shape (such as molecular connectivity indices and molecular kappa shape indices), 97 descriptors in the class of electro-topological state (such as electro-topological state indices), 22 descriptors in the class of quantum chemical properties (such as atomic charges and molecular dipole moment), and 25 descriptors in the class of geometrical properties (such as solvent accessible surface area and hydrophobic region). The 189 descriptors were computed from the optimized 3D structure of each compound using our own designed molecular descriptor computing program, but not all of these descriptors are essential for the classification of TACE inhibitor and non-inhibitor. The remaining redundant and unimportant descriptors in these 189 descriptors were further eliminated by RFE method (details in the latter subsection).

2.4. Machine learning and feature selection methods

2.4.1. Support vector machine (SVM)

The theory of SVM has been extensively described in the literature [57]. Thus, only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from machine learning theory. In linearly separable cases, SVM constructs a hyperplane which separates two different classes of vectors with a maximum margin. With regard to nonlinearly separable problems, SVM projects the input feature vectors into a high-dimensional feature space and searches for a linear optimal separating hyperplane (decision boundary) in the new feature space by using a kernel function. We constructed the SVM models by using the Gaussian radial basis kernel function which has been extensively introduced in different studies with good performances [58,59]. The Gaussian kernel can be represented by

$$K(x_i, x_j) = \exp(-||x_j - x_i||^2/2\sigma^2) \tag{1}$$

Then a linear SVM is applied to this high-dimensional feature space and the classification function is given by

$$f(x) = sign(\sum \alpha_i y_i K(x, x_i) + b) \tag{2}$$

where α_i is the Langrangian multiplier. A positive or negative value of $f(x)$ denotes that the vector x belongs to the positive or negative class, respectively.

Table 1
Molecular descriptors used in this work.

Descriptor class	Number of descriptor in class	Descriptors
Simple molecular properties	18	Molecular weight, numbers of rings, rotatable bonds, H-bond donors, and H-bond acceptors, element counts
Molecular connectivity and shape	27	Molecular connectivity indices, valence molecular connectivity indices, molecular shape Kappa indices, Kappa alpha indices, flexibility index
Electrotopological state	97	Electrotopological state indices, and atom type electrotopological state indices, Weiner index, centric index, Altenburg index, Balaban index, Harary number, Schultz index, PetitJohn R2 index, PetitJohn D2 index, mean distance index, PetitJohn I2 index, information Weiner, Balaban rmsd index, graph distance index
Quantum chemical properties	22	Polarizability index, hydrogen bond acceptor basicity (covalent HBAB), hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index, most positive charge on H, C, N, O atoms, most negative charge on H, C, N, O atoms, most positive and negative charge in a molecule, LSum of squares of charges on H, C, N, O and all atoms, mean of positive charges, mean of negative charges, mean absolute charge, relative positive charge, relative negative charge
Geometrical properties	25	Length vectors (longest distance, longest third atom, 4th atom), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively charged atoms, molecular rugosity, molecular globularity, hydrophilic region, hydrophobic region, capacity factor, hydrophilic–hydrophobic balance, hydrophilic intery moment, hydrophobic intery moment, amphiphilic moment

2.4.2. *Feature selection method*

Feature selection methods have been introduced for the selection of features meaningful for discriminating two categorical datasets by removing redundant and irrelevant molecular descriptors. Recently, the recursive feature elimination (RFE) method has been gain great popularity for feature selection in many biochemical classification fields due to its effectiveness for selection of molecular descriptors [45,60–62]. Thus, the RFE method was used in this work to extract the most important subset of features relevant to the classification of TACE inhibitors and improve the prediction accuracies of the ML models by using this descriptor subset. The details of the implementation of this method can be found in the literatures [47,56].

In our study, we developed a SVM model combined with RFE to select the most relevant descriptors for classification by 5-fold cross-validation method. This SVM model is trained by using a Gaussian kernel function with an adjustable parameter σ. We need to find a specific value of σ that gives the best prediction accuracy from sequential variation of σ. The feature selection procedure can be demonstrated by the following steps: (1) for a fixed σ, the SVM model is trained by using the complete set of features (189 descriptors). (2) Computing the ranking criterion score DJ(i) [47,56] for each feature in the current set. All of the computed DJ(i) is subsequently ranked in descending order. (3) Abandon of the m features with smallest criterion scores, $m = 5$ is used in our work. (4) The SVM classification model is retrained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of 5-fold cross-validation. Then the first to fourth steps are repeated for other values of σ. These procedures continue, until the set of features and parameter σ that give the best prediction accuracy are selected.

2.4.3. *k-Nearest neighbor (k-NN)*

In k-NN, the Euclidean distance between an unclassified vector $\mathbf{x}$ and each individual vector $\mathbf{x}_i$ in the training set is calculated by using the following formula:

$$D = \sqrt{||\boldsymbol{x} - \boldsymbol{x}_i||^2} \tag{3}$$

A total of k number of vectors nearest to the vector $\mathbf{x}$ are used to determine the class of that unclassified vector. The class of the majority of the k-nearest neighbors is decided as the predicted class of the unclassified vector $\mathbf{x}$.

2.4.4. *C4.5 decision tree (DT)*

C4.5 DT is a branch-test-based classifier [63]. A branch of a decision tree corresponds to a group of classes and a leaf represents a specific class. A decision node specifies a test to be conducted on a single attribute value, with one branch and its subsequent classes as possible outcomes. C4.5 DT uses recursive partitioning to examine every attribute of the data and rank them according to their ability to partition the remaining data, thereby constructing a decision tree. A vector $\mathbf{x}$ is classified by starting at the root of the tree and moving through the tree until a leaf is encountered. At each non-leaf decision node, a test is conducted to move into a branch. Upon reaching the destination leaf, the class of the vector $\mathbf{x}$ is predicted to be that associated with the leaf.

2.4.5. *Back-propagation neural network (BPNN)*

The BPNN is a generalization of the delta rule used for training multi-layer feed-forward neural networks with non-linear units. It is simply a gradient descent method design to minimize the total error (or mean error) of the output computed by the network. In the network, there is an input layer, an output layer, with one hidden layers in between them.

The network functions as follows: each node i in the input layer has a signal x_i as network's input, multiplied by a weight value between the input layer and the hidden layer. Each node j in the hidden layer receives the signal $In(j)$ according to

Eq. (4)

$$In(j) = b_1 + \sum_{i=1}^{n} x_i w_{ij} \tag{4}$$

Then passed through the sigmoid activation function

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

The output of the activation function $f(In(j))$ is then broadcast all of the neurons to the output layer

$$y_k = b_2 + \sum_{j=1}^{m} w_{jk} f(In(j)) \tag{6}$$

where b_1 and b_2 are the biases in the hidden layer and the output layer. The output value will be compared with the target; in this paper, we used the mean square error (*MSE*) as the error function:

$$\mathrm{MSE} = \frac{\sum_k Y_k - O_k}{N \times K} \tag{7}$$

where N is the number of training patterns, K is the number of nodes in output layer, Y_k and O_k are the output value and the target value, respectively. The gradient descent method combined with a momentum term was used to search for the global optimum of the network weights, and partial derivatives $\partial E(t)/\partial w(t)$ are computed for each weight in the network. The weight will adjust according to the following expression:

$$\Delta w(t+1) = \eta \frac{\partial E(t)}{\partial w(t)} + \alpha \Delta w(t) \tag{8}$$

where t is the number of epochs, η is the learning rate, $\alpha \Delta w(t)$ is the momentum term, $\alpha (\alpha \in (0, 1))$ is referred to the momentum coefficient.

2.5. Performance evaluation

As in the case of all classification methods [64], the performance of ML methods can be assessed by the quantity of true positive samples (*TP*), true negative samples (*TN*), negative samples predicted as false positive (*FP*) and positive samples predicted as false negatives (*FN*). Several criteria for evaluating prediction performance include sensitivity *SE* (prediction accuracy for inhibitors), specificity *SP* (prediction accuracy for non-inhibitors). The overall prediction accuracy (*Q*) and Matthews correlation coefficient (*C*) are also used to evaluate the prediction accuracies which are given below:

$$SE = TP/(TP + FN) \tag{9}$$

$$SP = TN/(TN + FP) \tag{10}$$

$$Q = \frac{TP + TN}{TP + FN + TN + FP} \tag{11}$$

$$C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{12}$$

3. Results and discussions

3.1. Structural diversity of compounds

Structural diversity of the whole dataset can be estimated by the *DI* value, which is the average value of the dissimilarity between pairs of compounds in the dataset [65]

$$DI = \frac{\sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} diss(i, j)}{N(N-1)} \tag{13}$$

where $diss(i, j)$ is a measure of dissimilarity between compound i and j and N is the number of compounds in the dataset. The structural diversity of a dataset increases with increasing *DI* value. The measure of dissimilarity $diss(i, j)$ is computed by using the Tanimoto coefficient [66] in this study:

$$diss(i, j) = 1 - \frac{\sum_{d=1}^{l} X_{di} Y_{dj}}{\sum_{d=1}^{l} (X_{di})^2 + \sum_{d=1}^{l} (X_{dj})^2 - \sum_{d=1}^{l} X_{di} X_{dj}} \tag{14}$$

where l is the number of descriptors computed for the compounds in the dataset, X_{di} and X_{dj} are the values of dth descriptor for compounds i and j, respectively. The value of *DI* ranges from 0.0 to 1.0. The value of 1.0 for *DI* denotes that the dataset is sufficiently diverse for the given molecular descriptors, the contrary indicates that all the compounds are fully identical. Obviously, the closer the value of *DI* is to 1.0, the more diverse is the dataset. In this work, the computed value of *DI* for the dataset of 443 TACE inhibitors and 759 non-inhibitors is 0.428, which indicates that the structures of the investigated compounds are sufficiently diverse.

Table 2
The accuracy of TACE inhibitors and non-inhibitors derived from SVM without the use of a feature selection method (SVM) and from SVM with the use of the feature selection method RFE (SVM + RFE) by using 5-fold cross-validation method[a].

Method	Cross-validation	TACE inhibitors			TACE non-inhibitors			*Q* (%)	*C*
		TP	*FN*	Accuracy *SE* (%)	*TN*	*FP*	Accuracy *SP* (%)		
SVM	1	63	2	96.92	102	1	99.03	98.2	0.962
	2	64	0	100.00	118	2	98.33	98.9	0.976
	3	57	3	95.00	116	1	99.14	97.7	0.949
	4	64	0	100.00	116	2	98.30	94.7	0.976
	5	70	1	98.59	95	2	97.93	98.9	0.963
	Average			98.10			98.55	98.4	0.966
	S.D.			2.1			0.5	0.5	0.011
SVM + RFE	1	63	2	96.92	103	0	100.00	98.8	0.975
	2	64	0	100.00	120	1	100.00	100.0	1.000
	3	59	1	98.33	116	1	99.15	98.9	0.975
	4	63	1	98.44	117	1	99.15	98.9	0.976
	5	70	1	98.59	94	3	96.91	97.6	0.952
	Average			98.45			99.04	98.8	0.976
	S.D.			1.1			1.3	0.8	0.017

[a] The results are given in *TP* (true positive), *FN* (false negative), *TN* (true negative), *FP* (false positive), *Q* (overall accuracy), and *C* (Matthews correlation coefficient). *SE* (sensitivity) is the prediction accuracy for TACE inhibitors and *SP* (specificity) is the prediction accuracy for non-inhibitors. Statistical significance is indicated by S.D. (standard deviation). The number of TACE inhibitors or non-inhibitors is *TP* + *FN* or *TN* + *FP*.

Table 3
The comparison of the prediction accuracies of TACE inhibitor and non-inhibitors from different statistical learning approaches by using an external validation set with all 189 descriptors and the 23 RFE-selected descriptors[a].

Method	TACE inhibitor accuracy (%)		Non-inhibitor accuracy (%)		Overall accuracy (%)		MCC	
	ML	ML + RFE	ML	ML + RFE	ML	ML + RFE	ML	ML + RFE
SVM	95.80	96.64	97.06	99.51	96.60	98.45	0.927	0.967
BPNN	96.64	94.96	98.04	99.02	97.52	97.52	0.947	0.947
k-NN	93.28	98.32	98.53	98.53	96.59	98.45	0.927	0.967
C4.5 DT	99.16	98.32	96.57	97.55	97.52	97.83	0.948	0.954

[a] The different machine learning (ML) methods include C4.5 decision tree (C4.5), *k*-nearest neighbor (*k*-NN), back-propagation neural network (BPNN), and support vector machine (SVM). The accuracy of each method was estimated from an external independent validation set by using all the descriptors and the RFE-selected descriptors, respectively. ML denotes that all the descriptors were involved in the model, whereas ML + RFE denotes that the RFE-selected descriptors were used in the model.

3.2. Effect of feature selection (RFE method) on overall prediction accuracies

The prediction accuracies of the SVM models for TACE inhibitors and non-inhibitors were primarily assessed by means of 5-fold cross-validation using the RFE method (termed as SVM + RFE) and without using the RFE method (termed as SVM). The SVM model gives the best performance for the prediction of TACE inhibitors and non-inhibitors when the parameter σ of SVM was chosen to be 4.0. The average prediction accuracies of SVM without RFE for TACE inhibitors and non-inhibitors are 98.10% and 98.55%, and those by using SVM + RFE are 98.45% and 99.04%, respectively. The detailed results are given in Table 2. (The prediction accuracies from SVM + RFE are slightly improved compared to those derived from SVM without RFE, which indicates the usefulness of RFE in selecting the most important features to facilitate the classification for TACE inhibitors and non-inhibitors.) The performance of SVM approaches without feature selection by RFE have already achieved great prediction performance, this may indicate that these 189 descriptors could be competent to characterize most of the molecular features of TACE inhibitors and non-inhibitors, and there may not be enough space to significantly improve the prediction performance by RFE method. Although the overall prediction accuracy (Q) and Matthews correlation coefficient (C) of SVM model using the RFE method is at a similar level as those of SVM model using all the 189 descriptors (98.4% vs. 98.8% for overall prediction accuracy (Q) and 0.966 vs. 0.976 for Matthews correlation coefficient (C)), the redundancy of SVM model and irrelevancy of molecular descriptors were enormously declined as the number of molecular descriptors used in SVM models was reduced from 189 descriptors to 23 ones. Moreover, in the independent validation method, the prediction accuracies of all four machine learning models, SVM, BPNN, *k*-NN, and C4.5 DT were further improved by using these RFE selected descriptors. In conclusion, the RFE method is useful in selecting the most important features and removing irrelevant molecular descriptors to facilitate the classification for TACE inhibitors and non-inhibitors.

Table 3 gives the prediction accuracies of TACE inhibitors and non-inhibitors derived from all four machine learning methods SVM, BPNN, *k*-NN, and C4.5 DT by using RFE selected descriptors and independent validation method. For comparison, those by using all the 189 descriptors are also included in Table 3. The prediction accuracies of all four ML models for non-inhibitors are in the range of 96.57–98.53% by using all the 189 descriptors, and those are in the range of 97.55–99.51% by using the RFE-selected descriptors. The use of RFE remarkably improved the prediction accuracies of non-inhibitors. The prediction accuracies of TACE inhibitors range from 93.28% to 99.16% by using all the 189 descriptors, and those range from 94.96% to 98.32% by using the RFE-selected descriptors. The RFE method significant enhances the prediction accuracies of *k*-NN for TACE inhibitors, the prediction accuracies of SVM for non-inhibitors are also slightly improved by using this RFE selected descriptors. The analogical results can be obtained with respect to the overall prediction accuracy and Matthews correlation coefficient. Our study suggests that RFE is efficient for the elimination of redundant descriptors and the improvement of prediction performance.

3.3. Performances of machine learning methods

Table 4 gives the detailed comparison of the prediction accuracies of TACE inhibitors and non-inhibitors from all four ML methods (SVM, BPNN, *k*-NN, C4.5 DT) with the RFE selected descriptors. The parameters of *k*-NN, BPNN and SVM, which bear the best prediction performance, are also given in Table 4 and the structure of the pruned C4.5 decision tree is shown in Fig. 1. For TACE inhibitors, as shown in Table 4, the accuracies of these methods are in the range of 96.64–98.32% with *k*-NN and C4.5 DT giving the best accuracies at 98.32%. For non-inhibitors, the accuracies from these methods are in the range of 97.55–99.51% with SVM giving the best accuracy at 99.51%. Both *k*-NN and SVM model gives the best overall prediction accuracy.

There are four TACE inhibitors and one non-inhibitor misclassified by SVM model, which listed in Fig. 2. In addition, it should be noted that there are essential differences in constructing classification models with the different ML algorithms which can be seen from the training set and the testing set. Also, due to a very small deviations between prediction accuracies of the training and

Table 4
The comparison of the prediction accuracies of TACE inhibitors and non-inhibitors from different statistical learning approaches with the RFE-selected descriptors by using an external validation set[a].

Method	Parameter	Training set		Testing set				External validation set					
		Compounds	Errors (%)	TACE inhibitors		Non-inhibitors		TACE inhibitors			Non-inhibitors		
				TP	*FN*	*TN*	*FP*	*TP*	*FN*	*SE* (%)	*TN*	*FP*	*SP* (%)
SVM	$\sigma = 0.06$	531	0.20%	126	0	212	10	117	2	98.32	199	5	97.55
BPNN	$h = 10$	531	1.69%	124	2	218	4	115	4	96.64	200	4	98.04
k-NN	$k = 2pr$	531	–	125	1	218	4	117	2	98.32	201	3	98.53
C4.5 DT		531	0.00%	123	3	221	1	115	4	96.64	203	1	99.51

[a] C4.5 DT (C4.5 decision tree), *k*-NN (*k*-nearest neighbor), BPNN (back-propagation neural network), SVM + RFE (support vector machine and recursive feature elimination).

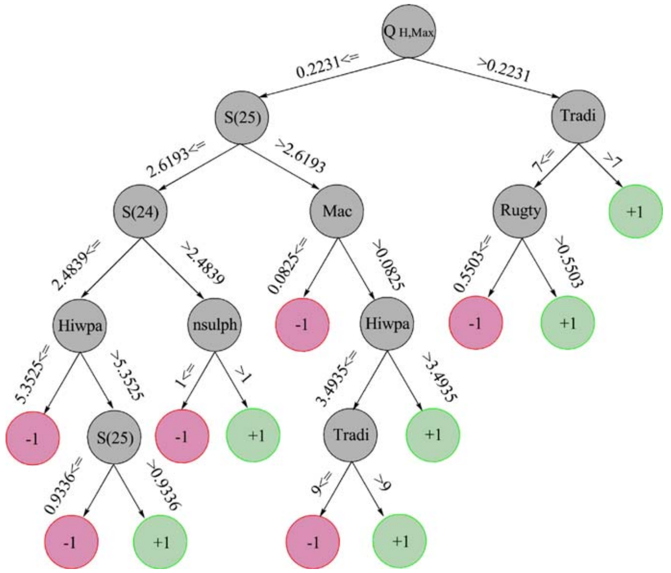
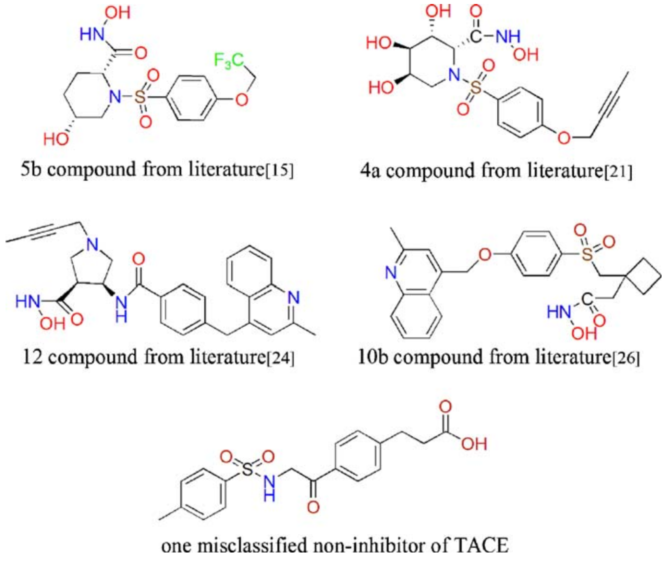


Fig. 1. The structure of the pruned C4.5 decision tree for the prediction of TACE inhibitors and non-inhibitors. Each node in the decision tree shows a special split attribute, the corresponding branch indicates the split value of attribute, and the notation "+1" or "−1" of a green or red leaf denotes an TACE inhibitor or non-inhibitor predicted, respectively.


5b compound from literature[15]

4a compound from literature[21]

12 compound from literature[24]

10b compound from literature[26]

one misclassified non-inhibitor of TACE

Fig. 2. The structures of misclassified four TACE inhibitors and one non-inhibitor by SVM in the independent validation method.

the external validation set, it can be concluded that these classification models are robust and reliable.

3.4. Relevance of selected molecular descriptors for predicting TACE inhibitors

A total of 23 molecular descriptors are selected by RFE as the most relevant descriptors for distinguishing between TACE inhibitors and non-inhibitors, which was given in Table 5. Some of these selected descriptors are correspondingly matched or partially matched with molecular features from other QSAR studies.

Despite a low sequence homology and divergent structural elements, the TACE structure has excellent similarity with MMP structures and more importantly the active site of TACE is reminiscent of the MMPs. Similar to MMPs, the requirements for a molecule to be an effective TACE are (1) the presence of a functional group, such as a hydroxamic group (–CONHOH), hydantoin group and sulfhydryl group (–SH), that may be able to chelate the active site Zn^{2+} ion of the enzyme (such a group is referred to as a zinc binging group, ZBG); and (2) several functional groups capable of hydrogen bonding with the enzyme backbone. Of the 23 RFE selected descriptors, *Mac* (mean absolute charges on O, N, and S atoms) and $Q_{O,max}$ (most positive charge on O atoms) were used to measure the strength of chelation between the active site Zn^{2+} and ZBG groups. *Mpc* (mean of positive charges) and $Q_{H,Max}$ (mean positive charge on H atoms) were used to represent the capability of hydrogen bonding. A specific polar functional group *S*(2) (Atom-type H Estate sum for = N**H)** describes a type of hydrogen bond donor. One simple molecular descriptor *nsulph*

Table 5
Molecular descriptors selected from the RFE feature selection method for classification of TACE inhibitors and non-inhibitors.

Molecular descriptor	Description	Matched or partially matched molecular descriptors used in previous QSAR models
Nsulph	Count of S atoms	Ss [42]
$3\chi c$	Simple molecular connectivity Chi indices for cluster	
$6\chi CH$	Simple molecular connectivity Chi indices for cycle of 6	$I_{1,pyr}$ [42], $I_{R,Ph}$ [42]
$4\chi vCH$	Valence molecular connectivity Chi indices for cycle of 4	
S(24)	Atom-type Estate sum for ≡C–	
S(13)	Atom-type H Estate sum for CH *n* (unsaturated)	Electrostatic field [43,44]
S(20)	Atom-type Estate sum for =CH–	$I_{R2\text{-}CC}$ [42]
Tradi	PetitJohn R2 index	
S(21)	Atom-type Estate sum for: CH: (aromatic)	$I_{R,Ph}$ [42], hydrophobic field [43,44]
S(2)	Atom-type H Estate sum for =NH	$I_{1,\ NH}$ [42], hydrogen bond donor potential field [43,44]
S(10)	Atom-type H Estate sum for:CH: (sp2, aromatic)	Electrostatic field [43,44]
S(34)	Atom-type Estate sum for =N–	SN [42], electrostatic field [43,44], hydrogen bond acceptor potential field [43,44]
S(25)	Atom-type Estate sum for =C<	$I_{R2\text{-}CC}$ [42]
Mac	Mean absolute charge	Electrostatic field [43,44]
$Q_{O,Max}$	Most positive charge on O atoms	Electrostatic field [43,44]
μ	Molecular dipole moment	
Mpc	Mean of positive charges	Electrostatic field [43,44]
$Q_{H,Max}$	Most positive charge on H atoms	Electrostatic field [43,44]
Hiwpa	Amphiphilic moment	Clog*P* [42], hydrophobic field [43,44]
Rugty	Molecular rugosity	Steric field [43,44]
Capty	Capacity factor	Electrostatic field [43,44]
Hiwpb	Hydrophobic intery moment	Clog*P* [42], hydrophobic field [43,44]
Hiwpl	Hydrophilic intery moment	ClogP [42], hydrophobic field [43,44]

(Count of S atoms) is also selected, which may associate with non-hydroxamic ZBG group (–SH) and hydrogen bond acceptor (–SO_2–).

However, TACE has a very interesting feature of its active sites, wherein S1′ and S3′ subsites are merged to create an L-shaped S1′ binding cleft that opens up into the S3′ pocket. Both S1′ and S3′ are hydrophobic but are connected by a polar entrance. Thus, the characteristics of the TACE binding pocket are unique relative to MMPs. The acetylenic P1′ group of TACE inhibitors could fit in this L-shape S1′ binding cleft and confer TACE selectivity over MMPs. Some specific functional groups *S*(24) (Atom-type Estate sum for ≡C–), *S*(20) (Atom-type Estate sum for =CH–), *S*(25) (Atom-type Estate sum for =C<), *S*(13) (Atom-type H Estate sum for $C\mathbf{H}_n$ (unsaturated)) were selected to state this acetylene-derived linear substituent in the tunnel between S1′ and S3′ hydrophobic pockets of the receptor.

The quantitative structure–activity relationship (QSAR) study on zinc-containing metalloproteinase inhibitors [42] has shown that in most of the cases the activity of TACE inhibition is dependent only on the hydrophobic property. But in some cases there is a parabolic correlation between the TACE inhibition and Clog *P*, and in other cases it is only linear with negative coefficient of Clog *P*. In vitro case, the parabolic dependence of activity on hydrophobicity is assumed to be due to the limited bulk tolerance of the receptor sites, as there are no hydrophilic–hydrophobic cell membrane-like intervening barriers. The bulky substituents could produce the steric effects in TACE inhibition. So the molecular volume itself sometimes can account for the variation in the activity when it occurs with the negative coefficient in the correlations. Some 3D-QSAR studies [43,44] also suggested that the biological activity of TACE inhibitors have important contributions from steric and electrostatic interactions. Of the 23 selected descriptors, *Rugty* (molecular rugosity) is selected to represent the steric effect of a molecule which measures the ratio between the bare molecular surface area and molecular volume. *Hiwpa* (amphiphilic moment), *Hiwpb* (hydrophobic intergy moment), and *Hiwpl* (hydrophilic intergy moment) are the RFE-selected descriptors correlated to Clog *P* and other hydrophobic descriptors used in previous 3D-QSAR studies. Besides *Mac*, *Mpc*, $Q_{O,Max}$, and $Q_{H,Max}$, the descriptor *Capty* associated with concentration of polar interactions on molecular surface is also selected to characterize the electrostatic interactions between TACE and its inhibitors.

Not all bulky substituents produce the adverse effect in inhibiting activity of TACE, in some cases, the bulky substituents could also have the favorable effect for TACE inhibition. For instance, QSAR study on series of amino acid hydroxamates [42] showed that a pyridyl ring substituent would be advantageous for TACE inhibitions. Molecular modeling studies of homology model and X-ray crystal structure of TACE complexes also suggested that certain quinolines when properly placed could fit into the curvature of TACE S1′ pocket while clashing with the linear MMP S1′ pocket. Such substituents, which are planar and contain a nitrogen with a lone pair of electrons, may lead to a stronger electrostatic interactions of the compounds with the receptor.

${}^3\chi_C$ (simple molecular connectivity Chi indices for cluster), ${}^6\chi_{CH}$ (simple molecular connectivity Chi indices for cycles of 6), and ${}^4\chi^v{}_{CH}$ (valence molecular connectivity Chi indices for cycles of 4) describe simple and valence molecular connectivity for a cluster or cycle of atoms, which can be used to describe ring. *S*(34) (Atom-type Estate sum for =N–) is selected to represent lone pair of electrons at the nitrogen atom. Moreover, μ (molecular dipole moment), *Tradi* (PetitJohn R2 Index), *S*(21) (Atom-type Estate sum for: CH: (aromatic)), and *S*(10) (Atom-type H Estate sum for :C**H**: (sp^2, aromatic)) are also selected. Overall, these 23 descriptors seem to be capable of competently describe most of the molecular features of TACE inhibitors.

4. Conclusion

This study shows that ML approaches, particularly SVM, are potentially useful for the prediction of TACE inhibitors from a diverse set of compounds. Feature selection method, recursive feature elimination, has been found to be useful for improving the performance and selecting of informative features for distinguishing TACE inhibitors and non-inhibitors by removing redundant and irrelevant molecular descriptors. In addition, the analysis of these RFE-selected descriptors can provide useful clues to the structural and physicochemical features contributing to a specific property, this may facilitate the discovery of common pharmacophore for TACE inhibitors.

Acknowledgments

This project has been supported by the National Natural Science Foundation of China (Grant No. 20773089) and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (Grant No. 20071118-18-15).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2009.08.001.

References

[1] B. Aggarwal, W. Kohr, P. Hass, B. Moffat, S. Spencer, W. Henzel, T. Bringman, G. Nedwin, D. Goeddel, R. Harkins, Human tumor necrosis factor production, purification, and characterization, J. Biol. Chem. 260 (1985) 2345–2354.

[2] M.H. Bemelmans, L.J. van Tits, W.A. Buurman, Tumor necrosis factor: function, release and clearance, Crit. Rev. Immunol. 16 (1996) 1–11.

[3] B.B. Aggarwal, K. Natarajan, Tumor necrosis factors: developments during the last decade, Euro. Cytokine. Net. 7 (1996) 93–124.

[4] R.A. Black, C.T. Rauch, C.J. Kozlosky, J.J. Peschon, J.L. Slack, M.F. Wolfson, B.J. Castner, K.L. Stocking, P. Reddy, S. Srinivasan, N. Nelson, N. Bolani, K.A. Schooley, M. Gerhart, R. Devis, J.N. Fitzner, R.S. Johnson, R.J. Paxton, C.J. March, D.P. Cerretti, A metalloproteinase disintegrin that release tumor necrosis factor-α from cells, Nature 385 (1997) 729–733.

[5] L. Killar, J. White, R. Black, J. Peschon, Adamalysins. A family of metzincins including TNF-α converting enzyme (TACE), Ann. N.Y. Acad. Sci. 878 (1999) 442–452.

[6] L.W. Moreland, S.W. Baumgartner, M.H. Schiff, E.A. Tindall, R.M. Fleischmann, A.L. Weaver, R.E. Ettlinger, S. Cohen, W.J. Koopman, K. Mohler, M.B. Widmer, C.M. Blosch, Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein, N. Engl. J. Med. 337 (1997) 141–147.

[7] B.A. Beutler, The role of tumor necrosis factor in health and disease, J. Rheumatol. 26 (1999) 16–21.

[8] J.M. Clements, J.A. Cossins, G.M. Wells, D.J. Corkill, K. Helfrich, L.M. Wood, R. Pigott, G. Stabler, G.A. Ward, A.J. Gearing, K.M. Miller, Matrix metalloproteinase expression during experimental autoimmune encephalomyelitis and effects of a combined matrix metalloproteinase and tumor necrosis factor-α inhibitor, J. Neuroimmunol. 74 (1997) 85–94.

[9] F.R. Cochran, M.P. Vitek, Neuroinflammatory mechanisms in Alzheimer's disease: new opportunities for drug discovery, Exp. Opin. Invest. Drugs 5 (1996) 449–455.

[10] R. Reiss, D. Makoff, H. Rodriguez, S. Graham, J. Mittleman, S. Jordan, Encephalopathy and cerebral infarction in OKT3-treated patients with concomitant elevation of cerebrospinal fluid tumour necrosis factor-α, Nephrol. Dial. Transpl. 8 (1993) 464–468.

[11] M. Feldmann, R.N. Mani, Anti-TNFα therapy of rheumatoid arthritis: what have we learned? Annu. Rev. Immuunol. 19 (2001) 163–196.

[12] M. Matsumoto, S. Mariathasan, M. Nahm, F. Baranyay, J. Peschon, D. Chaplin, Role of lymphotoxin and the type I TNF receptor in the formation of germinal centers, Science 271 (1996) 1289–1291.

[13] N. Shakoor, M. Michalska, C.A. Harris, J.A. Block, Drug-induced systemic lupus erythematosus associated with etanercept therapy, Lancet 359 (2002) 579–580.

[14] K. Maskos, C. Fernandez-Catalan, R. Huber, G.P. Bourenkov, H. Bartunik, G.A. Ellestad, P. Reddy, M.F. Wolfson, C.T. Rauch, B.J. Castner, R. Davis, H.R.G. Clarke, M. Petersen, J.N. Fitzner, D.P. Cerretti, C.J. March, R.J. Paxton, R.A. Black, W. Bode, Crystal structure of the catalytic domain of human tumor necrosis factor-α converting enzyme, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 3408–3412.

[15] M.A. Letavic, M.Z. Axt, J.T. Barberia, T.J. Carty, D.E. Danley, K.F. Geoghegan, N.S. Halim, L.R. Hoth, A.V. Kamath, E.R. Laird, L.L. Lopresti-Morrow, K.F. McClure, P.G. Mitchell, V. Natarajan, M.C. Noe, J. Pandit, L. Reeves, G.K. Schulte, S.L. Snow, F.J. Sweeney, D.H. Tan, C.H. Yu, Synthesis and biological activity of selective pipecolic acid-based TNF-α Converting Enzyme Inhibitors, Bioorg. Med. Chem. Lett. 12 (2002) 1387–1390.

[16] Z.R. Wasserman, J.J. Duan, M.E. Voss, C.B. Xue, R.J. Cherney, D.J. Nelson, K.D. Hardman, C.P. Decicco, Identification of a selectivity determinant for inhibition of tumor necrosis factor-α converting enzyme by comparative modeling, Chem. Biol. 10 (2003) 215–223.

[17] J.I. Levin, J.M. Chen, M.T. Du, F.C. Nelson, L.M. Killar, S. Skala, A. Sung, G. Jin, R. Cowling, D. Barone, C.J. March, K.M. Mohler, R.A. Black, J.S. Skotnicki, Anthranilate sulfonamide hydroxamate TACE inhibitors. Part 2. SAR of the acetylenic P1′ group, Bioorg. Med. Chem. Lett. 12 (2002) 1199–1202.

[18] K. Park, A. Aplasca, M.T. Du, L. Sun, Y. Zhu, Y. Zhang, J.I. Levin, Design and synthesis of butynyloxyphenyl β-sulfone piperidine hydroxamates as TACE inhibitors, Bioorg. Med. Chem. Lett. 16 (2006) 3927–3931.

[19] J.I. Levin, J.M. Chen, L.M. Laakso, M. Du, J. Schmid, W. Xu, T. Cummons, J. Xu, G. Jin, D. Barone, J.S. Skotnicki, Acetylenic TACE inhibitors. Part 3. Thiomorpholine sulfonamide hydroxamates, Bioorg. Med. Chem. Lett. 16 (2006) 1605–1609.

[20] J.I. Levin, J.M. Chen, L.M. Laakso, M. Du, X. Du, A.M. Venkatesan, V. Sandanayaka, A. Zask, J. Xu, W. Xu, Y. Zhang, J.S. Skotnicki, Acetylenic TACE inhibitors. Part 2: SAR of six-membered cyclic sulfonamide hydroxamates, Bioorg. Med. Chem. Lett. 15 (2005) 4345–4349.

[21] T. Tsukida, H. Moriyama, Y. Inoue, H. Kondo, K. Yoshino, S. Nishimura, Synthesis and biological activity of selective azasugar-based TACE inhibitors, Bioorg. Med. Chem. Lett. 14 (2004) 1569–1572.

[22] J.J. Duan, L. Chen, Z.R. Wasserman, Z. Lu, R.Q. Liu, M.B. Covington, M. Qian, K.D. Hardman, R.L. Magolda, R.C. Newton, D.D. Christ, R.R. Wexler, C.P. Decicco, Discovery of γ-lactam hydroxamic acids as selective inhibitors of tumor necrosis factor α converting enzyme: design, synthesis, and structure–activity relationships, J. Med. Chem. 45 (2002) 4954–4957.

[23] J.J. Duan, L. Chen, Z. Lu, C.B. Xue, R.Q. Liu, M.B. Covington, M. Qian, Z.R. Wasserman, K. Vaddi, D.D. Christ, J.M. Trzaskos, R.C. Newton, C.P. Decicco, Discovery of β-benzamido hydroxamic acids as potent, selective, and orally bioavailable TACE inhibitors, Bioorg. Med. Chem. Lett. 18 (2008) 241–246.

[24] X.T. Chen, B. Ghavimi, R.L. Corbett, C.B. Xue, R.Q. Liu, M.B. Covington, M. Qian, K.G. Vaddi, D.D. Christ, K.D. Hartman, M.D. Ribadeneira, J.M. Trzaskos, R.C. Newton, C.P. Decicco, J.J. Duan, A new 4-(2-methylquinolin-4-ylmethyl)phenyl P1′ group for the beta-amino hydroxamic acid derived TACE inhibitors, Bioorg. Med. Chem. Lett. 17 (2007) 1865–1870.

[25] R.J. Cherney, B.W. King, J.L. Gilmore, R.Q. Liu, M.B. Covington, J.J. Duan, C.P. Decicco, Conversion of potent MMP inhibitors into selective TACE inhibitors, Bioorg. Med. Chem. Lett. 16 (2006) 1028–1031.

[26] C.B. Xue, X.T. Chen, X. He, J. Roderick, R.L. Corbett, B. Ghavimi, R.Q. Liu, M.B. Covington, M. Qian, M.D. Ribadeneira, K. Vaddi, J. Trzaskos, R.C. Newton, J.J. Duan, C.P. Decicco, Synthesis and structure–activity relationship of a novel sulfone series of TNF-α converting enzyme inhibitors, Bioorg. Med. Chem. Lett. 14 (2004) 4453–4459.

[27] C.B. Xue, X. He, J. Roderick, R.L. Corbett, J.J. Duan, R.Q. Liu, M.B. Covington, M. Qian, M.D. Ribadeneira, K. Vaddi, D.D. Christ, R.C. Newton, J.M. Trzaskos, R.L. Magolda, R.R. Wexler, C.P. Decicco, Rational design, synthesis and structure–activity relationships of a cyclic succinate series of TNF-α converting enzyme inhibitors. Part 2. Lead Optimization, Bioorg. Med. Chem. Lett. 13 (2003) 4299–4304.

[28] J.F. Fisher, S. Mobashery, Recent advances in MMP inhibitor design, Cancer Metast. Rev. 25 (2006) 115–136.

[29] N. Kamei, T. Tanaka, K. Kawai, K. Miyawaki, A. Okuyama, Y. Murakami, Y. Arakawa, M. Haino, T. Harada, M. Shimano, Reverse hydroxamate-based selective TACE inhibitors, Bioorg. Med. Chem. Lett. 14 (2004) 2897–2900.

[30] U.K. Bandarage, T. Wang, J.H. Come, E. perola, Y. Wei, B.G. Rao, Novel thiol-based TACE inhibitors. Part 2. Rational design, synthesis and SAR of thiol-containing aryl sulfones, Bioorg. Med. Chem. Lett. 18 (2008) 44–48.

[31] B.G. Rao, U.K. Bandarage, T. Wang, J.H. Come, E. Perola, Y. Wei, S.K. Tian, J.O. Saunders, Novel thiol-based TACE inhibitors: Rational design, synthesis, and SAR of thiol-containing aryl sulfonamides, Bioorg. Med. Chem. Lett. 17 (2007) 2250–2253.

[32] J.E. Sheppeck II, J.L. Gilmore, A. Tebben, C.B. Xue, R.Q. Liu, C.P. Decicco, J.J. Duan, Hydantoins, trizzolones, and imidazolines as selective non-hydroxamate inhibitors of tumor necrosis factor-α converting enzyme (TACE), Bioorg. Med. Chem. Lett. 17 (2007) 2769–2774.

[33] J.J. Duan, L. Chen, Z. Lu, B. Jiang, N. Asakawa, J.E. Sheppeck II, R.Q. Liu, M.B. Covington, W. Pitts, S.H. Kim, C.P. Decicco, Discovery of low nanomolar non-hydroxamate inhibitors of tumor necrosis factor-α converting enzyme (TACE), Bioorg. Med. Chem. Lett. 17 (2007) 266–271.

[34] J.E. Sheppeck II, J.L. Gilmore, A. Yang, X.T. Chen, C.B. Xue, J. Roderick, R.Q. Liu, M.B. Covington, C.P. Decicco, J.J. Duan, Discovery of novel hydantoins as selective non-hydroxamate inhibitors of tumor necrosis factor-α converting enzyme (TACE), Bioorg. Med. Chem. Lett. 17 (2007) 1413–1417.

[35] J.E. Sheppeck II, A. Tebben, J.L. Gilmore, a. Yang, Z.R. Wasserman, C.P. Decicco, J.J. Duan, A molecular modeling analysis of novel non-hydroxamate inhibitors of TACE, Bioorg. Med. Chem. Lett. 17 (2007) 1408–1412.

[36] A. Zask, J. Kaplan, X. Du, G. MacEwan, V. Sandanayaka, N. Eudy, J. Levin, G. Jin, J. Xu, T. Cummons, D. Barone, S. Ayral-Kaloustian, J. Skotnicki, Synthesis and SAR of diazepine and thiazepine TACE and MMP inhibitors, Bioorg. Med. Chem. Lett. 15 (2005) 1641–1645.

[37] G.R. Ott, N. Asakawa, Z. Lu, R.Q. Liu, M.B. Covington, K. Vaddi, M. Qian, R.C. Newton, D.D. Christ, J.M. Traskos, C.P. Decicco, J.J. Duan, α,β-Cyclic-β-benzamido hydroxamic acids: Novel templates for the design, synthesis, and evaluation of selective inhibitors of TNF-α converting enzyme (TACE), Med. Chem. Lett. 18 (2008) 694–699.

[38] J.S. Condon, D. Joseph-McCarthy, J.I. Levin, H.G. Lombart, F.E. Lovering, L. Sun, W. Wang, W. Xu, Y. Zhang, Identification of potent and selective TACE inhibitors via the S1 pocket, Bioorg. Med. Chem. Lett. 17 (2007) 34–39.

[39] A. Huang, D. Joseph-McCarthy, F. Lovering, L. Sun, W. Wang, W. Xu, Y. Zhu, J. Cui, Y. Zhang, J.I. Levin, Structure-based design of TACE selective inhibitors: Manipulations in the S1′–S3′ pocket, Bioorg. Med. Chem. Lett. 15 (2007) 6170–6181.

[40] Z. Lu, G.R. Ott, R. Anand, R.Q. Liu, M.B. Covington, K. Vaddi, M. Qian, R.C. Newton, D.D. Christ, J. Trzaskos, J.J. Duan, Potent, selective, orally bioavailable inhibitors of tumor necrosis factor-α converting enzyme (TACE): Discovery of indole, benzofuran, imidazopyridine and pyrazolopyridine P1′ substituents, Bioorg. Med. Chem. Lett. 18 (2008) 1958–1962.

[41] G.R. Ott, N. Asakawa, Z. Lu, R. Anand, R.Q. Liu, M.B. Covington, K. Vaddi, M. Qian, R.C. Newton, D.D. Christ, J.M. Trzaskos, J.J. Duan, Potent, exceptionally selective, orally bioavailable inhibitors of TNF-α Converting Enzyme (TACE): Novel 2-substituted-1*H*-benzo [*d*]imidazol-1-yl)methyl)benzamide P1′ substituents, Bioorg. Med. Chem. Lett. 18 (2008) 1577–1582.

[42] S.P. Gupta, Quantitative structure–activity relationship studies on zinc-containing metalloproteinase inhibitors, Chem. Rev. 107 (2007) 3042–3087.

[43] P.R. Murumkar, S.D. Gupta, V.P. Zambre, R. Giridhar, M.R. Yadav, Development of Predictive 3D-QSAR CoMFA and CoMSIA models for β-aminohydroxamic acid-derived tumor necrosis factor-α converting enzyme inhibitors, Chem. Biol. Drug. Des. 73 (2009) 97–107.

[44] P.R. Murumkar, R. Giridhar, M.R. Yadav, 3D-quantitative structure–activity relationship studies on benzothiadiazepine hydroxamates as inhibitors of tumor necrosis factor-α converting enzyme, Chem. Biol. Drug. Des. 71 (2008) 363–373.

[45] Y. Xue, H. Li, C.Y. Ung, C.W. Yap, Y.Z. Chen, Classification of a diverse set of *Tetrahymena pyriformis* toxicity chemical compounds from molecular descriptors by statistical learning methods, Chem. Res. Toxicol. 19 (2006) 1030–1039.

[46] F. Luan, H.T. Liu, W.P. Ma, B.T. Fan, Classification of estrogen receptor-β ligands on the basis of their binding affinities using support vector machine and linear discriminant analysis, Euro. J. Med. Chem. 43 (2008) 43–52.

[47] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, Y.Z. Chen, Prediction of P-glycoprotein substrates by a support vector machine approach, J. Chem. Inf. Comput. Sci. 44 (2004) 1497–1505.

[48] M.W.B. Trotter, S.B. Holden, Support vector machines for ADME property classification, QSAR Comb. Sci. 22 (2003) 533–548.

[49] R.O. Duda, P.E. Hart, D.G. Stork, Unsupervised Learning and Clustering, In Pattern Classification, second ed., John Wiley & Sons, New York, 2001, pp. 517.

[50] L.Y. Han, X.H. Ma, H.H. Lin, J. Jia, F. Zhu, Y. Xue, Z.R. Li, Z.W. Cao, Z.L. Ji, Y.Z. Chen, A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor, J. Mol. Graph. Model. 26 (2008) 1276–1286.

[51] C.Y. Liew, X.H. Ma, X. Liu, C.W. Yap, CVM model for virtual screening of Lck inhibitors, J. Chem. Inf. Model. 49 (2009) 877–885.

[52] CambridgeSoft Corporation, ChemDraw, 7.0.1 edn., CambridgeSoft Corporation, Cambridge, MA, USA, 2007.

[53] Corina, Version 3.4; Molecular Networks, GmbH Computerchemie, Germany, 2006.

[54] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.

[55] J.Y. Hu, T. Aizawa, Quantitative structure–activity relationships for estrogen receptor binding affinity of phenolic chemicals, Water Res. 37 (2003) 1213–1222.

[56] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inform. Comp. Sci. 44 (2004) 1630–1638.

[57] V.N. Vapnic, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[58] M.W.B. Trotter, B.F. Buxton, S.B. Holden, Support vector machines in combinatorial chemistry, Meas. Cont. 34 (2001) 235–239.

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

[60] H. Yu, J. Yang, W. Wang, J. Han, Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines, Proc. IEEE Comput. Soc. Bioinform. Conf. 2 (2003) 220–228.

[61] I. Guyon, J. Weston, S. Barnhill, V. Vapnic, Gene selection for cancer classification using support vector machine, Mach. Learn. 46 (2002) 389–422.

[62] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machine, J. Chem. Inf. Model. 45 (2005) 982–992.

[63] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

[64] P. Baldi, S. Brunak, Y. Chauvin, C.A. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.

[65] J.J. Perez, Managing molecular diversity, Chem. Soc. Rev. 34 (2005) 143–152.

[66] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.