Quantitative structure–activity relationship and combinatorial design of 1,3,4-oxadiazole-based thymidine phosphorylase inhibitors as potential anti-cancer agents

Shalini Bajaj, Piyush Ghode, Jagadish Singh, Partha Pratim Roy and Sanmati Kumar Jain*

SLT Institute of Pharmaceutical Sciences, Guru Ghasidas Vishwavidyalaya, Bilaspur 495 009, India

The 3D quantitative structure–activity relationship model representing $r^2 = 0.8605$, $q^2 = 0.8193$ and $\text{pred\_}r^2 = 0.6847$ respectively, was generated for thymidine phosphorylase (TP) inhibitory activity of some 1,3,4-oxadiazole derivatives. Electronegative substituents at R^1 and less steric bulk with electropositive substituents at R^2 were found to be favourable for TP inhibition. The activity prediction of a combinatorial library of 1629 compounds resulted in 50 molecules whose predicted activity was comparable to the most active compound in the dataset and within the model's applicability domain. Among them six molecules showed favourable interactions with the active site of TP proposing potential anticancer activity of the title compounds.

Keywords: Anti-cancer therapy, docking, combinatorial library, 1,3,4-oxadiazole, 3D-QSAR.

DURING the last few decades, target-based drug design for anti-cancer therapy has gained increasing attention. In cancer patients, the level of thymidine phosphorylase (TP) is elevated[1] leading to cell proliferation. The TP inhibitors, through inhibition of 2′-deoxy-D-ribose-1-phosphate, repress vascular endothelial growth factor (VEGF) production which in turn suppresses the formation of new blood vessels through inhibition of metalloproteinase secretion, proliferation, differentiation, angiogenesis and thus averts cancer metastasis[2–4].

The quantitative structure–activity relationship (QSAR) has become an increasingly practicable approach in chemical and biological sciences through establishment of a statistical relationship between molecular features (descriptors) and activities of assorted compounds with similar scaffolds. The postulation of correlation between the descriptors and chemical/biological properties is the crux behind the wide applicability of this approach[5–7].

In continuation of our efforts to explore TP inhibitory potential of 1,3,4-oxadiazole nucleus, a 3D-QSAR study was performed on 3,5-disubstituted-1,3,4-oxadiazole-2-thione derivatives[8] using VLife Molecular Design suite 3.5 (MDS)[9].

Binding mode of ligand to the TP receptor was predicted by docking through Python Prescription (PyRx) using AutoDock Vina[10,11]. Combinatorial library was generated by swapping the substituents of selected series using lead grow tool of VLife MDS.

Materials and methods

The workflow for exploration of new 1,3,4-oxadiazole derivatives through QSAR model development, validation, molecular docking and combinatorial library in the present work is shown in the following flow chart.
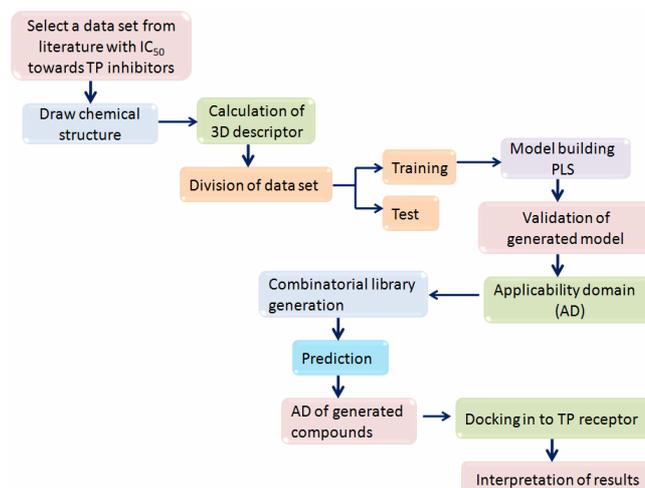
Data mining and preparation

A dataset of 1,3,4-oxadiazole derivatives with concentration inhibiting 50% of the enzyme (IC_{50} value (μM)) was chosen for QSAR study[8] and the reported IC_{50} values were transformed to negative logarithms [$-\log IC_{50}/pIC_{50}$(moles)] (Table 1).

*For correspondence. (e-mail: ks17jain03@gmail.com)

Table 1. Structure and thymidine phosphorylase inhibitory activity of 1,3,4-oxadiazole derivatives

Compound no.	R^1	R^2	Observed activity pIC$_{50}$ (mole)	PLSR model Predicted activity	PLSR model *Residual activity
1	phenyl	2-OCH₃-phenyl	4.059	4.114	−0.055
2	phenyl	3-OCH₃-phenyl	4.391	4.336	0.055
3	phenyl	2-Cl-phenyl	3.929	4.208	−0.279
4	phenyl	2,4-dimethylphenyl	4.391	4.391	0.000
5	3,4,5-trimethoxyphenyl	2-CH₃-phenyl	3.931	3.972	−0.041
6	3,4,5-trimethoxyphenyl	phenethyl	4.159	4.230	−0.071
7	2-OH-phenyl	2-Cl-phenyl	3.884	4.026	−0.142
8	4-CH₃-phenyl	2-OCH₃-5-NO₂-phenyl	4.842	4.850	−0.008
9	4-CH₃-phenyl	4-CH₃-phenyl	4.262	4.140	0.122
10	4-CH₃-phenyl	phenyl	3.982	4.008	−0.026
11	4-CH₃-phenyl	2,3-dimethylphenyl	4.625	4.494	0.131
12	4-CH₃-phenyl	3,4-dichlorophenyl	3.874	3.862	0.012
13	4-CH₃-phenyl	2,4-dimethylphenyl	4.08	4.200	-0.120

(*Contd*)

Table 1. (*Contd*)

Compound no.	R^1	R^2	Observed activity pIC$_{50}$ (mole)	PLSR model	
				Predicted activity	*Residual activity
14	H$_3$C— (phenyl)	H$_3$CO (phenyl)	4.074	3.994	0.080
15	Br (phenyl)	Cl (phenyl)	4.661	4.713	−0.052
16	Br (phenyl)	Br— (phenyl)	4.58	4.452	0.128
17	Br (phenyl)	OCH$_3$ (phenyl)	4.754	4.699	0.055
18	Cl— (phenyl)	CH$_3$ (phenyl)	4.235	4.033	0.202
19	Cl— (phenyl)	Br— (phenyl)	3.761	3.760	0.001
20	Cl— (phenyl)	Cl (phenyl)	4.334	4.021	0.313
21	Cl— (phenyl)	OCH$_3$ (phenyl)	4.052	4.011	0.041

*Residuals = Obs. pIC$_{50}$ − Pred. pIC$_{50}$.
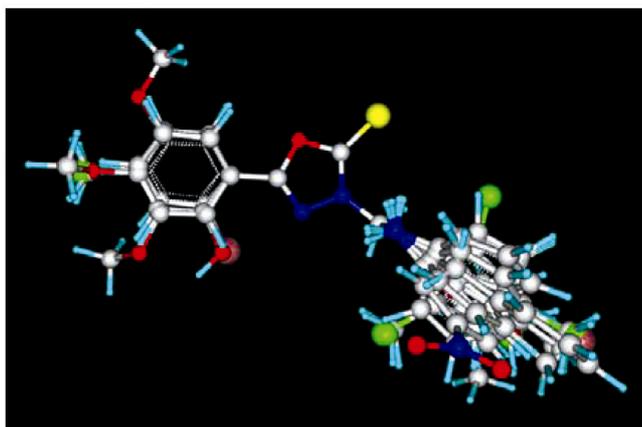
Figure 1. Template based alignment of all the 1,3,4-oxadiazole derivatives.

All the compounds were energy-minimized with a dielectric constant of 1, 10,000 number of cycles, modified Q_{eq} equilibrium charge method based on atomic electronegativity and 0.01 kcal/mol Å as root mean square gradient using Merck molecular force field (MMFF)[12]. All molecules of the dataset were aligned using 5-methyl-3-((methylamino) methyl)-1,3,4-oxadiazole-2(3H)-thione template[13]. 3-(((2-methoxy-5-nitrophenyl) amino) methyl)-5-(*p*-tolyl)-1,3,4-oxadiazole-2(3*H*)-thione (compound **8**) was chosen as a reference molecule owing to its highest activity among the series (Figure 1).

Molecular descriptors

Following template-based alignment; the interactions of an SP^3 carbon probe on a rectangular grid provided with

steric, electrostatic and hydrophobic descriptors of molecules[14]. The cut-off values for steric and electrostatic interactions were set to 30 and 10 kcal/mol respectively[15].

Division of the dataset

For internal and external validation of the model, the training and test sets were chosen in the ratio of ~80 : 20%. For the QSAR model to be truly predictive in the range of its descriptor and activity space, the training set must comprise compounds with maximum and minimum activities. Determination of various descriptive statistical parameters facilitated the choice of best representatives of training and test sets[16,17].

Development of statistically significant models

The development of 3D-QSAR models involved partial least squares regression (PLSR) with stepwise forward–backward method (SW FW–BW) for selection of appropriate descriptors[18,19]. The stepwise forward backward regression involves a combination of forward selection and backward elimination of predictive variables (descriptors) in the model until no term is left out to meet the specified statistical significance criteria[20].

PLSR involves finding linear regression using a small number of latent variables derived from a large number of original descriptors[21]. Thus, PLSR is the method of choice in most cases where the predictive variable space is very large as compared to the response variable space. The main advantage of PLSR lies in its ability to find linear relationship even if the variables have very little contribution to the first few principal components[22,23]. The selection of statistically significant models involved determination of squared correlation coefficient (r^2), cross-validated squared correlation coefficient (q^2), predicted correlation coefficient (pred_r^2) and external validation parameters ZScore r^2, ZScore q^2, ZScore pred_r^2.

Molecular docking

Molecular docking was performed on ligand bound TP structure (protein data bank ID: 1UOU). The important residues of TP responsible for the activity are SER117, SER217, HIS116, LYS221, ARG202 and TYR199[24]. The ligands were prepared by computing the charges and setting the number of torsions. The macromolecule was prepared by removal of ligands and water molecules (if any) present in the crystal structure, addition of polar hydrogens and grid generation around the active site. Vina generates multiple conformers for a molecule and the result of docking is expressed in terms of binding affinity (kcal/mol) and a lower scoring conformation is ranked higher. The generated conformers were loaded in PyMOL to visualize the binding modes with the receptor[25,26].

Combinatorial library generation

A library of compounds with different substituents at various sites of title compounds was generated using Leadgrow module of VLifeMDS. For this purpose, different substitutions of the dataset were scrambled to generate new structures. The generated structures were then filtered on the basis of Lipinski's rule[27]: hydrogen bond acceptor (HAc) ≤10; hydrogen bond donors (HDr) ≤5; slogp ≤5; molecular weight (MW) ≤500; rotatable bond (RtB) ≤10; polar surface area ≤140. The activity of generated compounds was predicted by the most significant 3D QSAR model.

Results and discussion

Interpretation of PLSR model

The term selection criteria for model generation were r^2, q^2 and pred_r^2 (Supplementary Table 1). Amongst these models, we discuss the most significant model (eq. (1)) on the basis of its robustness (Table 2)

$$pIC_{50} = -0.0724\ E\_427 - 21.4035\ S\_737 \\ + 0.3960\ E\_1122 + 4.1317. \qquad (1)$$

Squared correlation coefficient and cross-validated correlation coefficient were found to be 0.86 and 0.82, consecutively explaining the correlation and internal predictive potential of the model. The E_427 descriptor in eq. (1) has a negative coefficient value (–0.0724), implying that more electronegative groups favour TP inhibition. Thus, substitution of more electronegative group, e.g. bromine at R^1 will increase (compounds **17** and **15**) while a less electronegative group methoxy reduces TP inhibitory activity (compounds **6** and **5**). The parameters for steric field (S_737) with negative coefficient and electrostatic field (E_1122) with positive coefficient indicates that less bulky and less electronegative group is favourable on R^2. Figure 2 shows the stereo view of these interactions in the rectangular grid.

The residual values reflect the prediction potential of any model; lower the residual value higher is the statistical reliability of the model. The model indicates very low values of residuals (less significant differences) (Table 1). Figure 3 presents the fitness plot between experimental and predicted activities. Figure 4 depicts the contribution of all three descriptors towards the activity.

Model validation

The internal validation was performed to check the robustness and external validation was performed to determine the predictive ability of the model.

Leave-one-out cross-validation: Predictivity of the model was checked by predicted residual sum of square (PRESS) and leave-one-out cross-validation r^2 (LOO-q^2). Equation (2) represents the calculation of cross-validated r^2 (q^2)

$$r_{cv}^2 Vq^2 = 1 - \frac{\sum (Y_{obs(train)} - Y_{pred(train)})^2}{\sum (Y_{obs(train)} - \acute{Y}_{train})^2}. \qquad (2)$$

The equation can also be expressed as

$$q^2 = 1 - \frac{PRESS}{\sum (Y_{obs(train)} - \acute{Y}_{train})^2}. \qquad (3)$$

In eqs (2) and (3), $Y_{obs(train)}$ is observed and $Y_{pred(train)}$ is the predicted activities of training and $\acute{Y}_{train}$ indicates mean of activities of training set[28]. $q^2 > 0.5$ indicates the acceptability of the model.

Predicted r^2 (pred_r^2 or $q^2_{(F1)}$): Correlation of observed to predicted activity for the test set was checked by pred_r^2

$$pred_r^2 = 1 - \frac{\sum (Y_{obs(test)} - Y_{pred(test)})^2}{\sum (Y_{obs(test)} - \acute{Y}_{train})^2}. \qquad (4)$$

In eq. (4), $Y_{obs(test)}$ is observed and $Y_{pred(test)}$ is the predicted activity of the test set and $\acute{Y}_{train}$ is the mean activity of the training set. pred_$r^2 > 5$ indicates acceptable predictivity of the model.

Y-scrambling/Y-randomization: The *Y*-scrambling method was used to rule out any likeliness of possible chance correlation between predictive and response variables of the model. Random correlation is calculated by Z_{Score}[29] using the following formula

$$Z_{score} = (h - \mu)/\sigma. \qquad (5)$$

In eq. (5), *h* represents q^2 for the model and μ and σ represent average q^2 and standard deviation respectively for models developed through random datasets. The values of all validation parameters is shown in Table 2.

Applicability domain: Applicability domain (AD) is defined as the predictive and response variable space that any model belongs to[30]. Any modelled response predicted
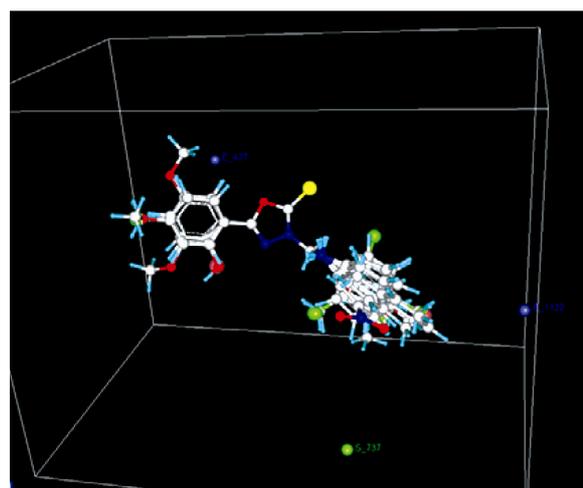
Table 2. Statistical and validation parameters of statistically significant PLSR model

Parameter	PLSR model
Training/test set selection method	Manual
Training set size (percentage)	16 (80%)
Test set size	5
Test set	12,13,16,18,9
Descriptors	E_427
	S_737
	E_1122
Degree of freedom	13
F test	40.1007
Coefficient	−0.0724
	−21.4035
	0.3960
Regression constant	4.1317
r^2	0.8605
r^2_se	0.1360
q^2	0.8193
q^2_se	0.1548
pred_r^2	0.6847
pred_r^2se	0.1473
ZScore r^2	4.44353
ZScore q^2	2.78738
ZScore pred_r^2	1.55711
Best Rand r^2	0.62703
Best Rand q^2	0.46504
Best Rand Pred_r^2	0.5946
Alpha Rand r^2	0.00015
Alpha Rand q^2	0.01
Alpha Rand Pred_r^2	0.1

Figure 2. Stereo view of field points around the molecule generated by 3D PLSR QSAR model.

Figure 3. Fitness plot for the actual versus predicted activities of 3D PLSR model.

through QSAR can be valid only for compounds belonging to the AD of that model and the activity of compounds outside the AD may not be predicted with equal confidence. For a successful PLSR analysis, the predictive variables should follow a normal distribution pattern.
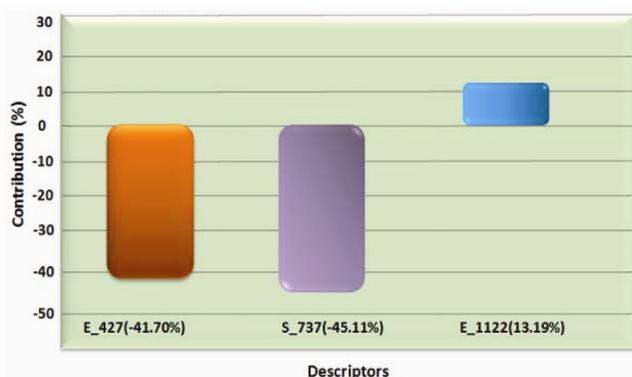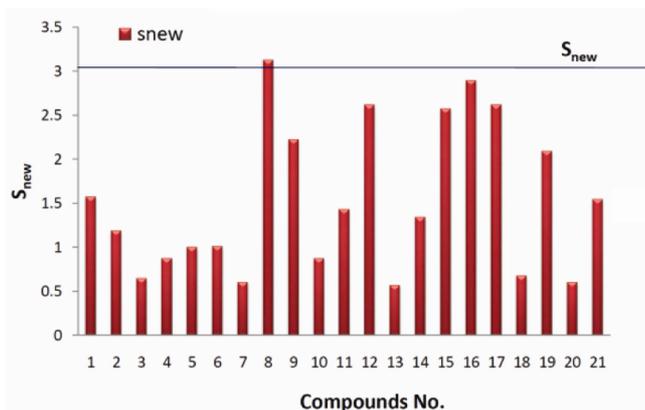

Figure 4. Contributory plot of 3D PLSR model.

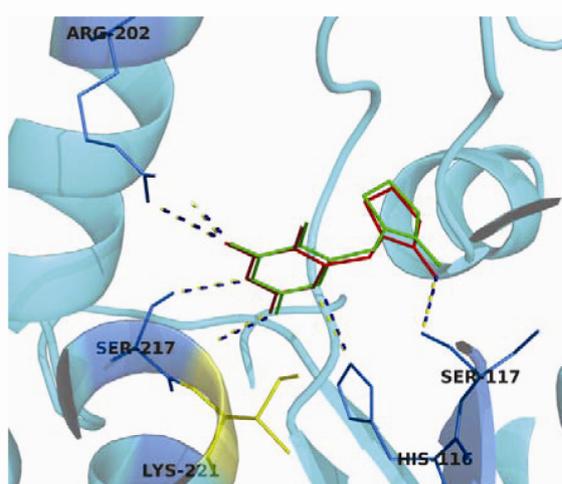**Figure 5.** Applicability domain of 3D PLSR model compounds.

Figure 6. Validation of the docking methodology (cocrystallized ligand: red, docked ligand: green).

It follows that mean ± 3 standard deviations (SD) should cover 99.7% population. Any compound not belonging to this criterion can be considered as an outlier.

Recently, Roy et al.[32] proposed a criterion for estimating AD by measuring the corresponding standardized value for predictive variable i of compound k (S_{ki}). There are three conditions regarding the minimum and maximum values for S_{ki}:

(i) $S_{ki} < 3$ for all the descriptors: The compound is within AD.
(ii) $S_{ki} > 3$ for all the descriptors: The compound is an X-outlier or is outside AD for training and test sets respectively.
(iii) For some descriptors $S_{ki} < 3$ and for others $S_{ki} > 3$: The compound shows similarity with most of the compounds according to some descriptors but dissimilarity to most according to others.

It calls for formulation of some measurement criterion for the third group of compounds. For instance, if we have the standard score (Z) = 1.28, it represents that 90% of the observations will occur below 1.28∗SD. Thus a probability of 90% of any compound not being an X-outlier calls for a sum of 1.28∗SD and mean of S_i values of all the
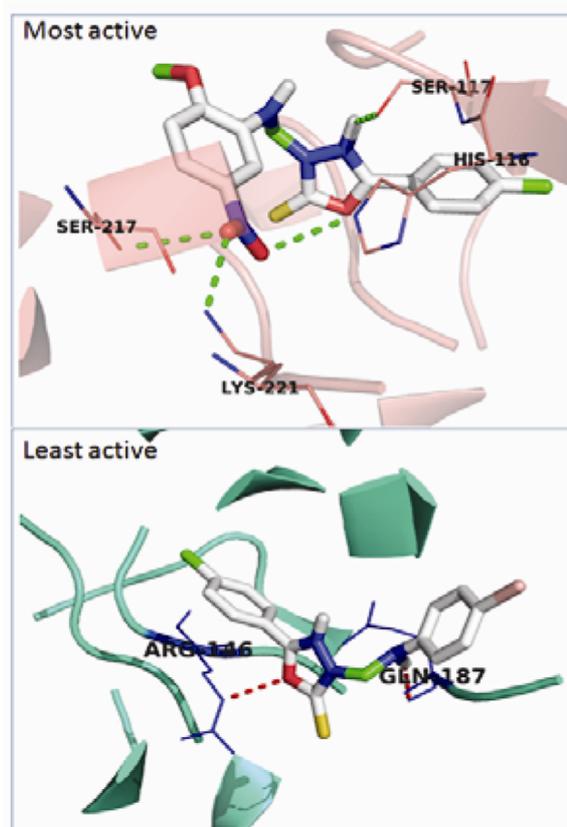

Figure 7. Docking of the most active and least active compound to active site of TP.

Table 3. Binding affinity of docked compounds

Ligand	Binding affinity (kcal/mol)	Binding residues
Crystallographic ligand	−8.2	SER117, SER217, LYS221, ARG202, HIS116
Most active comp. (8)	−7.3	SER117, SER217, LYS221, HIS116
Less active comp. (19)	−7	ARG146, GLN187
Combi lab comp. 66	−8.6	SER117, LYS115, GLY152, TYR199, SER126
Combi lab comp. 67	−8.5	SER117, LYS115, TYR199, GLY152, SER126
Combi lab comp. 138	−6.7	SER117, SER217, TYR199, LYS115, GLY152, SER126
Combi lab comp. 1279	−7.6	SER117, SER217, TYR199, LYS115, SER126
Combi lab comp. 1922	−6.1	SER117, SER217, LYS221, THR154, SER126
Combi lab comp. 2066	−6.6	SER117, SER217, LYS221, SER126, THR154
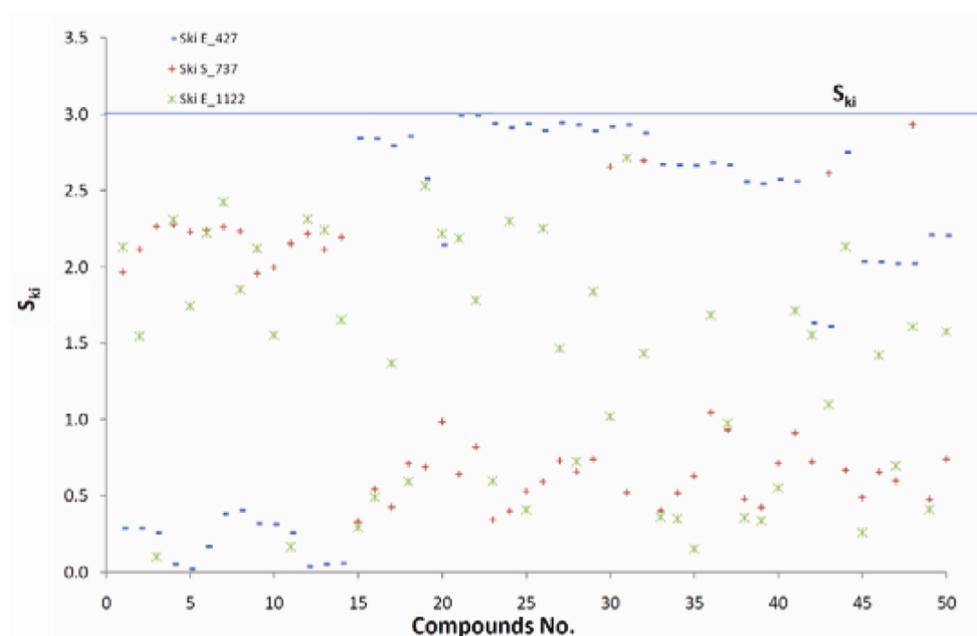
Figure 8. Applicability domain of compounds obtained through combinatorial library generation.

descriptors to be below 3. We can call it S_{new}. The values of S_{ki} and $S_{new(k)}$ were calculated using the formula

$$S_{ki} = (|X_{ki} - \acute{X}_i|)/\sigma_{Xi}. \tag{6}$$

$$S_{new(k)} = \acute{S}_k + 1.28 \times \sigma_{Sk}. \tag{7}$$

In eq. (6), k is the total number of compounds, i the total number of descriptors, S_{ki} the standardized descriptor i for compound k (from the training or test set), $\acute{S}_{ki}$ the original descriptor i for compound k (from the training or test set), $\acute{X}_i$ the mean value of the descriptor X_i for the training set compounds only and σ_{Xi} the standard deviation of the descriptor X_i for the training set compounds only. In eq. (7), $S_{new(k)}$ is the S_{new} value $\acute{S}_k$ is the mean of $S_{i(k)}$ values and σ_{Sk} is the standard deviation of $S_{i(k)}$ values of the compound k.

Compound **8** of the training set shows S_{ki} and $S_{new(k)}$ more than 3, therefore compound **8** may be considered as the outlier (Figure 5).

Analysis of binding mode by molecular docking

All the compounds were subjected to molecular docking studies to explore the binding site interactions with TP. The co-crystallized ligand was redocked into the enzyme binding pocket for validation of the docking protocol. Vina was able to generate the binding interactions of the docked pose identical to that of the co-crystallized ligand (Figure 6). In order to search for the optimum protein–ligand interaction, a set of compounds had been docked to the protein active site. Few representatives especially the highest and the lowest active compounds were studied for their interactions with TP as shown in Figure 7. As expected, all the compounds exhibited hydrogen bond interactions in the binding pocket. The most active compound (compound **8**) forms hydrogen bond closely with SER117, SER217, LYS221 and HIS116. Compound **19** that exhibited low inhibition potency towards TP as evidenced by experimental and 3D-QSAR studies, reflects less binding affinity (ARG146, GLN187). Binding
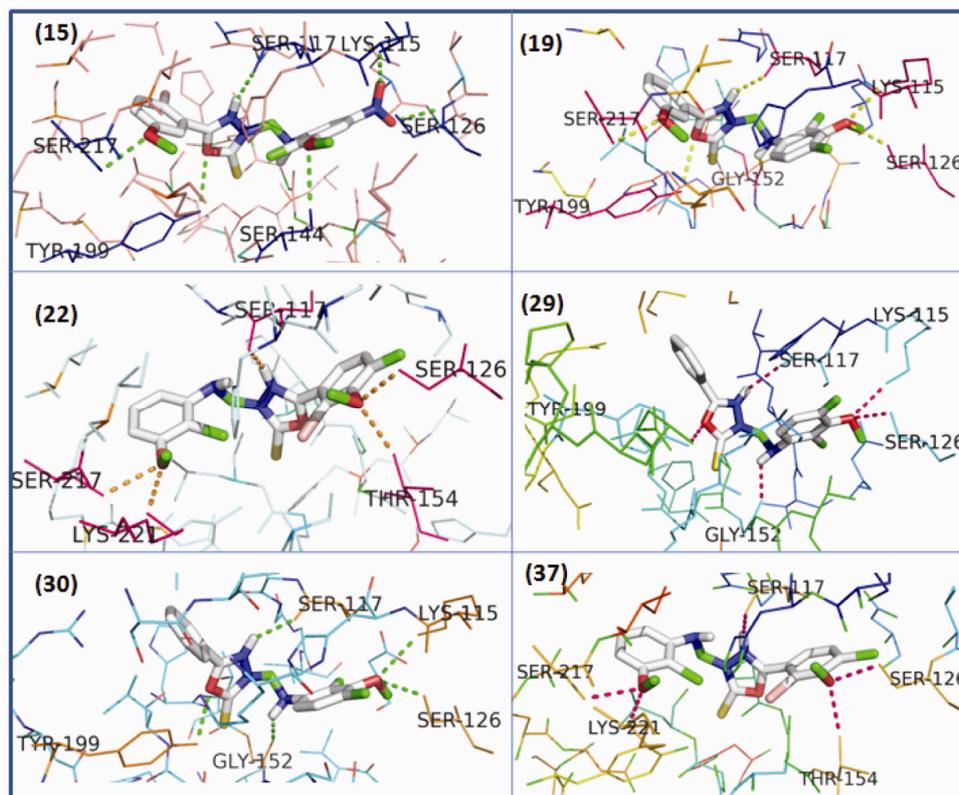
Figure 9. Docking of combinatorial library compounds to active site of TP.

affinities of docked compounds are presented in Table 3.

Combinatorial library

Combinatorial library was generated using the substituents of the model's training set. The generated compounds were filtered on the basis of Lipinsky's rule of five. Amongst the 1629 generated compounds, 696 compounds were within AD according to the standardization approach (Figure 8). Therefore the predictions of these compounds through the model can be regarded as reliable. Among these, 50 compounds were found to have biological activity compared to or greater than the most active compound of selected series (Supplementary Table 2). To establish the pre-eminence of molecules over the dataset compounds, they were docked in the TP receptor, amongst which docking of a few compounds is shown in Figure 9 and their binding affinities are presented in Table 3.

Conclusion

An attempt was made for identification of crucial structural features of 5-substituted-1,3,4-oxadiazole-2-thione derivatives for effective TP inhibition is made. The

QSAR analysis advocates the substitution of electronegative groups around R^1, and less bulky and less electronegative group around R^2 for enhanced biological activity. Combinatorial library was generated by scrambling the substituents followed by activity prediction of the new compounds. Docking study helped in predicting the ligand–receptor interaction of title compounds. The compounds with promising predicted biological activities were docked to TP active site to deduce their binding pattern. On that account, the assumptions of this study may be instrumental in finding new anti-cancer leads with high selectivity and potency.

1. Pauly, J. L., Schuller, M. G., Zelcer, A. A., Kirss, T. A., Gore, S. S. and Germain, M. J., Identification and comparative analysis of thymidine phosphorylase in the plasma of healthy subjects and cancer patients. *J. Natl. Cancer Inst.*, 1977, **58**, 1587–1590.
2. Matsushita, S. *et al.*, The effect of thymidine phosphorylase inhibitor on angiogenesis and apoptosis in tumors. *Cancer Res.*, 1999, **59**, 1911–1916.
3. Focher, F. and Spadari, S., Thymidine phophorylase: a two-face janus in anticancer chemotherapy. *Curr. Cancer Drug Targets*, 2001, **1**, 141–153.
4. Sivridis, E., Giatromanolaki, A., Papadopoulos, I., Gatter, K. C., Harris, A. L. and Koukourakis, M. I., Thymidine phosphorylase expression in normal, hyperplastic and neoplastic prostates: correlation with tumour associated macrophages, infiltrating lymphocytes, and angiogenesis. *Brit. J. Cancer*, 2002, **86**, 1465–1467.

5. Ferreira, M. M. C., Multivariate QSAR. *J. Braz. Chem. Soc.*, 2002, **13**, 742–753.

6. Abdullahi, A. D. *et al.*, Novel insight into the structural requirements of P70S6K inhibition using group-based quantitative structure activity relationship (GQSAR). *J. Appl. Pharm. Sci.*, 2014, **4**(06), 16–24.

7. Jain, S. V., Bhadoriya, K. S., Bari, S. B., Sahu, N. K. and Ghate, M., Discovery of potent anticonvulsant ligands as dual NMDA and AMPA receptors antagonists by molecular modelling studies. *Med. Chem. Res.*, 2012, **21**, 3465–3484.

8. Shahzad, S. A. *et al.*, Synthesis and biological evaluation of novel oxadiazole derivatives: A new class of thymidine phosphorylase inhibitors as potential anti-tumor agents. *Bioorganic Med. Chem.*, 2014, **22**, 1008–1015.

9. VLifeMDS 3.5, Molecular Design Suite, VLife Sciences Technology Pvt Ltd, Pune, India, 2009; www.vlifesciences.com

10. Trott, O. and Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J. Comp. Chem.*, 2010, **31**, 455–461.

11. Dallakyan, S. and Olson, A. J., Small-molecule library screening by docking with PyRx. *Meth. Mol. Biol.*, 2015, **1263**, 243–250.

12. Halgren, T. A., Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comp. Chem.*, 1996, **17**, 553–586.

13. Ajmani, S., Jhadav, K. and Kulkarni, S. A., Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J. Chem. Inf. Model.*, 2006, **46**, 24–31.

14. Ghosh, P. and Bagchi, M. C., Comparative QSAR studies of nitrofuranyl amide derivatives using theoretical structural properties. *Mol. Simul.*, 2009, **35**(14), 1185–1200.

15. Clark, M., Cramer, R. D. and Van, O. N., Validation of the general purpose Tripos 5.2 force field. *J. Comp. Chem.*, 1989, **10**, 982–1012.

16. Golbraikh, A. and Tropsha, A., Predictive QSAR modeling based on diversity sampling of experimental datasets for training and test set selection. *J. Comp. Aided Mol. Des.*, 2002, **16**, 357–369.

17. Sahu, N. K., Sharma, M. C., Mourya, V. and Kohli, D. V., QSAR studies of some side chain modified 7-chloro-4-aminoquinolines as antimalarial agents. *Arabian J. Chem.*, 2014, **7**, 701–707.

18. Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P., Optimization by simulated annealing. *Science*, 1983, **220**, 671–680.

19. Scior, T., Medina-Franco, J., Do, Q. T., Martínez-Mayorga, K., Yunes, R. J. and Bernard, P., How to recognize and workaround pitfalls in QSAR studies: a critical review. *Curr. Med. Chem.*, 2009, **16**, 4297–4313.

20. Armstrong, N. A., *Pharmaceutical Experimental Design and Interpretation*, CRC Press, Taylor & Francis, 2006.

21. Hoskuldsson, A., PLS regression methods. *J. Chemom.*, 1988, **2**, 211–228.

22. Martens, H. and Naes, T., *Multivariate Calibration*, Chichester, Wiley, 1989.

23. Palyulin, V. A., Radchenko, E. V. and Zefirov, N. S., Molecular field topology analysis method in QSAR studies of organic compounds. *J. Chem. Inf. Comp. Sci.*, 2000, **40**, 659–667.

24. Norman, R. A. *et al.*, Crystal structure of human thymidine phosphorylase in complex with a small molecule inhibitor. *Structure*, 2004, **12**, 75–84.

25. Seeliger, D. and Groot, B. L., Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J. Comp. Aided Mol. Des.*, 2010, **24**, 417–422.

26. Osterberg, F., Morris, G. M., Sanner, M. F., Olson, A. J. and Goodsell, D. S., Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. *Proteins: Struct. Funct. Genet.*, 2002, **46**, 34–40.

27. Lipinski, C. A., Lombardo, F., Dominy, B. W. and Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, 1997, **23**, 3–25.

28. Golbraikh, A. and Tropsha, A., Beware of q^2. *J. Mol. Graph Model.*, 2002, **20**, 269–276.

29. Zheng, W. and Tropsha, A., Novel variable selection quantitative structure–property relationship approach based on the *k*-nearest neighbor principle. *J. Chem. Inf. Comp. Sci.*, 2000, **40**, 185–194.

30. Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M. and Gramatica, P., Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.*, 2003, **111**, 1361–1375.

31. Snedecor, G. W. and Cochran, W. G., *Statistical Methods*, Oxford and IBH, New Delhi, 1967.

32. Roy, K., Supratik, K. and Pravin, A., On a simple approach for determining applicability domain of QSAR models. *Chemometr. Intell. Lab. Syst.*, 2015, **145**, 22–29.