J. Indian Chem. Soc.,
Vol. 96, June 2019, pp. 685-691

QSAR modeling and in silico designing of Tumor-Associated Carbonic Anhydrases XII inhibitors

Shalini Singh

QSAR & Cheminformatics Laboratory, Department of Chemistry, Bareilly College, Bareilly-243 001, Uttar Pradesh, India

E-mail: shalinisingh_15@yahoo.com

Manuscript received online 23 March 2018, accepted 22 May 2019

Substituted coumarin such as sulfocoumarins (1,2-benzoxathiine-2,2-dioxides) possessing are the most important class of Potent and Isoform-Selective Inhibitors of Tumor-Associated Carbonic Anhydrases CA XII.

I have attempted to build QSAR models to explore the correlations between the calculated molecular descriptors on the pool of 16 compounds and their experimental CA XII inhibitory activities. The quality of prediction is high enough (SE = 0.1291, r^2 = 0.98, F = 212.7398, Q = 0.7963). The virtual molecular fragment that lead to a significant increase of the inhibitor activity of hCA XII is C_2HN_3. The virtual fragments, Br atom and NO_2 leads to a significant decrease of the inhibitor activity value. The innovation of this work consists in not only exploring the structural attributes of bioactive molecules but in predicting in silico the structures of twenty six new compounds which may show Tumor-Associated Carbonic Anhydrases XII (CA XII) inhibitory activity. The analogs of the lead molecule are generated by replacing selected fragments that have similar shape and electrostatics. The molecules of the prediction set include many molecules having high computed activity.

Keywords: Carbonic Anhydrase inhibitors, PRECLAV, Tumor-Associated Carbonic Anhydrases XII, sulfocoumarins.

Introduction

Coumarins such as[1], a natural product secluded from the Australian plant Leionema ellipticum, P. G. Wilson (Rutaceae), or the simple unsubstituted coumarin[2], were detected to act as effective inhibitors of the metalloenzyme carbonic anhydrase (CA, EC 4.2.1.1)[1–3]. Substituted coumarin such as sulfocoumarins (1,2-benzoxathiine-2,2-dioxides) possessing are the most important[4] class of Potent and Isoform-Selective Inhibitors of Tumor-Associated Carbonic Anhydrases CA XII[4].

Known to be a vital feature of the tumor microenvironment, tumor hypoxia is a result of uncontrolled tumor growth outpacing the rate of vascular proliferation and of architecturally defective microcirculation.

The hCA IX and hCA XII are in drudged in the tumor acidification processes, providing H^+ ions to an extracellular milieu by means of the CO_2 hydration reaction to bicarbonate and protons. The pH of tumorsis is in fact more acidic by 0.5–1.0 pH unit than that of the surrounding normal tissue[5], and this acidic environment seems to play a very important role both in the growth, dissemination and propagation of tumor cells and in their no responsiveness to chemo- and radiotherapy[5,6]. As a result, targeting the tumor microenvironment via CA IX and hCA XIII inhibition constitutes an attractive new approach for the administration of hypoxic tumors[7]. The potential use of carbonic anhydrase inhibitors as antitumor agents opens, a new important research direction[5,6].

The goals of our QSAR[8–12] study are the identification of molecular features (significant molecular fragments included) having largest influence on biochemical activity and the estimation of activity for some not yet synthesized molecules in prediction set.

Methods and formula:

Series of with sulfocoumarins was prepared and assayed as inhibitors of carbonic anhydrase (CA, EC 4.2.1.1), CA XII by Supuran *et al*.[4]. In my QSAR study I used, as calibration set, sulfocoumarins 16 derivatives. As prediction set, used 26 molecules having various structures, including sulfocoumarins derivatives. The goals of our QSAR study are the identification of molecular features (significant molecular fragments included) having largest influence on biochemical activity and the estimation of activity for some not yet synthesized molecules in prediction set. The dependent property in our QSAR study is 'activity'. The values of the activities are

K_1 (μM), i.e. A = log (1000/C), the chemical structures and the observed (experimental) activities of the molecules in calibration set are presented in Fig. 1 and Table 1. The data were taken from the literature[2].

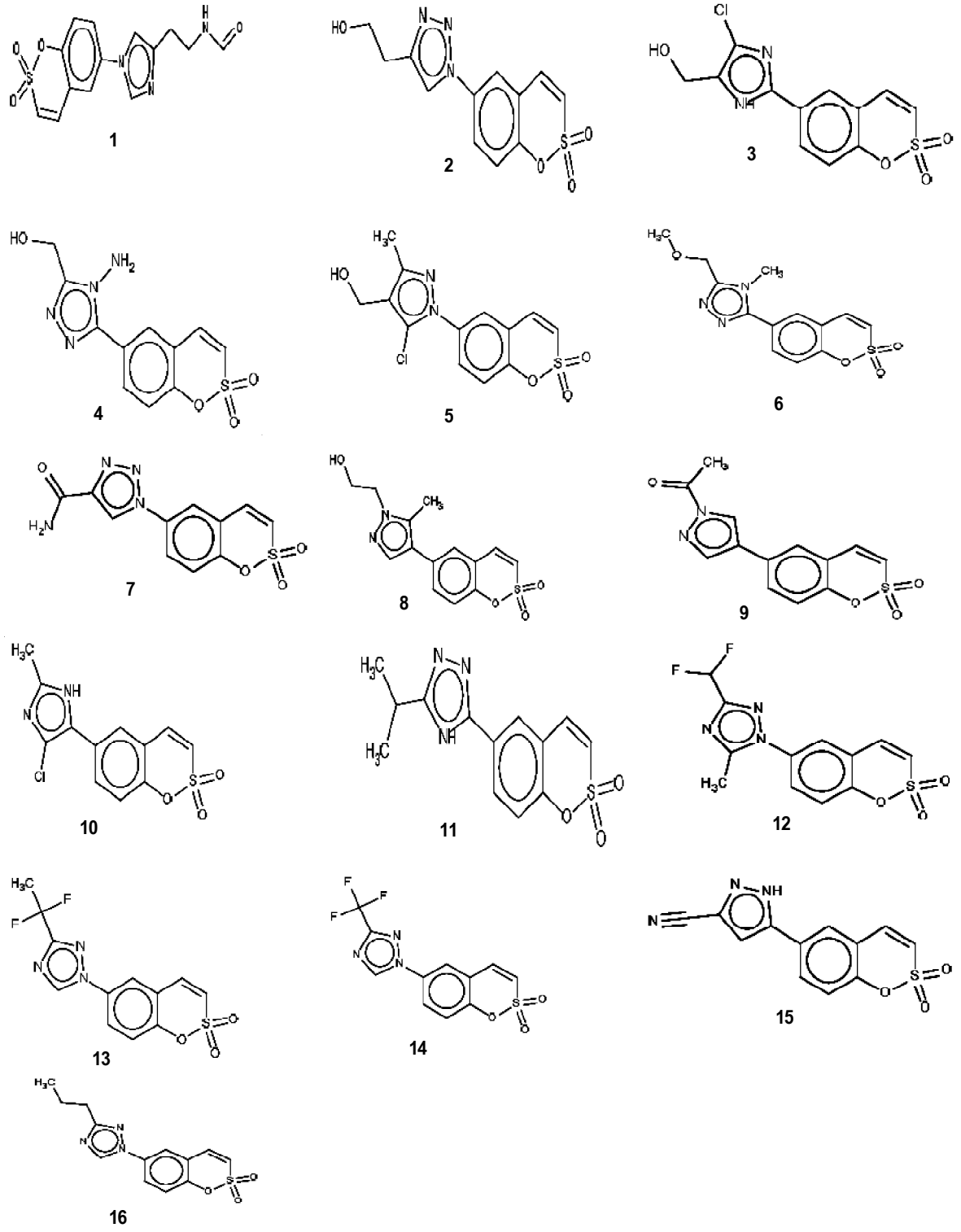

Fig. 1

Table 1. Value of the predictors used in QSAR study of calibration set and prediction set and CA XII activities (in μM and A = log 1000/C), Estimated activities, hat diagonal, Standardized Residual, |*RStudent*| of the calibration set molecules **1-16** with predicted value (A) of the not yet synthesized ones **1-27**

Compd.	Obsd. K_a (μM)	A (Obsd.)	A (Est.)	Residual	|*RStudent*|	Hat diagonal	Compd.	A (predicted value)	Hat diagonal	Compd.	A (predicted value)	Hat diagonal
1	0.234	3.476	3.46	0.016	0.114	0.165	**1**	4.529	0.079	**16**	4.806	0.115
2	0.254	3.44	3.563	–0.123	–0.898	0.113	**2**	4.556	0.083	**17**	4.69	0.156
3*	0.717	2.99	3.165	–0.175	–1.767	0.444	**3**	4.408	0.229	**18**	4.573	0.148
4	4.51	2.191	2.129	0.062	0.867	0.760	**4**	4.583	0.159	**19**	4.38	0.075
5	3.16	2.345	2.345	0	0.866	1	**5**	4.44	0.126	**20**	4.566	0.144
6*	0.023	4.483	4.459	0.024	0.166	0.073	**6**	4.348	0.098	**21**	4.728	0.129
7	0.032	4.34	4.181	0.159	1.176	0.096	**7**	4.385	0.121	**22**	**5.225**	0.379
8	0.012	4.766	4.597	0.169	1.273	0.111	**8**	4.378	0.072	**23**	**5.225**	0.379
9*	0.013	4.731	4.688	0.044	0.305	0.095	**9**	4.246	0.133	**24**	**5.009**	0.670
10	0.009	4.891	5.132	–0.241	–2.049	0.161	**10**	4.291	0.089	**25**	4.22	0.077
11	0.016	4.641	4.727	–0.086	–0.616	0.115	**11**	4.693	0.105	**26**	4.728	0.129
12	0.007	5	4.836	0.164	1.326	0.218	**12**	4.787	0.104			
13	0.014	4.699	4.807	–0.108	–0.860	0.256	**13**	4.627	0.192			
14	0.039	4.254	4.348	–0.094	–0.666	0.099	**14**	4.922	0.149			
15*	0.013	4.731	4.717	0.014	0.103	0.194	**15**	4.698	0.129			
16	0.021	4.523	4.348	0.175	1.318	0.098						

*Molecules of test set.

Prediction set (design of new compounds):

The structure of the molecules in prediction set (not yet synthesized or not assayed) is presented in Fig. 2. The molecules **1-26** are sulfocoumarins derivatives and their activity has been estimated as excellent activity and generated by Brood[13] software. Brood uses the shape and attachment geometry of the query fragment to identify a family of similar fragments.

MOPAC created out files of the molecule; based on the output, the PRECLAV[14] software calculated, for each molecule, more than 1000 whole molecule descriptors, specific to this program.

The program PRECLAV computes type (1) multilinear QSARs.

$$A = C_0 + \sum_{i=1} C_k \bullet D_k \quad (1)$$

where A is (the value of) activity; C_0 is the free term (intercept); C_i are coefficients (weighting factors); D_i are (the value of) significant descriptors; k is the number of descriptors.

The square of Pearson linear correlation r^2 of observed/computed values, the Fisher function F, the standard error of estimation SEE, and the quality function Q (4) are criteria for the quality of prediction for the molecules in calibration set.

$$F = r^2/(1 - r^2) \bullet (N - p)/p \quad (2)$$

$$SEE = [(\Sigma\Delta^2)/(N - 1)]1/2 \quad (3)$$

$$Q = r^2 \bullet (1 - p/N) \quad (4)$$

where p is number of descriptors; N is number of molecules in the calibration set; Δ is difference.

The descriptors included in the best (by Q function) QSAR are named 'predictors'. The relative utility of predictors is computed by the formula (5).

$$U = (R^2 - r^2)/(1 - r^2) \quad (5)$$

where R^2 is the square of Pearson correlation between the observed values and the computed values (using p predictors) r^2 is the square of Pearson correlation between the observed values and the calculated values (using the p – 1 predictors, i.e. the QSAR equation without the analyzed predictor).

After computation of U (5) for each predictor, the values of U are normalized by the highest of them (the highest value for U becomes 1000). The predictors with high enough value

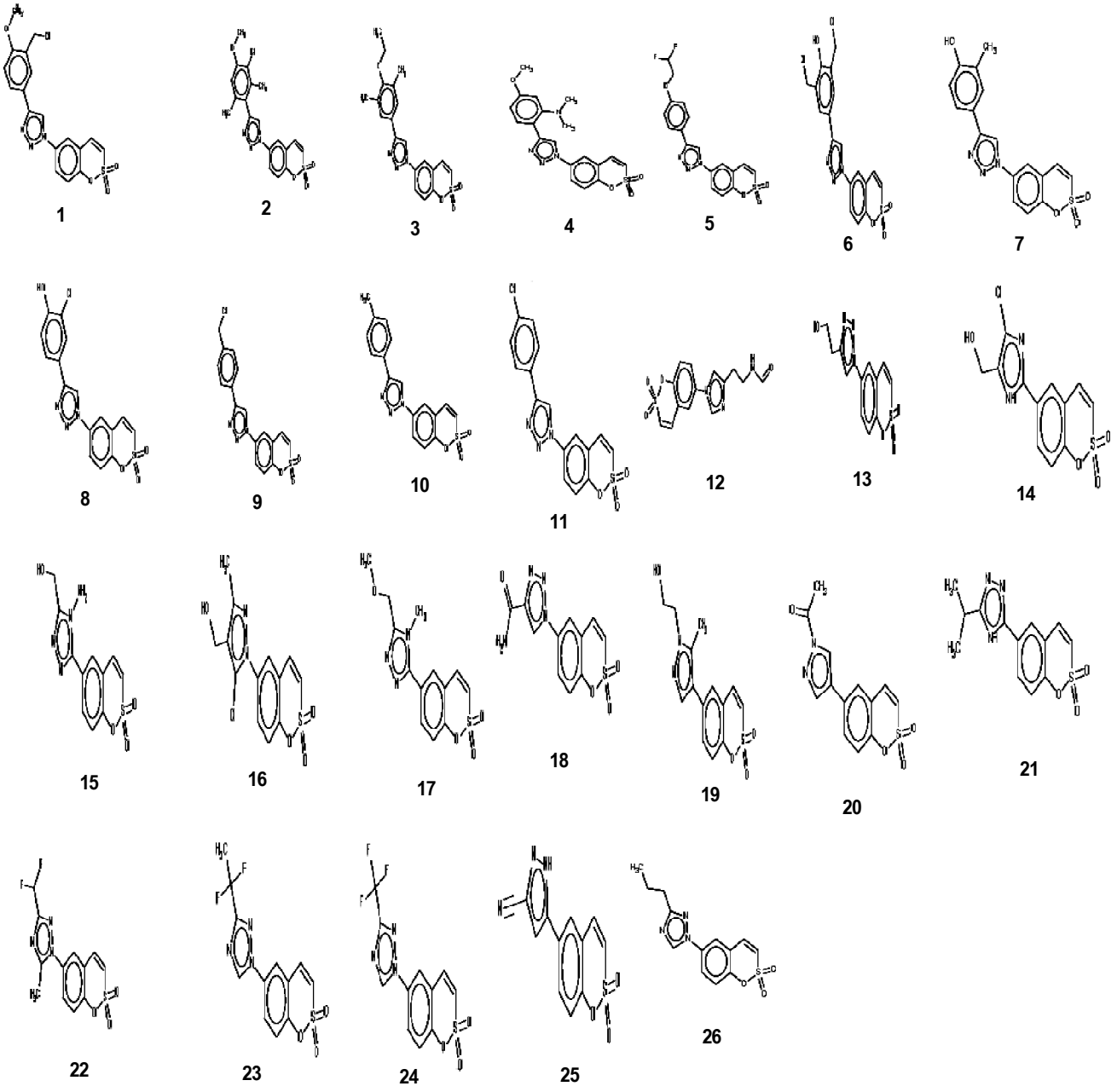

Fig. 2

of U (U > 500) can be considered 'with high relative utility'. PRECLAV (**5**) calculates square of cross-validated correlation r^2_{CV} using LHO (Leave Half Out) method. However, this usual method is applied after ordering of molecules in calibration set according to the observed values of activity. Therefore, the cross-validated function r^2_{CV} is a measure of homogeneity of calibration set from the point of view of predictors' set, i.e. from the point of view of structure-activity relationship. A low value (< 0.4) of r^2_{CV} means 'the QSAR for molecules having high values of activity and the QSAR for molecules having low values of activity include the same descriptors, but very different weighting factors'. Actually, the computation of r^2_{CV} is a very drastic 'internal validation test'. After computing the A_{calc} values of the activity for the prediction set molecules, the program computes the average value $A_{calc}{}^{m}$ and the standard deviations of the estimated values.

The program considers 'high values' the values fulfilling the criterion (6) and 'low values' the values fulfilling the criterion (7).

$$A_{calc} > A_{calc}^{m} + 0.5 \bullet s \quad (6)$$

$$A_{calc} < A_{calc}^{m} - 0.5 \bullet s \quad (7)$$

Applicability of domain and detection of outliers:

A QSAR model can be used for selection new compounds if its domain of application is defined. The need to exemplify the model applicability domain is also reflected in the OECD guidelines for QSAR model validation[15,16]. QSAR model should only be used for making predictions of compounds fall within the specified domain may be considered reliable[12]. Extent of extrapolation is one simple approach to define the applicability of the domain. It is based on the calculation of the hat diagonal (leverage) h_i for each chemical, where the QSAR model is used to predict its activity:

$$h_i = ¼x_i^T(X^TX)^{-1}x_i \quad (9)$$

In eq. (9), x_i is the descriptor-row vector of the query molecule and X is the $k \times n$ matrix containing the k descriptor values for each one of the n training molecules. A hat diagonal (leverage) value $>3(k + 1)/n$ leverage warning limit is considered large. To visualize the applicability of domain of a developed QSAR model, William plot was used. In the William plot, |*RStudent*| versus leverage values (h_i) are plotted. This plot could be used for an immediate and simple graphical detection of both the response outliers and structurally influential compounds in a model. It must be noted that compounds with high value of leverage and good fitting in the developed model can stabilize the model. On the other hand, compounds with bad fitting in the developed model may be outliers. Thus, combination of leverage and the |*RStudent*| could be used for assigning the applicability of domain.

Results and discussion

The statistical computations were conducted using the specific formulas and procedures of PRECLAV program algorithm. Using only the "significant" descriptors PRECLAV computes ten thousand QSAR type (1) multilinear equations. The quality of the obtained equation is reflected by the value of the *Q* function and also by values of some usual statistical functions. During the PRECLAV MLR analysis, I observed that the 3-parametric model has the highest value of the *Q* function for hCA XII inhibitors and also has the highest predictive power as follows:

Dependent property: Inhibition constant (*A*) for hCA XII

Molecules number in calibration set: 16

Number of "significant" descriptors in presence of prediction set = 261

A = 2.6381 – 0.5271 (mam) – 1.593 (lco) + 4.5288 (dva)

Average atomic mass (U = 1000); lco; number of coordinative bonds (U = 931); D3 = dva; Shannon entropy of atomic numbers (U=976)

r^2 = 0.98 , F = 212.7398, r^2_{CV} = 0.966 , SEE = 0.1291

SEE = standard error of estimation, r^2 = Pearson square correlation, F = Fisher function, r^2_{cv} = Pearson cross validated square correlation (Leave one out method).

According to algebraic sign of coefficients in QSAR formula and the value of utility *U* the main factor in influence on activity value is the average atomic mass parameter) mam so increase the Average atomic mass parameter descriptor of compound decrease the activity and because of this most active compound **12** is highest value of lowest value mam. Number of coordinative bonds; dva; Shannon entropy of atomic numbers also play effective role on activity.

Significant molecular fragments:

Fragment atoms	Specimen compounds	Correlation
C_2HN_3	8	0.814
Br atom	4	–0.5556
NO_2	5	–0.5105

The fragments C_2HN_3 is favorable to inhibitory activity and Br and NO_2 atom not favorable to activity.

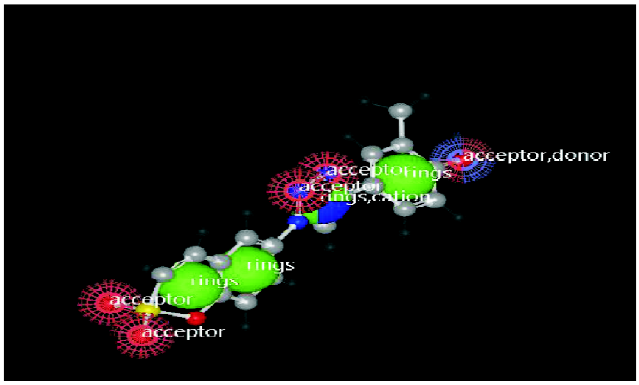

Fig. 3

I have developed a computer representation of the pharmacophore model; this also includes information on the available space at important substituent positions. Fig. 3 represent pharmacophore models with most active compound (**13**) which is generated by Brood. The model displays seven pharmacophore elements (three hydrogen bond donors and four hydrogen bond acceptors) which are used to develop and describe the interaction between ligands and the target receptor from the ligand point of view.

The highest four values in Table 1 are 'high' values according to the formula (6) and the smallest six values are 'low' values' according to the formula (7). If the molecules in the prediction set are not yet synthesized molecules, the molecules having 'high' estimated value are, as a rule, 'recommended for synthesis'. The molecules having 'low' estimated value are not 'recommended for synthesis'. In order to confirm our findings I have compared the estimated values of the activities with the experimental (observed) ones (Table 1). This has further been demonstrated in Fig. 2; a linear relationship between observed and estimated activities in a scatter plot indicates that linearity assumption is appropriate. I observed that the estimated activities are very close to the experimental activities.

External validation of the computation method:

In this work, the molecules with rank **3**, **6**, **9** and **15** for QSAR study constituted of the validation set and the remaining molecules formed the reduced calibration set. The validation set of 4 molecules (22% of the database) captured all the features and spanned the activity range of the entire dataset. The remaining 13 molecules formed the reduced calibration set. In the case when there is a validation set, the most important tool is the correlation between the estimated and experimental values of QSAR equation for the molecules in the validation set. In the presence of the validation set, I obtained the three parametric models for the reduced calibration set (for 13 molecules) with the predictors used in the above QSAR study and obtain results:

$r^2 = 0.98204; F = 145.81361; Se = 0.15211; r^2_{pred} = 0.94178$

Hence, I can state that the estimated value for the molecules in the validation set are close to the experimental ones and have ordered the molecules in a series alike sufficient to the actual CA XII activity value. This was confirmed by graph (Fig. 5) between observed and estimated value of calibration set and validation set. The predictive r^2 ($r^2_{pred} > 0.5$) parameter indicates significant ability of the developed model to predict the CA XII activity (log inhibition constant) of new compounds.

Applicability domain:

As discussed earlier, I used |*RStudent*| of observed inhibitory activity calculated by the obtained models and hat diagonal (leverage) for assigning applicability of domain (AD). Values of leverage could be calculated for both calibration set and prediction set compounds shown in Table 1. Applicability of domain for the developed model is shown in William plot (Fig. 4). Influential compounds are points with leverage value higher than the warning leverage limit. It can be seen

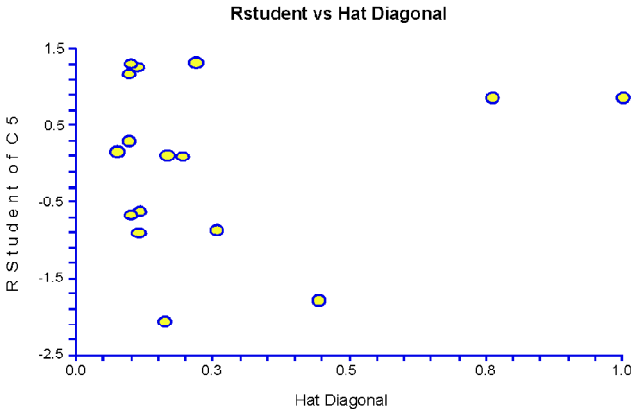

Fig. 4. |*RStudent*| of observed vs hat diagonal applicability domain.

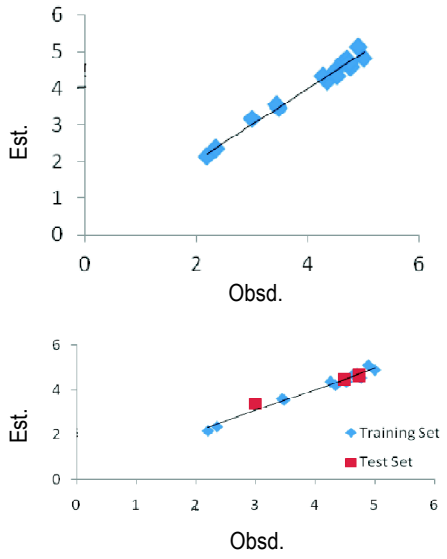

Fig. 5. Observed versus estimated inhibitory activity (A) of calibration set and training and test set.

in the William plot; all molecules in calibration set lie in the application domain of the developed model. None of the molecules have a hat diagonal (leverage) value higher than warning leverage limit (0.75) except compound **5** but |*RStudent*| is within limit so it is not considered to outliers. and also none of the molecules have higher |*RStudent*| than threshold limit |*RStudent*| < 2 except compound **5** but hat diagonal is within limit so it is not considered to outliers.

Conclusions

The virtual fragment C_2HN_3 favorable to the inhibitory activity. In calibration set; the increase the average atomic mass parameter descriptor of compound causes the decreases the inhibitory activity so this descriptor has greater influence on inhibitory activity value. The homogeneity of the calibration set, in fact the similarity of the molecules from the point of view of structure-activity relationship, seems to be 'good'. Many molecules in proposed prediction set have much higher computed activity than observed value.

Acknowledgement

This article is dedicated to the memory of the late Professor Padmakar V. Khadikar (1936-2012).

References

1. Vu. H. Pham and N. B. Quinn, *J. Biomol. Screening,* 2008, **13**, 265.
2. A. Maresca, C. Temperini, L. Pochet, B. Masereel, A. Scozzafava and C. T. Supuran, *J. Med. Chem.*, 2010, **53**, 335.
3. A. Maresca and C. T. Supuran, *Bioorg. Med. Chem. Lett.,* 2010, **20**, 4511.
4. T. V. Kaspars, K. Daniela, A. Leitans, J. A. Lends, A. R. Grandane, A. Scozzafava and C. T. Supuran, *J. Med. Chem.*, 2013, **56**, 293.
5. S. Pastorekova and J. Pastorek, "Cancer-Related Carbonic Anhydraselsozymes", in: 'Carbonic Anhydrase – Its Inhibitors and Activators', eds. C. T. Supuran, A. Scozzafava and J. Conway, CRC, Boca Raton, FL, USA, 2004.
6. C. C. Wykoff, N. J. Beasley, P. H. Watson, K. J. Turner, J. Pastorek, A. Sibtain, G. D. Wilson, H. Turley, K. L. Talks, P. H. Maxwell, C. W. Pugh, P. J. Ratcliffe and A. L. Harris, *Cancer Res.*, 2000, **60**, 7075.
7. A. Thiry, J. M. Dogne, B. Masereel and C. T. Supuran, *Pharmacol. Sci.*, 2006, **27**, 566.
8 S. Singh, *Med. Chem.,* 2012, **8**, 656.
9. S. Singh, *J of Genet Disord and Genet Med.*, 2017, **1**.
10. S. Singh, *J. Indian Chem. Soc.,* 2017, **94**, 1.
11. S. Singh, *J. Enzyme Inhibit. Med. Chem.*, 2016, **31**, 417.
12. S. Singh, *Bioorg. Med. Chem. Lett.*, 2015, **25**, 133.
13. Brood (version 2.0.0), Open Eye Science Software, 3600 Cerrillos Road, Suite 1107, Santa Fe, USA, 2010.
14. PRECLAV and DESCRIPT programes are available from Center of Organic Chemistry-Bucharest-Romanian Academy, ltarko@cco.ro; tarko_laszlo@yahoo.com.
15. Organization for Economic Co-operation and Development Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship QSAR Models. 2007, OECD Document ENV/JM/MONO (2007)2. 47.
16. A. Tropsha, P. Gramatica and V. K. Gombar, *Comb. Sci.,* 2003, **22**, 69.