vLife
adding value to life

Group Based QSAR (GQSAR): A Novel Ligand Design Tool for Medicinal Chemist

By

Dr. Subhash Ajmani
VLife Sciences Technologies Pvt Ltd.

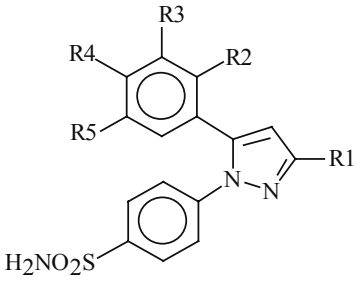
Objective

To perform G-QSAR analysis on the series of substituted 1,5-diphenylpyrazole Cox-2 inhibitors (NSAID) [1]. In the present study a conventional 2D QSAR analysis from whole molecular descriptors was also done and compared with the corresponding group based QSAR analysis.

Method

GQSAR is a novel group based QSAR approach recently developed by our group. As the name suggests fragmentation of the molecules to define various molecular sites (to be considered in the study) is a prerequisite for G-QSAR analysis. And hence allows to establish a correlation of chemical group variation at different molecular sites of interest with the biological activity. This method also includes interaction of various molecular sites (as cross terms) to analyze its effects in determining the biological activity.

Based on various groups at different substitution sites on common fragment of 1,5-diphenylpyrazole several group based descriptors were evaluated and work sheet was populated for GQSAR analysis. We have used 25 molecules as training set and 5 molecules as test set as described in the original paper [1].



Structure of Cox-2 Inhibitors

1 Descriptor calculation

For QSAR analysis various 2D descriptors (a total of 296) like element counts, molecular weight, molecular refractivity, logP, topological index, electro-topological index, Baumann alignment independent topological descriptors etc. were calculated using VLifeMDS software [2].

In the G-QSAR approach, the descriptors of the molecules (same as in QSAR) were calculated for various groups present at different substitution sites. The removal of the invariable descriptors resulted in a total of 105 descriptors for each group which can be used further. The datasheet prepared with group descriptors (105) at different (5) substitution sites resulted in total 279 descriptors instead of 525 (105x5) after removing the invariable columns. Since the same descriptors are calculated for various groups at different sites the following nomenclature is used for naming a descriptor at a particular position for e.g. R1_Mol.Wt. represents the molecular weight of the group present at R1 substitution site. Following formula was used for calculation of interaction/cross terms of the various group descriptors at different substituent sites e.g.:

R3_slogp*R4_Mol.Wt. = R3_slogp x R4_Mol.Wt.

Where, R3_slogp corresponds to value of slogP of the group at R3 substitution site and R4_Mol.Wt. is the value of molecular weight of the group at R4 substitution site.

The calculation of the interaction/cross terms resulted in ~5000 descriptors after removing the invariable descriptors.

2 Model Building

2.1 QSAR model

At first all the descriptors were subjected to stepwise forward multiple linear regression analysis which resulted in a 5 descriptors model with significant r^2 (0.93) and q^2 (0.87) but poor external prediction r^2 (pred_r^2 = -0.32) (regression

data not shown). Then PLS regression method was applied on selected set of 8 descriptors which resulted in a statistically significant model with 6 PLS components as reported in table 1. The pair-wise inter-correlation (r^2) between descriptors in the model is less than 0.5 indicating that chosen descriptors are not highly dependent.

2.2 Group based QSAR (G-QSAR) model

The stepwise multiple linear regression analysis resulted in a significant G-QSAR model with 5 descriptors. The descriptors and the statistical parameters of the model are reported in table 1. The pair-wise inter-correlation (r2) between descriptors in the model is less than 0.5.

2.3 Group based QSAR with interaction terms (G-QSAR_IT) model

At first various variable selection methods like stepwise forward, forward backward and simulated annealing coupled with multiple regressions were applied for building model, which resulted in various models with good statistical parameters for training and internal validation but poor external validation parameters. Then PLS was applied on a selected set of 12 descriptors which includes both group based and interaction term descriptors and that resulted in a statistically significant G-QSAR model with 4 PLS components as reported in table 1. The pair-wise inter-correlation (r^2) between descriptors in the model is less than 0.6. Figure 1 shows plot of observed versus predicted activity by G-QSAR_IT model.

Results and Discussion

The above study resulted in better QSAR, G-QSAR and G-QSAR_IT models using simple 2D descriptors as compared to the various 3DQSAR models reported in the paper [5]. It can be seen from the table 3 that both the G-QSAR and G-QSAR_IT models are comparable/better than the conventional QSAR method.

The conventional QSAR model (table 3) is statistically significant and indicates the significance of basic molecular properties such as hydrogen bond acceptor counts, hydrogen bond donor counts, log partition coefficient (slogP) etc., however it does not show the site where variation is required leading to difficulty in interpretation. In order to get better insights of group descriptors important in explaining variation of activity G-QSAR and G-QSAR_IT models were developed.

It can be seen from the table 3 and figure 3 that substitution sites R2, R3 and R4 were found to be playing major role (R4 being the most important) in G-QSAR and G-QSAR_IT model and this in line with the amount of variation in chemical substitution at the various substitution sites. The R1 and R5 site descriptors do not appear in models since the variation of groups at those sites are not significant.

Although the descriptors appearing in QSAR and G-QSAR models are not the same, they share a similarity in terms chemical nature, e.g. hydrogen bond acceptor counts in QSAR models is similar to Oxygen counts in G-QSAR and G-QSAR_IT models. From cross terms it could be seen that R3 and R4 group interactions are mainly responsible for activity variation. It is the combination of hydrophobicity at R3 and bulk/ hydrophobicity of substituent at R4 which influence the activity. Unlike conventional QSAR models, new models also provide information about the important substitution site(s) along with their chemical nature. This information could provide useful indication for design of new molecules.

Summary

In the present study we have demonstrated application of partitioning of molecular descriptor information into the substituent group based descriptors for performing GQSAR analysis. In addition, we have shown the role of cross terms (i.e product of group based descriptors) in the improvement of G-QSAR models. This combination of methods allows a better interpretation of the models in terms of the contribution of each individual substituent site and/or their interactions.

The methodology was used for dataset of Cox-2 inhibitors by evaluating simple 2D descriptors to generate QSAR, G-QSAR and G-QSAR_IT models using multiple regression and partial least squares regression methods. All the developed models were found to have better predictive ability than the earlier reported 3D-QSAR models using

CoMFA, MFA, RSA approaches [1]. However, these models are comparable to the 3D-QSAR models generated using k-Nearest Neighbor Molecular Field Analysis (k-NNMFA) [3].

Finally, the G-QSAR methodology provides an approach for better understanding of the structure activity relationship both in terms of identifying important chemical variations at specific substitution sites and also by providing quantitative model for prediction of activities of the new designed molecules.

References

[1] Desiraju, G. R.; Gopalakrishnan B.; Jetti, R. K. R.; Raveendra, D.; Sarma, J. A. R. P.; Subramanya, H. S., Molecules 2000, 5, 945-955.

[2] VLifeMDS: Molecular Design Suite developed by VLife Sciences Technologies Pvt. Ltd., Pune, India 2006.

[3] Ajmani, S.; Jadhav, K.; Kulkarni, S.A. J.Chem. Inf. Mod. 2006, 46, 24-31.

Table 1: Statistical parameters and descriptors obtained for QSAR, G-QSAR and G-QSAR_IT models for Cox-2 inhibitors

	QSAR	G-QSAR	G-QSAR_IT
components	6	-	3
n (train/test)	25/5	25/5	25/5
descriptors (k)	8	5	10
r2	0.932	0.926	0.93
q2	0.745	0.866	0.88
pred_r2	0.864	0.809	0.89
r2_SE	0.361	0.366	0.35
q2_SE	0.698	0.492	0.45
Zscore_r2	4.641	7.279	5.37
Zscore_q2	3.477	1.817	4.02
best_ran_r2	0.335	0.229	0.44
best_ran_q2	-0.768	-1.196	-1.28
alpha_r2	0.0001	0.0000	0.00
alpha_q2	0.001	0.05	0.00
F-test	41.118	47.551	88.89
Descriptor_1	H-AcceptorCount	R2_SsCH3E-index	R2_Mol.Wt.
Descriptor_2	H-DonorCount	R3_RotatableBondCount	R2_SsCH3E-index
Descriptor_3	slogp	R4_chi2	R4_chi2
Descriptor_4	chiV4pathCluster	R4_OxygensCount	R4_OxygensCount
Descriptor_5	CarbonsCount	R4_SsCH3E-index	R4_SsCH3E-index
Descriptor_6	NitrogensCount		R1_0PathCount*R4_XlogP
Descriptor_7	FluorinesCount		R3_slogp*R4_Mol.Wt.
Descriptor_8	T_2_T_4		R3_HydrogensCount *R4_k1alpha
Descriptor_9			R4_polarizabilityAHC*R4_chi0
Descriptor_10			R4_T_T_T_0*R5_smr

Figure 1: Plot of observed versus predicted activity of Cox-2 inhibitors dataset