Journal of Molecular Graphics and Modelling 48 (2014) 87–95

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Models for H_3 receptor antagonist activity of sulfonylurea derivatives

Naveen Khatri, A.K. Madan*

Faculty of Pharmaceutical Sciences, Pt. B. D. Sharma University of Health Sciences, Rohtak 124001, India

ARTICLE INFO

Article history:
Received 27 April 2013
Received in revised form 7 December 2013
Accepted 17 December 2013
Available online 25 December 2013

Keywords:
Histamine H_3 receptor antagonist
Sulfonylurea derivatives
Decision tree
Dragon software
Moving average analysis
Random forest

ABSTRACT

The histamine H_3 receptor has been perceived as an auspicious target for the treatment of various central and peripheral nervous system diseases. In present study, a wide variety of 60 2D and 3D molecular descriptors (MDs) were successfully utilized for the development of models for the prediction of antagonist activity of sulfonylurea derivatives for histamine H_3 receptors. Models were developed through decision tree (DT), random forest (RF) and moving average analysis (MAA). Dragon software version 6.0.28 was employed for calculation of values of diverse MDs of each analogue involved in the data set. The DT classified and correctly predicted the input data with an impressive non-error rate of 94% in the training set and 82.5% during cross validation. RF correctly classified the analogues into active and inactive with a non-error rate of 79.3%. The MAA based models predicted the antagonist histamine H_3 receptor activity with non-error rate up to 90%. Active ranges of the proposed MAA based models not only exhibited high potency but also showed improved safety as indicated by relatively high values of selectivity index. The statistical significance of the models was assessed through sensitivity, specificity, non-error rate, Matthew's correlation coefficient and intercorrelation analysis. Proposed models offer vast potential for providing lead structures for development of potent but safe H_3 receptor antagonist sulfonylurea derivatives.

1. Introduction

The search for new, effective and safe drugs has now become increasingly complex and refined. Two distinct characteristics clearly define the modern age of the pharmaceutical industry: "*competitiveness*" and "*high cost*". To alleviate these problems, efforts have been directed to reduce the cost and time span required for the discovery of a new drug molecule. More and more computer approaches are now being developed so as to lower the cost and reduce the time cycle for discovering a new drug. The step of lead discovery is considered a bottle-neck of the drug discovery process. It was reported that only one potential lead can be identified by random screening of approximately twenty thousand chemicals. Therefore, the efficiency of mere random screening looks buried in the deep sea. Computer aided drug design (CADD) approaches, which aims to help the rapid and efficient discovery of drug leads may be broadly classified into three categories i.e. combinatorial chemistry based approaches, receptor structure based drug design and (quantitative) structure activity relationship [(Q)SAR] based drug design [1–3].

(Q)SAR based drug design is one of the most standard and authoritative approach used in drug design. History of this approach may be traced back in the year of 1868 when Crum-Brown and Fraser published the first equation in the field of QSAR, which set forth the idea that the biological activity of a compound is a function of its structural properties. Nearly hundred years later, Hansch and Fujita published the extra thermodynamic approach (also called Hansch approach), which says that the activity of a drug is related to three descriptors, namely the hydrophobicity parameter, the electrostatic parameter and the stereo parameter. Modern (Q)SAR studies use much more complicated descriptors to describe the structural features of chemical entities like topostructural (TS), topochemical (TC), topological charge indices, walk and path counts, information based indices, 3D-MoRSE descriptors and Eigenvalue-based indices, etc [4–6].

Molecular descriptors (MDs) are the numbers that contain structural information derived from the structural representation of the molecule under consideration. MDs can be calculated with the help of topological matrix, which in turn can be obtained from a hydrogen suppressed molecular graph [7]. However, only a small fraction of MDs have been successfully employed in (Q)SAR studies. Like MDs, selection of a proper statistical approach plays a vital role in development of a good (Q)SAR model. Decision tree (DT), random forest (RF), multiple linear regression (MLR), partial least squares (PLS), principal component analysis (PCA), genetic algorithm (GA), artificial neural network (ANN) and moving average analysis (MAA) are some of the statistical approaches which have been successfully explored in SAR studies to construct models [8–11].

* Corresponding author. Tel.: +91 98963 46211; fax: +91 1262213202.
E-mail address: madan_ak@yahoo.com (A.K. Madan).

1093-3263/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2013.12.005

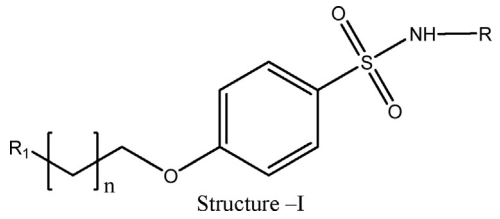

Fig. 1. Basic structure of sulfonamide and sulfonylurea derivatives [15].

Four histamine receptor subtypes, each belonging to the class of G-protein-coupled receptors (GPCRs) have been reported in the literature. Antagonists for the histamine H_1 receptor (e.g. *L*-cetirizine) and the histamine H_2 receptor (*e.g.* cimetidine) have been successfully used for the treatment of allergic conditions and gastric ulcer respectively. The histamine H_3 receptor (H_3R) is considered to be an attractive target in the field of drug discovery because of its expression in brain regions that are critical for cognition (cortex and hippocampus), sleep and homeostatic regulation (hypothalamus) and it also act as a heteroreceptor modulating the release of several important neurotransmitters that are involved in processes like cognition, mood and sensory gating. In addition, the H_3R acts as an autoreceptor regulating the release and synthesis of histamine, which is an important neurotransmitter that plays a role in impulsivity, vigilance, attention and feeding/weight regulation. Therefore, antagonists for the H_3R are currently under investigation in several therapeutic areas including sleep disorders, epilepsy, Alzheimer's disease, energy homeostasis, obesity and cognitive disorders [12–15].

In the present study, MDs of wide nature were successfully utilized for the development of models for prediction of histamine H_3 receptor antagonist activity of sulfonylurea derivatives using DT, RF and MAA.

2. Methodology

2.1. Dataset

A dataset comprising of 49 sulfonylurea derivatives reported by Ceras et al. was selected for developing the models for the present investigation [15]. The basic structure of sulfonylurea derivatives has been illustrated in Fig. 1 and the various substituents enlisted in Table 3. The antagonist H_3R activity of sulfonylurea derivatives has been reported in terms of IC_{50} by Ceras et al. [15]. All the analogues possessing IC_{50} values of $\leq$1.0 μM were considered to be active and analogues possessing IC_{50} values of >1.0 μM were considered to be inactive for the purpose of present study.

2.2. Molecular descriptors

4885 2D and 3D molecular descriptors of diverse nature were used to capture the structural characteristics of the compounds from all aspects. All the molecular structures were built using ChemBio3Dultra (version 11.0.1), energies were minimized and structures were saved as Sybyl2(mol2) files. These mol2 files were used for calculating MDs using Dragon software (version 6.0.28). The MDs used in the current study include constitutional, physicochemical, topostructural, topochemical, topological charge indices, walk and path counts, geometrical descriptors, GETAWAY descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors and information based indices etc. Most of these MDs have been described in the Dragon software and in textbooks by Todeschini and Consonni [16,17]. All the MDs were examined manually and MDs expressing zero values or similar values for different molecules i.e. having high degeneracy were omitted from the pool of 4885 MDs. For the remaining MDs, a pair wise correlation analysis was carried out and one of any two descriptors with $r \geq 0.97$ was omitted from the study. This was done to minimize redundant information and collinearity between MDs [18]. Finally, 60 MDs were employed for the development of models using DT and RF. Subsequently, 4 non-correlating/poorly correlating MDs were employed for development of models through MAA.

3. Statistical methods

DT, RF, MLR and MAA were selected for modeling due to their simplicity and promising results reported during recent past. DT by their association with logic-based and expert systems distinguishes itself from other classification and regression algorithms. Compared with many other modeling techniques, such as ANN and SVM, etc. it is relatively simple and yields easily interpretable results [19,20]. RF is based on an ensemble of hundreds or thousands of classification trees called "forests", such that each tree grows on the value of an input random vector, independently introduced and with the same distribution for all the trees in a forest [21]. MLR, however, did not yield any satisfactory result. MAA models are unique and differ widely from conventional QSAR models. Both systems of modeling have their advantages and limitations. This modeling system has the distinct advantage of identification of narrow active range(s), which may be erroneously skipped during routine regression analysis in conventional correlation modeling [22].

3.1. Decision tree

DT is a supervised rule based method that provides both classification and predictive functions simultaneously. It was first applied by Morgan and Sonquist in 1963 but it gained popularity after the work of Breinman et al. in 1984, who introduced the classification and regression tree method [23,24]. DT works by finding some features from the pool of descriptors for molecules in each class using training set and based on these features some rules are created called the nodes or leafs of the tree. A single DT was grown to identify the importance of various descriptors and for the prediction of antagonist H_3R activity of sulfonylurea derivatives used for the present study. The index value for each descriptor was assigned that represents a margin dividing the compounds of dataset into active and inactive with regard to antagonist H_3R activity. Then, a single descriptor is identified that split the entire training set into two or more homogenous subsets and shows the lowest possible false assignment before being chosen as parent node. The molecules at each parent node are classified, based on the descriptor value, into two child nodes and resulting child nodes or subsets are split into sub-subsets, generally using different descriptors. The prediction for a molecule reaching a given terminal node is obtained through majority vote of the molecules reaching the same terminal node in training set. In this manner, DT created an interactive branching topology in which the branch taken at each intersection is determined by a rule related to a descriptor of the molecule and lastly, each terminating leaf of the tree is assigned to a particular category i.e. A (active) or B (inactive) [11]. In this study, R program (version 2.10.1) along with the RPART library was used to grow DT.

3.2. Random forest

RF is an ensemble of unpruned classification trees created by using bootstrap samples of the training data to construct multiple trees (forests) and random subsets of variables to define the best split at each node, hence the name "*random*" forests [25]. In the

present study, the RFs were grown with the R program (version 2.10.1) using the random forest library.

3.3. Moving average analysis

MAA was utilized so as to facilitate construction of single MD based models for antagonist H_3R activity of sulfonylurea derivatives. For the selection and evaluation of range specific features, exclusive activity ranges were discovered from the frequency distribution of therapeutic response level. This was accomplished by plotting the relationship between descriptor values and antagonist H_3R activity and then identifying the active range by analyzing the resultant data by maximization of moving average with respect to active compounds (<35% = inactive, 35–65% = transitional and >65% = active) [26]. Subsequently, a biological activity was assigned to each analogue involved in the dataset and compared with the reported antagonist H_3R activity (Table 3). Average values of IC_{50} and selectivity index (SI) were calculated for each range of the proposed models.

3.4. Model validation

One of the most important steps to produce reliable and useful (Q)SAR models is the validation which includes determining the predictive power of the model [27]. The validation of the DT based models and self-consistency test was performed by 10-fold cross validation (CV) method. The folds were obtained by random splitting method. The performance of proposed models was evaluated by calculating sensitivity, specificity, non-error rate (arithmetic mean of sensitivity and specificity), overall accuracy of prediction and Matthews's correlation coefficient (MCC) [28–31].

The sensitivity and specificity may be calculated as per the well known and widely reported expression:

$$\text{Sensitivity} = \frac{P^T}{P^T + N^F},$$

$$\text{Specificity} = \frac{N^T}{N^T + P^F}$$

where the true positive (P^T) is the number of compounds correctly predicted as active, false negative (N^F) is the number of compounds incorrectly predicted as inactive, true negative (N^T) is the number of compounds correctly predicted as inactive, false positive (P^F) is the number of compounds incorrectly predicted as active [28,29].

Because of unavailability of perfect method for describing the confusion matrix of true and false positives or negatives by a single number, MCC is generally regarded as being one of the best statistical techniques which account for both over and under prediction. MCC takes both sensitivity and specificity into account and its value ranges from −1 to +1. Higher values of MCC indicate better predictions [32,33]. Statistical significance of MDs used in building predictive models was also assessed by intercorrelation analysis. The degree of correlation was evaluated by the correlation coefficient '*r*'. Pairs of MDs with $r \geq 0.97$ are considered to be highly inter-correlated while those with $0.90 \leq r \leq 0.97$ are appreciably correlated, MDs with $0.50 \leq r \leq 0.89$ are treated to be weakly correlated and finally the pairs of MDs with low *r*-values (<0.50) are not inter-correlated [34].

4. Results and discussion

Dissipating proprietary products, low productivity, rising R&D costs and dwindling pipelines are the driving challenges for the pharmaceutical industry [35]. Use of machine learning approaches provides a solution to these challenges, as (Q)SAR models can be effectively utilized to minimize the time and resource requirement for the drug development process. The objective of this study was to establish relationship between antagonist H_3R activity of sulfonylurea derivatives and their chemical structures by developing suitable models using diverse classification techniques i.e. DT, RF and MAA. Decision tree was built from a set of 60 MDs. The MD at root node is most important and the importance of MD decreases as the length of tree increases [8,11]. The classification of sulfonylurea derivatives as inactive and active using a single tree has been depicted in Fig. 2. The DT identified 3 indices i.e. A50 (mean information index on atomic composition, AAC), A55 (Geary autocorrelation – lag 6/weighted by ionization potential, GATS6i), and A10 (complementary information content (neighborhood symmetry of 1-order), CIC1) as the most important MDs. In the training set the decision tree classified the analogues with an accuracy of 96%.

A50 or AAC descriptor (mean information index on atomic composition) developed by Dancoff et al. is calculated from hydrogen included molecular formula and is the mean of total information content. It may be expressed as per the following:

$$\text{AAC} = -\sum_x \frac{N_x}{N_0} \times \log_2 \frac{N_x}{N_0} = -\sum_x \text{pro}_x \times \log_2 \text{pro}_x$$

where N_0 represents total number of atoms in a molecule, N_x represents number of atoms of type '*x*' and pro_x is the probability of randomly selecting atom of type '*x*' [16,17,36].

A50 encodes relationship between total number of atoms and types of atoms in a molecule. Its value is directly proportional to the types of equivalent classes of and inversely proportional to number of atoms in a molecule. Analysis of data furnished in Table 3 reveals that optimum number of non-hydrogen atoms for active analogues is 20–21 for sulfonamides and 29–32 for sulfonylurea analogues. Any deviation from these numbers will result in descriptor value to be outside the active range with consequent loss of biological activity. Similarly, any halogenated-R substitution will increase the number of equivalent classes in sulfonylurea analogues leading to increase in the descriptor value to be outside the active range. This will result in loss of biological activity.

The sensitivity, specificity, non-error rate, overall accuracy of prediction and MCC of the tenfold cross validated set was of the order of 73.0%, 92.0%, 82.5%, 88.0% and 0.65 respectively (Table 1). High value of MCC simply indicates robustness of proposed DT based model.

The RFs were grown utilizing 60 MDs. The RF classified sulfonylurea analogues with regard to antagonist H_3R activity of sulfonylurea derivatives with an accuracy of 88% and out-of-bag (OOB) estimate of error was 12%. The sensitivity, specificity, non-error rate, accuracy of prediction and MCC value of RF based model for the ten fold cross validated set were found to be 63.6%, 95.0%, 79.3%, 88.0% and 0.63 respectively (Table 1). High value of MCC simply indicates robustness of proposed RF based model.

Four models using A10, A12, A55 and A58 were developed through MAA. A55 and A10 were also identified as most important MDs by decision tree. Two more MDs i.e. A12 and A58 were also used to develop MAA based models for predicting H_3 receptor antagonist activity of sulfonylurea derivatives.

One needs use of numerous models derived from non-correlating/poorly correlating MDs in order to reduce very large library of compounds to a handful of compounds for synthesis and biological screening in a cost effective manner. For synergistic use of proposed models in screening of large number of compounds for identification of lead or active molecules, it is necessary that the descriptor value of identified or designed compounds should fall in the active ranges of each model. Finally, compounds that fall

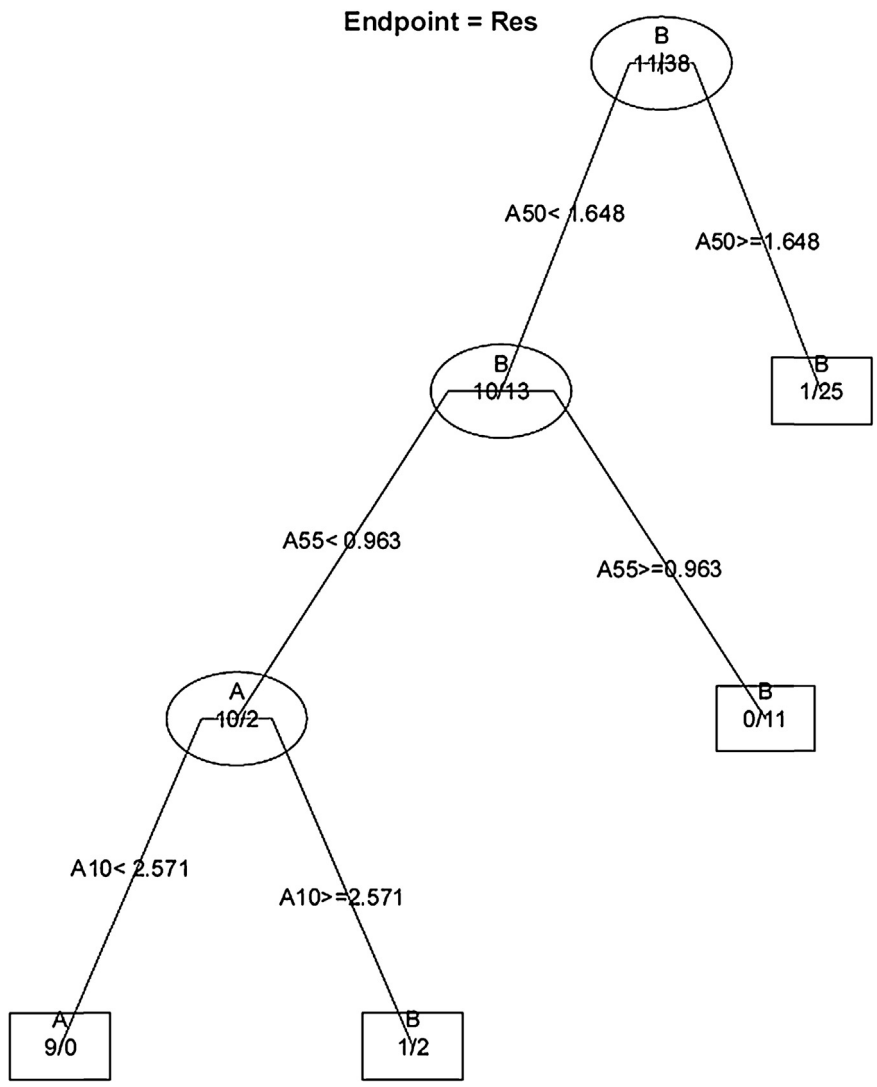

Fig. 2. A decision tree for distinguishing active sulfonylurea derivatives (A) from inactive derivatives (B).

Table 1
Confusion matrix for H_3 receptor antagonist activity of sulfonylurea derivatives and recognition rate of decision tree and random forest based models.

Model	Description	Ranges	Number of compounds predicted		Sensitivity	Specificity	Non error rate (%)	Overall accuracy of prediction (%)	MCC
			Active	Inactive					
Decision Tree	Training set	Active	10	01	91.0	97.0	94.0	96.0	0.88
		Inactive	01	37					
	Cross validated set	Active	08	03	73.0	92.0	82.5	88.0	0.65
		Inactive	03	35					
Random Forest		Active	07	04	63.6	95.0	79.3	88.0	0.63
		Inactive	02	36					

in the active ranges of each model may have a potential to act as H_3 receptor antagonists. Such compounds, which are predicted to be active using the proposed models, should be synthesized and subsequently tested for the desired activity. Thus, the proposed modeling studies can play a vital role for the design and development of new molecules having improved profile in terms of H_3 receptor antagonist activity. Such an approach will naturally reduce time, minimize animal sacrifice and significantly curtail wasteful expenditure and energy.

A10 or CIC1 descriptor [Complementary information content (neighborhood symmetry of 1-order)] developed by Magnuson et al. is an information descriptor based on neighbor degrees and edge multiplicity. It may be expressed as per the following:

$$\mathrm{CIC1} = \log_2 n\mathrm{AM} - \mathrm{IC1}$$

where nAM represents total number of atoms in a molecule and IC1 is neighborhood information content index [16,17,37].

A10 encodes relationship between total number of atoms and neighborhood information content. Its value is directly proportional to number of atoms and inversely proportional to number of types of equivalent classes of atoms in a molecule. Analysis of data

Table 2
Proposed MAA based models for the prediction of H_3 receptor antagonist activity of sulfonylurea derivatives.

Molecular descriptor	Nature of range in proposed model	Molecular descriptor value	Number of compounds in each range		Sensitivity	Specificity	Non error rate (%)	Overall accuracy of prediction (%)	MCC	Average IC_{50} (μM) of correctly predicted compounds in each range	Average SI values of correctly predicted compounds in each range
			Total	Correctly predicted							
A10	Lower inactive	<2.293	17	17	85.7	94.0	90.0	92.0	0.75	7087	9.5
	Transitional	2.293–2.416	11	NA						922.1	132.1
	Active	2.417–2.491	8	6						0.42	361.2
	Upper inactive	>2.491	13	12						10852.1	10.73
A12	Lower inactive	<1.439	29	26	63.6	97.0	80.3	90.0	0.69	5412.5	8.09
	Active	1.439–1.548	8	7						0.46	234.9
	Upper inactive	>1.548	12	11						10923.6	10.9
A55	Lower inactive	<0.918	19	17	75.0	100	87.5	95.0	0.84	1795.6	7.01
	Active	0.918–0.927	6	6						0.49	195.45
	Transitional	0.928–0.958	7	NA						14291.7	243.47
	Upper inactive	>0.958	17	17						7666.3	9.2
A58	Lower inactive	<20.704	20	18	62.5	94.0	78.3	88.0	0.60	13911.5	7.3
	Transitional	20.704–22.6	7	NA						9.8	256
	Active	22.61–25.446	7	5						0.52	172.1
	Upper inactive	>25.446	15	14						742.1	11.2

NA: not applicable.

Table 3
Relationship between molecular descriptors and H_3 receptor antagonist activity.

Sr. no	Substituent (R_1)	Substituent (R)	Number of carbons in alkane chain (n)	A50	A10	A12	A55	A58	H_3 receptor antagonist activity				
									Predicted				
									A50	A10	A12	A55	A58
1-A	Piperidin-1-yl	H	2	1.626	2.304	1.118	0.804	32.453	+	±	–	–	–
2-A	Pyrrolidin-1-yl	H	3	1.626	2.304	1.023	0.923	24.94	+	±	–	+	+
3-A	Piperidin-1-yl	H	3	1.594	2.476	0.962	0.923	25.14	+	+	–	+	+
4-A	Piperidin-1-yl	N-phenylformamide	2	1.64	2.461	1.517	0.925	23.151	+	+	+	+	+
5-A	Pyrrolidin-1-yl	N-(4-acetylphenyl)formamide	2	1.671	2.293	1.464	0.927	15.25	–	±	+	+	–
6-A	Pyrrolidin-1-yl	N-(4-methylphenyl)formamide	2	1.64	2.339	1.481	0.918	22.611	+	±	+	+	+
7-A	Pyrrolidin-1-yl	N-(1-napthyl)formamide	2	1.625	2.456	1.681	0.948	21.003	+	+	–	±	±
8-A	Pyrrolidin-1-yl	N-[4-(N,N′-dimethylamino)phenyl]formamide	2	1.647	2.491	1.548	0.945	16.903	+	+	+	±	–
9-A	Piperidin-1-yl	N-(4-acetylphenyl)formamide	2	1.647	2.417	1.439	0.927	22.166	+	+	+	+	±
10-A	Piperidin-1-yl	N-(4-methylphenyl)formamide	2	1.616	2.466	1.464	0.917	25.446	+	+	+	–	+
11-A	Piperidin-1-yl	N-benzhydrylformamide	2	1.56	2.824	1.459	0.958	20.704	+	–	+	±	±
12-I	Pyrrolidin-1-yl	H	1	1.702	1.938	0.744	0.621	22.303	–	–	–	–	±
13-I	Piperidin-1-yl	H	1	1.662	2.123	0.941	0.686	23.103	–	–	–	–	+
14-I	(4-Ethoxycarbonyl)piperidin-1-yl	H	1	1.676	2.204	0.797	0.909	17.74	–	–	–	–	–
15-I	(4-Ethoxycarbonyl)piperidin-1-yl	H	2	1.647	2.352	0.906	0.968	16.833	+	±	–	–	–
16-I	(4-Ethoxycarbonyl)piperidin-1-yl	H	3	1.62	2.497	0.832	1.048	18.935	+	–	–	–	–
17-I	Pyrrolidin-1-yl	N-isopropylformamide	1	1.679	2.234	1.574	0.948	12.824	–	–	–	±	–
18-I	Pyrrolidin-1-yl	N-phenylformamide	1	1.694	2.217	1.344	0.847	19.624	–	–	–	–	–
19-I	Pyrrolidin-1-yl	N-cyclohexylformamide	1	1.623	2.651	1.637	0.898	15.465	+	–	–	–	–
20-I	Pyrrolidin-1-yl	N-(2,5-dichlorophenyl)formamide	1	1.89	1.942	1.083	0.859	32.962	–	–	–	–	–
21-I	Pyrrolidin-1-yl	N-(4-trifluromethylphenyl)formamide	1	1.916	2.01	1.045	0.52	29.179	–	–	–	–	–
22-I	Piperidin-1-yl	N-isopropylformamide	1	1.65	2.38	1.661	0.965	18.784	–	±	–	–	–
23-I	Piperidin-1-yl	N-phenylformamide	1	1.666	2.337	1.442	0.87	25.46	–	±	+	–	–
24-I	Piperidin-1-yl	N-cyclohexylformamide	1	1.598	2.794	1.726	0.917	20.346	+	–	–	–	–
25-I	Piperidin-1-yl	N-(2,5-dichlorophenyl)formamide	1	1.856	2.078	1.163	0.882	32.353	–	–	–	–	–
26-I	Piperidin-1-yl	N-(4-trifluromethylphenyl)formamide	1	1.883	2.135	1.082	0.54	29.046	–	–	–	–	–
27-I	(4-Ethoxycarbonyl)piperidin-1-yl	N-phenylformamide	1	1.677	2.401	1.119	0.984	16.38	–	±	–	–	–
28-I	Pyrrolidin-1-yl	N-isopropylformamide	2	1.65	2.38	1.749	1.021	20.834	–	±	–	–	±
29-I	Pyrrolidin-1-yl	N-cyclohexylformamide	2	1.598	2.794	1.814	0.97	18.631	+	–	–	–	–
30-I	Pyrrolidin-1-yl	N-(2,5-dichlorophenyl)formamide	2	1.856	2.078	1.248	0.939	32.817	–	–	–	±	–
31-I	Pyrrolidin-1-yl	N-(4-trifluromethylphenyl)formamide	2	1.883	2.135	1.124	0.569	23.704	–	–	–	–	+
32-I	Piperidin-1-yl	N-isopropylformamide	2	1.623	2.524	1.715	1.015	20.457	–	–	–	–	–
33-I	Piperidin-1-yl	N-(2,5-dichlorophenyl)formamide	2	1.824	2.215	1.226	0.937	35.738	–	–	–	±	–
34-I	Piperidin-1-yl	N-(4-trifluromethylphenyl)formamide	2	1.852	2.263	1.096	0.576	37.245	–	–	–	–	–
35-I	(4-Ethoxycarbonyl)piperidin-1-yl	N-phenylformamide	2	1.655	2.508	1.189	1.018	18.465	–	–	–	–	–
36-I	Pyrrolidin-1-yl	N-isopropylformamide	3	1.623	2.524	1.614	1.103	16.906	+	–	–	–	–
37-I	Pyrrolidin-1-yl	N-phenylformamide	3	1.64	2.461	1.43	1.007	20.787	+	+	–	–	±
38-I	Pyrrolidin-1-yl	N-cyclohexylformamide	3	1.576	2.93	1.696	1.045	16.433	+	–	–	–	–
39-I	Pyrrolidin-1-yl	N-(2,5-dichlorophenyl)formamide	3	1.824	2.215	1.156	1.021	35.899	–	–	–	–	–
40-I	Pyrrolidin-1-yl	N-(4-trifluromethylphenyl)formamide	3	1.852	2.263	1.034	0.624	31.938	–	–	–	–	–
41-I	Piperidin-1-yl	N-isopropylformamide	3	1.599	2.663	1.556	1.092	20.456	+	–	–	–	–
42-I	Piperidin-1-yl	N-phenylformamide	3	1.616	2.584	1.377	1.001	20.226	+	–	–	–	–
43-I	Piperidin-1-yl	N-cyclohexylformamide	3	1.555	3.059	1.643	1.039	15.976	+	–	–	–	–
44-I	Piperidin-1-yl	N-(2,5-dichlorophenyl)formamide	3	1.794	2.351	1.111	1.014	37.199	–	±	–	–	–
45-I	Piperidin-1-yl	N-(4-trifluromethylphenyl)formamide	3	1.823	2.39	0.996	0.629	33.279	–	±	–	–	–
46-I	(4-Ethoxycarbonyl)piperidin-1-yl	N-phenylformamide	3	1.634	2.616	1.115	1.077	15.84	+	–	–	–	–
47-I	Pyrrolidin-1-yl	N-(3-trifluromethylphenyl)formamide	2	1.883	2.135	0.862	0.862	30.296	–	–	–	–	–
48-I	Piperidin-1-yl	N-(2-trifluromethylphenyl)formamide	2	1.852	2.263	1.045	0.681	33.059	–	–	–	–	–
49-I	Piperidin-1-yl	N-(4-methoxyphenyl)formamide	2	1.651	2.491	1.298	0.939	22.458	–	+	–	±	±

Note: + Active, – inactive, ± transitional.
Analogs from Sr. No. 1-A to 11-A reportedly having IC_{50} values ≤1.0 μM [15] were considered to be active and analogs from Sr. No. 12-I to 49-I reportedly having IC_{50} values >1.0 μM [15] were considered to be inactive.

furnished in Tables 2 and 3 reveals following information. Firstly, optimum number of non-hydrogen atoms for active analogues is 20–21 for sulfonamides and 29–32 for sulfonylurea analogues. Any deviation from these numbers will result in descriptor value to be outside the active range and the analogues will most probably be inactive. Secondly, in case of sulfonamides, ethoxycarbonyl substitution on amino ring results in deviation from optimum descriptor value with consequent loss of biological activity. Accordingly, any halogenated-*R* substitution will increase the number of equivalent classes in sulfonylurea analogues leading to decrease in the descriptor value to be outside the active range with consequent loss of biological activity. If in *R* substituent, aromatic benzene ring is replaced by cyclohexane moiety, the type of equivalent classes of atoms remains the same but it increases the number of atoms (6 hydrogen). This leads to increase in descriptor value to be outside the active range, resulting in loss of biological activity in sulfonylurea derivatives.

A12 or *GATS7e descriptor* (Geary autocorrelation – lag 7/weighted by atomic Sanderson electronegativities) is a 2D autocorrelations molecular descriptor which deals with distribution of atomic Sanderson electronegativities in a topological molecular structure. It is calculated from an H-included molecular graph. It may be expressed as per the following:

$$GATS7e = \frac{(1/2\Delta)\sum_{i=1}^{n}\sum_{j=1}^{n}\delta ij(e_i - e_j)^2}{(1/n-1)\sum_{i=1}^{n}(ei - \bar{e})^2}$$

where e_i is atomic Sanderson electronegativity, $\bar{e}$ is the average value of atomic Sanderson electronegativity in the molecule, n is the number of atoms, δ_{ij} is the Kronecker delta (equal to one if $\delta_{ij} = 7$, zero otherwise, δ_{ij} being the topological distance between two considered atoms). Δ is the sum of the Kronecker deltas, that is, the number of atom pairs at distance equal to 7 [17,38].

The value of A12 is directly proportional to number of atoms and inversely proportional to number of atom pairs at a distance equal to 7. As the number of types of hetero atoms in a molecule increases, deviation from average atomic Sanderson electronegativity also increases and results in decrease of descriptor value. Analysis of data furnished in Tables 2 and 3 reveals the following information. Firstly, optimum number of non-hydrogen atoms for active analogues is 20–21 for sulfonamides and 29–32 for sulfonylurea derivatives. Any deviation from these numbers will lead descriptor value to be outside the active range and analogues will most probably be inactive. Secondly, any halogenated-R substitution will increase the average atomic Sanderson electronegativities in sulfonylureas leading to decrease in the descriptor value to be outside the active range with consequent loss of biological activity. If in R substituent, aromatic benzene ring is replaced by cyclohexane moiety, the type of atoms having atomic Sanderson electronegativities remains same but it increases the number of atoms (6 hydrogen) resulting in an increase in descriptor value to be outside the active range with consequent loss of biological activity in sulfonylurea derivatives.

A55 or *GATS6i descriptor* (Geary autocorrelation – lag 6/weighted by ionization potential) is a 2D autocorrelations molecular descriptor which deals with distribution of atomic ionization potentials along a topological molecular structure. It is calculated from an H-included molecular graph. It may be expressed as per the following:

$$GATS6i = \frac{(1/2\Delta)\sum_{i=1}^{n}\sum_{j=1}^{n}\delta ij(i_i - i_j)^2}{(1/n-1)\sum_{i=1}^{n}(i_i - \bar{i})^2}$$

where i_i is atomic ionization potential, $\bar{i}$ is the average value of atomic ionization potentials on the molecule, n is the number of atoms, δ_{ij} is the Kronecker delta (equal to one if $\delta_{ij} = 6$, zero otherwise, δ_{ij} being the topological distance between two considered atoms). Δ is the sum of the Kronecker deltas, that is, the number of atom pairs at distance equal to 6 [17,38].

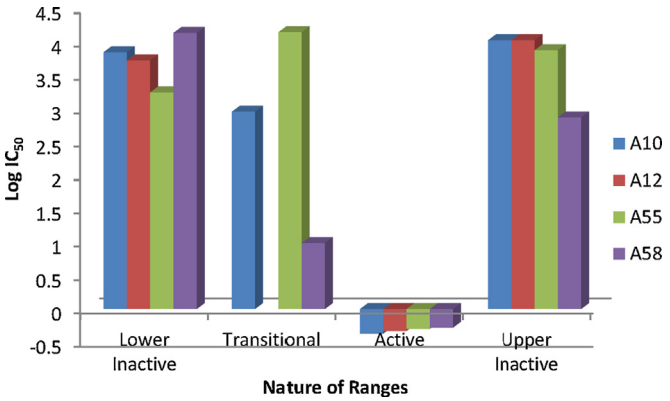

Fig. 3. Average log IC_{50} values of H_3 receptor antagonist activity of correctly predicted sulfonylurea derivatives in various ranges of proposed MAA based models.

Its value is directly proportional to number of atoms and inversely proportional to number of atom pairs at a distance equal to 6. As the number of types of hetero atoms in a molecule increases, deviation from average atomic ionization potential also increases and results in decrease of descriptor value. Analysis of data furnished in Tables 2 and 3 reveals similar results as obtained from 'A12' descriptor.

A58 or *DISPm descriptor* (d COMMA2 value/weighted by atomic masses) is a comparative molecular moment analysis descriptor which gives the magnitude of the displacement between the molecular centroid and property field centre. It was developed by B.D. Silverman [17,39] and may be expressed as per the following:

$$DISPm = \vec{a} - \vec{p}_c = \frac{\vec{M}_1}{M_0}$$

where $\vec{a}$ is the property field centre from an arbitrary location, $\vec{p}_c$ is the property field vector to the centroid of the structure. $\vec{M}_1$ is the first order moment and M_0 is molecular weight of the molecule.

A58 is a 3D descriptor and it encodes relationship between the molecular centroid and property field centre. Analysis of data furnished in Tables 2 and 3 reveals following information. Firstly, optimum number of atoms for active analogues is 20–21 for sulfonamides and 29–32 for sulfonylurea derivatives. Any deviation from this number will lead descriptor value to be outside the active range and analogues will most probably be inactive. Secondly, propoxy chain linkage between the basic amine ring and benzene sulfonamide or benzene sulfonylurea moiety is required for activity. Any deviation from this chain length results in deviation of property field centre from the molecular centroid with consequent loss of biological activity.

Simple MDs are best suited for reverse engineering and screening whereas complex MDs are generally best suited for screening purpose.

Accuracy of prediction for MAA based models varied from 88.0% to 95.0% indicating high predictability (Table 2). The sensitivity, specificity, non-error rate and MCC value of MAA based models varied from 62.5% to 85.7%, 94.0% to 100%, 78.3% to 90.0% and 0.60 to 0.84 respectively (Table 2). High values of MCC simply indicate robustness of proposed MAA based models. The average IC_{50} value of the correctly predicted analogs in the active ranges in MAA based models varied from 0.42 μM to 0.52 μM. Such a low average IC_{50} value signifies high potency of the active ranges (Fig. 3).

Drug safety evaluation is the key part of drug discovery and development process to identify those that have an appropriately balanced safety–efficacy profile for a given indication [40]. The

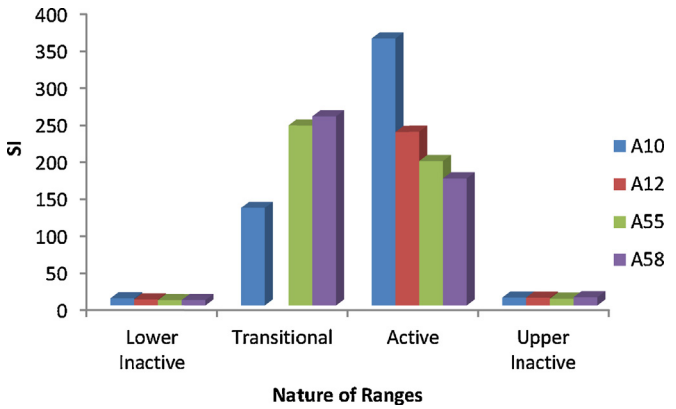

Fig. 4. Average SI values of correctly predicted sulfonylurea derivatives in various ranges of proposed MAA based models.

therapeutic index (TI), certain safety factor (CSF), protective index (PI), therapeutic window (TW) and SI are some of such important parameters that can be used to achieve this balance [41]. SI is calculated for a drug molecule in case of cell studies and it may be defined as the ratio of CC_{50} to EC_{50}. CC_{50} and EC_{50} represent cytotoxic and effective concentrations respectively. It is an indirect measure of safety of drug. High value of SI simply indicates low toxicity and more safety. High value of SI is a desirable property for any drug candidate so as to minimize toxicity [42]. Human ether-a go-go-related gene (HERG) is strongly expressed in the heart. HERG encodes a K^+ channel; many of the reported H_3R antagonists have shown substantial HERG channel inhibition. This inhibition poses a potential safety risk of cardiac toxicity and has become a critical issue for regulatory agencies and the pharmaceutical industry [14,15,43]. To address this problem SI values for sulfonylurea derivatives had been calculated by taking the ratio of HERG channel inhibiting activity (IC_{50}) to H_3R antagonist activity (IC_{50}) from the reported data [15]. Active ranges of proposed MAA based models exhibited high degree of selectivity towards H_3 receptors compared to HERG channel inhibition as indicated by greater value of SI for active ranges compared to inactive ranges (Fig. 4). As a consequence active ranges identified by MAA models have both the desired requirement of a drug molecule i.e. high potency and safety. Model validation by confusion matrix shows the specificity of the proposed models of the order of 94%–100% (Table 2). The value of MCC varied from 0.6 to 0.84. High values of MCC indicate robustness of the proposed MAA based models (Table 2).

The results of intercorrelation analysis (Table 4) reveals that the pairs A12:A55, A12:A58 and A55:A58 were not correlated while the pair A10:A12, A10:A55 and A10:A58 were found to be weakly correlated.

The present modeling studies may be of great utility for providing lead molecules through exploitation of active ranges in the single MD based models. Proposed models are unique and differ widely from conventional QSAR models. Both system of modeling have their advantages and limitations. In the instant modeling, the system adopted has distinct advantage of identification of narrow active ranges, which may be erroneously skipped during regression analysis in conventional QSAR. Since the ultimate goal of modeling is to provide lead structures, therefore, active ranges of the proposed models can play vital role in providing lead structures [22]. Therefore, active ranges of proposed models can naturally play a vital role in providing lead structures.

Table 4
Intercorrelation matrix.

	A10	A12	A55	A58
A10	1.00	0.62	0.58	−0.53
A12		1.00	0.48	−0.45
A55			1.00	−0.48
A58				1.00

5. Conclusion

In the present study, diverse 2D and 3D MDs were successfully utilized for the development of diverse models for prediction of histamine H_3 receptor antagonist activity of sulfonylurea derivatives using DT, RF and MAA. Models based on DT and RF show accuracy of prediction up to the order of 88%. The overall accuracy of prediction of MAA based models varied from 88% to 95%. The sensitivity, specificity, non-error rate and MCC value of MAA based models varied from 62.5%–85.7%, 94%–100%, 78.3%–90% and 0.60–0.84 respectively. High values of sensitivity, specificity, non-error rate and MCC for the proposed models indicate the robustness of the proposed models. High predictability amalgamated with high potency and safety in the active ranges offer proposed MAA based models a vast potential for providing lead structures for development of sulfonylurea derivatives as potent but safe Histamine H_3 receptor antagonists.

References

[1] G.R. Marshall, Computer-aided drug design, Annual Review of Pharmacology and Toxicology 27 (1987) 193–213.
[2] G.H. Loew, H.O. Villar, I. Alkorta, Strategies for indirect computer-aided drug design, Pharmaceutical Research 10 (1993) 475–486.
[3] A.V. Veselovsky, A.S. Ivanov, Strategy of computer-aided drug design, Current Drug Targets – Infectious Disorders 3 (2003) 33–40.
[4] C.D. Sellasie, History of quantitative structure–activity relationships, in: D.J. Abraham (Ed.), Burger's Medicinal Chemistry and Drug Discovery, 6th ed., A John Wiley and sons, New York, 2003, pp. 1–48.
[5] A. Crum-Brown, T.R. Fraser, On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia, Transactions of the Royal Society of Edinburgh 25 (1868) 151–203.
[6] C. Hansch, T. Fujita, Rho–delta–pi. A method for correlation of biological activity and chemical structure, Journal of American Chemical Society 86 (1964) 1616–1626.
[7] J. Gálvez, M. Gálvez-Llompart, R. García-Domenech, Introduction to molecular topology: basic concepts and application to drug design, Current Computer Aided Drug Design 8 (2012) 196–223.
[8] A. Asikainen, M. Kolehmainen, J. Ruuskanen, K. Tuppurainen, Structure-based classification of active and inactive estrogenic compounds by decision tree, LVO and κNN methods, Chemosphere 62 (2006) 658–673.
[9] E.C.M. Nascimento, J.B.L. Martins, Electronic structure and PCA analysis of covalent and non-covalent acetylcholinesterase inhibitors, Journal of Molecular Modeling 17 (2011) 1371–1379.
[10] A. Fassihi, R. Sabet, QSAR study of p56lck protein tyrosine kinase inhibitory activity of flavonoid derivatives using MLR and GA-PLS, International Journal of Molecular Sciences 9 (2008) 1876–1892.
[11] H. Dureja, S. Gupta, A.K. Madan, Topological models for prediction of pharmacokinetic parameters of cephalosporins using random forest, decision tree and moving average analysis, Scientia Pharmaceutica 76 (2008) 377–394.
[12] S.J. Hill, C.R. Ganellin, H. Timmerman, J.C. Schwartz, N.P. Shankley, J.M. Young, W. Schunack, R. Levi, H.L. Haas, International Union of Pharmacology. XIII. Classification of histamine receptors, Pharmacology Review 49 (1997) 253–278.
[13] M.B. Passani, J.S. Lin, A. Hancock, S. Crochet, P. Blandina, The histamine H_3 receptor as a novel therapeutic target for cognitive and sleep disorders, Trends in Pharmacological Sciences 25 (2004) 618–625.
[14] M. Wijtmans, R. Leurs, D.I. Esch, Histamine H_3 receptor ligands break ground in a remarkable plethora of therapeutic areas, Expert Opinion on Investigational Drugs 16 (2007) 967–985.
[15] J. Ceras, N. Cirauqui, S. Pérez-Silanes, I. Aldana, A. Monge, S. Galiano, Novel sulfonylurea derivatives as H_3 receptor antagonists. Preliminary SAR studies, European Journal of Medicinal Chemistry 52 (2012) 1–13.
[16] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, 2000.
[17] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, vol. I/II, 2nd ed., Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009.
[18] R. Dutt, A.K. Madan, Models for the prediction of PPARs agonistic activity of indanylacetic acids, Medicinal Chemistry Research 22 (2013) 3213–3228.
[19] H. Almuallim, T.G. Dietterich, Learning Boolean concepts in the presence of many irrelevant features, Artificial Intelligence 69 (1/2) (1994) 279–306.

[20] N.S. Sethi, A review on computational methods in developing quantitative structure-activity relationship (QSAR), International Journal of Drug Research and Technology 2 (4S) (2012) 313–341.
[21] L. Breiman, Random forests, Machine Learning 45 (2001) 5–35.
[22] H. Dureja, A.K. Madan, Prediction of h5-HT2A receptor antagonistic activity of arylindoles: computational approach using topochemical descriptors, Journal of Molecular Graphics and Modelling 25 (2006) 373–379.
[23] J.N. Morgan, J.A. Sonquist, Problems in the analysis of survey data and a proposal, Journal of the American Statistical Association 58 (1963) 415–434.
[24] L. Breinman, J.H. Friedman, R.A. Olshen, J.C. Stone, Classification and Regression Trees, Wandsworth, Pacific Grove, CA, USA, 1984.
[25] A.M. Prasad, L.R. Iverson, A. Liaw, Newer classification and regression tree techniques: bagging and random forests for ecological prediction, Ecosystems 9 (2006) 181–199.
[26] S. Gupta, M. Singh, A.K. Madan, Predicting anti-HIV activity: computational approach using novel topological indices, Journal of Computer-Aided Molecular Design 15 (2001) 671–678.
[27] A. Tropsha, P. Gramatica, V. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR & Combinatorial Science 22 (2003) 69–76.
[28] L. Han, Y. Wang, S.H. Bryant, Developing and validating predictive decision tree models from mining chemical structural fingerprints and high throughoutput data in PubChem, BMC Bioinformatics 9 (2008) 401.
[29] C. Lamanna, M. Bellini, A. Padova, G. Westerberg, L. Maccari, Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process, Journal of Medicinal Chemistry 51 (2008) 2891–2897.
[30] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Analytical Methods 5 (2013) 3790–3798.
[31] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta 405 (1975) 442–451.
[32] P. Baldi, S. Bruank, Y. Chauvin, C.A.F. Andersen, H. Nielsen, Assessing the accuracy of prediction algorithms for classification: an overview, Bioinformatics 16 (2000) 412–424.
[33] http://en.wikipedia.org/wiki/Matthews correlation coefficient (accessed 15.06.13).
[34] N. Trinajstic, S. Nikolic, S.C. Basak, I. Lukovits, Distance indices and their hypercounterparts: intercorrelation and use in the structure property modeling, SAR and QSAR in Environmental Research 12 (2001) 31–54.
[35] I. Khanna, Drug discovery in pharmaceutical industry: productivity challenges and trends, Drug Discovery Today 17 (2012) 1088–1102.
[36] S.M. Dancoff, H. Quastler, The information content and error rate of living things, in: H. Quastler (Ed.), Essays on the Use of Information Theory in Biology, University of Illinois Press, Urbana, IL, 1953.
[37] V.R. Magnuson, D.K. Harriss, S.C. Basak, Topological indices based on neighbor symmetry: chemical and biological applications, in: R.B. King (Ed.), Chemical Applications of Topology and Graph Theory, Elsevier, Amsterdam, 1983, pp. 178–191.
[38] R.C. Geary, The contiguity ratio and statistical mapping, The Incorporated Statistician 5 (1954) 115–145.
[39] B.D. Silverman, Three-dimensional moments of molecular property fields, Journal of Chemical Information and Computer Sciences 40 (2000) 1470–1476.
[40] P.Y. Muller, M.N. Milton, The determination and interpretation of the therapeutic index in drug development, Nature Reviews Drug Discovery 11 (2012) 751–761.
[41] M.E. Barka, A.W. Hayes, in: A.W. Hayes (Ed.), Principles and Methods of Toxicology, CRC Press Taylor and Francis Group, Boca Raton, 2001, pp. 1131–1141.
[42] A.K. Madan, S. Bajaj, H. Dureja, in: B. Reisfeld, A.N. Mayeno (Eds.), Computational Toxicology, vol. 2, Humana Press, Springer Science, NY, 2013, pp. 99–102.
[43] M.E. Curran, I. Splawski, K.W. Timothy, G.M. Vincent, E.D. Green, M.T. Keating, A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome, Cell 80 (1995) 795–803.