Pharmacophore-similarity-based QSAR (PS-QSAR) for group-specific biological activity predictions

Sivakumar Prasanth Kumar, Yogesh T. Jasraia*, Himanshu A. Pandya and Rakesh M. Rawal

Department of Bioinformatics, Applied Botany Centre (ABC), Gujarat University, Ahmedabad 380009, Gujarat, India; Division of Medicinal Chemistry and Pharmacogenomics, Department of Cancer Biology, The Gujarat Cancer and Research Institute, Ahmedabad 380 016, Gujarat, India

Communicated by Ramaswamy H. Sarma

(Received 7 June 2013; accepted 25 September 2013)

Recent technological breakthroughs in medicinal chemistry arena had ameliorated the perspectives of quantitative structure–activity relationship (QSAR) methods. In this direction, we developed a group-based QSAR method based on pharmacophore-similarity concept which takes into account the 2D topological pharmacophoric descriptors and predicts the group-specific biological activities. This activity prediction may assist the identification of certain pharmacophore features encoded by respective fragments toward activity improvement and/or detrimental effects. We termed this method as pharmacophore-similarity-based QSAR (PS-QSAR) and studied the activity contribution of fragments from 3-hydroxypyridinones derivatives possessing antimalarial activities.

Keywords: group-based QSAR; topological pharmacophore descriptors; 2D QSAR; pharmacophore-similarity-based QSAR (PS-QSAR); antimalarials

Introduction

Quantitative structure–activity relationship (QSAR) methods have been applied in numerous scientific disciplines such as computational drug design, predictive toxicology models, and high-throughput screening. Three-dimensional QSAR, although provide hints about the areas of steric and electrostatic requirements of the molecules, possesses known inherent problems such as the need for feasible molecular alignment and bioactive conformations, while 2D-QSAR has been considered efficient due to ease and robustness in calculating 2D descriptors (Gasteiger & Engel, 2003). Hansch method utilizes molecular descriptors contributing substituent properties such as Hammett and Taft constants, that are related to chemical environment and steric properties of groups (Kubinyi, 1993). These substituent constants are believed to be independent of each other, and their interactions are completely neglected in the QSAR development process (Ajmani, Jadhav, & Kulkarni, 2009). Hence, there is a growing demand to use advanced QSAR approaches in order to connect these chemical information to biological response.

The primary objective of any type of advanced QSARs is to develop a direct relationship from mechanistic chemistry using productive descriptors such as pharmacophore fingerprints (McGregor & Muskal, 1999), molecular interaction fields (Modh, Kumar, Jasrai, & Chikhalia, 2013), grid cell occupancy descriptors (da Cunha, Mancini, & Ramalho, 2012), geometric and electronic features (Ajmani et al., 2009) to enhance our understanding of mechanism (Hansch, Kurup, Garg, & Gao, 2001). The interpretation of the statistically significant models highlights the importance of core structure to design potent agents of biological interest (Devinyak, Zimenkovsky, & Lesyk, 2012). In our presented method, the importance of fragments/substructures with respect to biological activities can be studied by pharmacophore/similarity concept which signifies the activity contribution rate.

Group-based QSAR approach attempts to predict the biological activities specific to substitution sites (R-groups) in the molecular dataset. This approach is available in VLife Molecular Design Suite (VLifeMDS; academic licensed software) and takes advantage of fragment interaction descriptors. These fragment interaction descriptors are developed to surmount the potential drawback of Hansch method and to guide site-specific optimization toward designing new molecules (Ajmani et al., 2009). Analogously, fragment-similarity-based QSAR (FS-QSAR) predicts the biological activities of fragments in the test molecules by comparing the fragments present in the training set molecules by fragment/similarity concept (Myint, Ma, Wang, & Xie, 2011). We present here a group-based QSAR approach using 2D atom pair topological pharmacophoric descriptors and...
predict the biological activities of the R-groups in the test set molecules using pharmacophore/similarity concept. We also employed Z-score and percentile metrics to obtain information about the contribution of each R-group toward the whole molecule biological activities which can assist the role of individual R-group constituting enhancing or detrimental effects. This method is considered as an extension to VLifeMDS group-based QSAR approach (VLifeMDS, 2010) and extensively utilizes topological pharmacophoric descriptors in place of fragment interaction descriptors to build QSAR model. With ongoing efforts on development of computational strategies on medicinal plants-based drug discovery, structural bioinformatics and related methods (Kumar, Jasrai, Pandya, George, & Patel, 2013; Kumar, Patel, Kapopara, Jasrai, & Pandya, 2012), we present here pharmacophore-similarity-based QSAR (PS-QSAR) for group-specific biological activity predictions. In comparison to FS-QSAR approach which estimates similarity among fragments using BCUT matrices and Tanimoto coefficient (Myint et al., 2011), PS-QSAR approach is based on pharmacophore-similarity idea, i.e. R-groups diversified by a atom or a group of atoms but constitute pharmacophore features similar to those encoded by training set R-group molecules. This approach will be very useful for QSAR modeling for two cases. (i) QSAR modeling on a molecular dataset enriched with analogues and (ii) QSAR modeling for active molecules whose molecular target (protein) crystal structures are not available. Besides, PS-QSAR requires a common substructure or a main scaffold (from here, it will be called a template) to distinguish various R-groups present in the dataset to generate fragment molecules.

Materials and methods

Molecular dataset

Dehkordi, Liu, and Hider (2008) synthesized a series of 3-hydroxypyridinones (HPOs) derivatives containing 19 molecules and evaluated its antimalarial properties which act as iron chelators. Hershko et al. (1991) and Pattanapanyasat et al. (2001) also studied the effects of various 3-hydroxypyridin-4-ones (CP compounds) as iron chelators to inhibit the growth of Plasmodium falciparum in vitro. The molecules were drawn in Marvin Sketch program 5.12.2 (ChemAxon LLC, 2012), geometrically optimized using YASARA Structure (Krieger, Darden, Nabuurs, Finkelstein, & Vriend, 2004; academic licensed software) and exported in structure data format (SDF).

Training and test set

The HPO dataset was divided into training and test set to develop PS-QSAR model. About 20% of the molecules (4 out of 19) were chosen as test set molecules, while rest of the molecules were used to train QSAR model. Saghaie, Sakhi, Sabzyan, Shahlaei, and Shamshirian (2013) developed QSAR models using multiple linear regressions (MLR) and principal component regression (PCR) regression statistical methods with the same dataset. They applied the Kennard and Stone algorithm (1969) to split the HPO dataset into training and test sets. This dataset selection was used to develop PS-QSAR model. The experimental antimalarial activities of HPO derivatives were reported by Dehkordi et al. (2008). In addition, we selected 5 CP compounds viz. CP38, CP40, CP51, CP96, and CP110 studied by Hershko et al. (1991) and Pattanapanyasat et al. (2001), as external test set to evaluate the QSAR models predictive ability. These biological activities expressed in IC50 values were converted into its logarithmic scale (pIC50) and subsequently used as dependent variable for QSAR modeling.

Generation of fragment molecules

Dehkordi et al. (2008) designed HPO derivatives by considering 3-hydroxypyridinone as template (Figure 1) and substituted various chemical moieties at R1, R2, and R6 positions. Since PS-QSAR approach utilizes fragment molecules generated from a molecular dataset as input for QSAR modeling, the original dataset must be chopped down to develop fragment molecules. The fragment molecules were generated by the following steps.

1. In the training set (viz. HPO derivative 1, 2, 3, 4, 5, 7, 8, 11, 12, 14, 15, 16, 17, 18 and 19), the substitution sites (R1 and R2, R6 was not considered due to less chemical complexities at this site) was defined in chemical space by aligning with the template (Figure 2) using VLifeMDS program (VLifeMDS, 2010).

2. Various chemical moieties at R1 and R2 sites were noted for each molecule in the training set. Training set molecules possess identical chemical
moieties at respective substitution sites, and these should be removed to produce a unique list constituting diverse fragments for R1 and R2 substitution positions.

(3) Fragment molecules were generated by sketching the template and attaching only one substitution site (For instance, template +R1 site while R2 and R6 sites were not drawn). Similarly, fragment molecules belonging to R2 site were generated. For sake of simplicity, the fragment molecules of R1 site was named as R1a, R1b, R1c, R1d, R1e, R1f, and R1g (Table 1) where each molecule constitute a diverse fragment (a–g) at R1 site. Similarly, the fragment molecules of R2 site were named as R2a, R2b, R2c, R2d, R2e, R2f, and R2g (Table 2).

Assignment of biological activities to fragment molecules

The assignment of biological activity to each fragment molecules was carried out by identifying the fragments representing the training molecules in the dataset at its respective substitution sites. The biological activities of the sorted training set are averaged and assigned to its fragment counterpart. For example, the biological activity of the fragment molecule R1d was calculated as follows.

\[
R1d \text{ (biological activity)} = \frac{\sum \text{biological activity} \times (HPO \text{ derv7, HPO derv14, HPO derv15, HPO derv16})}{n}
\]

Where molecules, namely HPO derivatives 7, 14, 15, and 16 possess the propylpiperidine group at R1 site, and \(n\) is the number of compounds possessing the propylpiperidine fragment in the training set at R1 site.

\[
R1d \text{ (biological activity)} = \frac{(4.6021 + 4.6676 + 5.0000 + 5.0969)}{4} = 4.84165
\]

Similarly, the biological activities of the entire fragment training molecules were computed by applying the above mathematical expression. The biological activities of these fragment molecules were termed as ‘scaled biological activities,’ as it was scaled from original HPO dataset molecules obliged with pharmacophore/similarity concept. Henceforth, this term represents the biological activities of the fragment molecules for ease in understanding.

Calculation of topological pharmacophoric atom pair descriptors

Topological pharmacophoric atom pair descriptors for training (R1a–R1g; R2a–R2g) and test set (HPO derv6-R1, derv6-R2, derv9-R1, derv9-R2, derv10-R1, derv10-R2, derv13-R1, and derv13-R2) fragment molecules were calculated using Perl scripts provided by MayaChemTools (Sud, 2012). The fragment molecules (R1a–R1g, R2a–R2g) were defined in SDF format. Active State Perl 5.16.3 (64-bit) was installed on a Windows Workstation equipped with Intel Xeon processors. The Perl script “TopologicalPharmacophore-AtomPairsFingerprints.pl” was executed using Perl compiler.

Neural network development and statistical analysis

Statistical analysis of the generated pharmacophoric descriptors was performed using principal component analysis (PCA) tool engineered in SPSS statistical package (SPSS Inc., 2007). The factor scores calculated by PCA were defined as independent variables, and its scaled biological activities was used as response variable for developing neural networks specific to R1 and R2 sites using JustNN program (version 4.0, Neural Planner Software). The network constituted 4 input (first 4 factor scores), 1 hidden and 1 output (biological activities) layers, respectively. The network was learned with a learning rate of .6 and momentum of .8 with optimization settings. The learning was halted when the average error of .01 was below the target error. We imposed no
### Table 1. R1-specific fragment molecules.

| S. No | Molecules          | R1-specific fragments representing training set molecules | Naming conventions | Calculated biological activity (log units) |
|-------|--------------------|----------------------------------------------------------|--------------------|------------------------------------------|
| 1     | HPO derv 1, 2      | R1a                                                      |                    | 4.1290                                   |
| 2     | HPO derv 3, 4      | R1b                                                      |                    | 4.0840                                   |
| 3     | HPO derv 5         | R1c                                                      |                    | 4.3010                                   |
| 4     | HPO derv 7, 14, 15, 16 | R1d                                                  |                    | 4.8417                                   |
| 5     | HPO derv 8         | R1e                                                      |                    | 4.3010                                   |

(Continued)
restrictions on the periodic stop of learning cycles for better optimization. The network predictive accuracy was measured by applying the factors scores (F1–F4) derived from pharmacophoric descriptors encoded by test set molecules. Z-score statistic and percentile calculations were performed in SPSS modules (SPSS Inc., 2007) to understand the contribution of various fragments in the molecules toward overall biological activities.

**Results and discussion**

**Calculation of topological pharmacophoric atom pair descriptors**

Three-dimensional pharmacophore models congregate the set of pharmacophore points along with their relative arrangement depicted in Euclidean distance between each point pair. Since no spatial information is available in the case of 2D pharmacophore models, topological relations are used to exemplify the relative position of pharmacophore points (Güner, 2000). Atom-pair-based 2D pharmacophore descriptors are defined as the collection of all atom–atom pharmacophore feature pairs along with their topological distances (Langer & Hoffmann, 2006). It was ensured that pharmacophoric descriptors should be mapped when a feature connected to another feature by ≤10 arbitrary bonds. For example, H-D10-HBA descriptor maps a hydrophobic (H) feature connected by 10 bond distance toward a hydrogen bond acceptor (HBA) feature in the molecule. Here, the HBA feature was mapped on template, while H feature was mapped on R1 group of the R1d molecule (Figure 3).

Atom-pair-based topological pharmacophore descriptors were calculated by specifying the fragment molecules in SDF format. Prior to pharmacophore descriptors calculations, all the molecules were geometrically optimized using AMBER03 force field (Duan et al., 2003) in YASARA Structure (Krieger et al., 2004) and exported. The perl script, TopologicalPharmacophore-AtomPairsFingerprints.pl was compiled using command line prompt in Windows kernel defining the input file name and its output containing counts of mapped pharmacophore descriptors in CSV (comma-delimited) format and subsequently employed to develop PS-QSAR model.

**Removal of inconsistent pharmacophore descriptors**

Inconsistent columns of atom-pair-based topological pharmacophore descriptors were recognized by calculating standard deviation (SD) of each descriptor encoded by the training set. Initially, descriptor columns enriched with value zero were removed due to its inconsistency as a variable in making up a QSAR model. To retain only noncorrelative descriptors to develop PS-QSAR model, descriptors whose SD ≥ .5 were considered. The selected list of topological pharmacophoric descriptors along with its computed descriptors values pertaining to R1- and R2-specific fragment molecules were listed in Tables 3–5, respectively.

**PCA to develop PS-QSAR model**

PCA is a mathematical procedure which orthogonally transforms a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables known as PCRs. This orthogonal transformation was made in such a way that the first PCR has the largest possible variance which may represent much of the variability in the data observations and subsequent
PCRs may constitute lower levels of variability (Pearson, 1901). Pearson-type PCA was performed separately for R1- and R2-specific fragment molecules using SPSS statistical package (SPSS Inc., 2007).

The Eigen values along with its variability and cumulative percentages calculated from pharmacophoric descriptors encoded by R1 fragment molecules revealed that factor 1 (F1) contains 52.621% variability of data points. The total variability across first two factors viz. F1 and F2 accounts to 79.611%. In the case of R2 fragment molecules, F1 contains 49.211% data point variability. The total variability among first two factors viz. F1 and F2 sums up to 77.628% (Table 6). It indicates that PCRs generated from first 2 factors are reliable to orthogonally map the descriptors space. It was observed that the descriptors were clustered nearby without any outliers and occupied near the origin (Figure 4). The correlation matrix among various pharmacophoric descriptors from R1- and R2-specific fragment training molecules are shown in Supplementary Tables S1 and S2. The correlation of selected pharmacophoric descriptors of R1 and R2 fragment molecules with its respective PCA factors are listed in Supplementary Tables S3 and S4.

**Factor scores calculation**

The results of a PCA are usually explained in terms of component scores. The component scores also known as factor scores represent the transformed variable values corresponding to a particular data point of a set of observations while factor loadings describes the weight by which each standardized original variable should be multiplied to achieve the component score (Shaw, 2003). Since the Eigen vector and factors transformations are calculated from the PCRs, it is a prerequisite to attain a good correlation among the factor scores with descriptors so as to ensure the mapping of local pharmacophore space encoded by the training fragment molecules. In R1 fragment molecules, the correlation of eight descriptors with F1 is above .5, while F2 do not possess any significant correlations. In the case of R2 fragment molecules, nine descriptors ensured more than .5 correlation with respect to F1, while F2 have five descriptors with correlations more than .5. Hence, the calculated factors describing data variability can be used as independent variables to train neural network to predict biological response of the HPO derivatives. The motive behind the consideration of factor scores as independent variables is due to the following two reasons: i. the topological pharmacophoric descriptors are mere counts of mapped features and exist as highly correlated variables among each other thereby limiting its role as independent variable for QSAR modeling and ii. each factor score represents a transformed value which in turn contains information from different descriptors, the loss of information was averted. The role of factor scores as independent variables to develop MLR-based QSAR was recently demonstrated by Shahlaei, Fassihi, and Nezami (2009) on 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives. The factor scores calculated for seven training fragment molecules vide R1a–R1g and R2a–R2g are summarized in Supplementary Table S5 and S6, respectively.

**Neural network to predict biological activities**

R1- and R2-specific neural networks were developed with 4-1-1 architecture using factor scores computed from respective PCAs of the training set as input layers, while the scaled biological activities were specified as response variable (output) (Figure 5). Each network was
developed with a learning rate of .6 and momentum of .8 with optimization settings. The R1-specific network was built in 152 learning cycles with a training error of .009987, while R2-specific network was made in 88 learning cycles with a training error of .009886. The relative importance of the input layer signifies the higher contribution rate of F1 (R1 network: 3.9221; R2 network: 3.2849) and F2 (R1 network: 3.1531; R2 network: 3.1531).

| S. No | Molecules          | R2-specific fragments representing training set molecules | Naming conventions | Calculated biological activity (log units) |
|-------|--------------------|----------------------------------------------------------|--------------------|------------------------------------------|
| 1     | HPO derv 1, 3, 5, 8, 17 | R2a                                                      | 4.1882             |
| 2     | HPO derv 2, 4, 7, 18   | R2b                                                      | 4.2800             |
| 3     | HPO derv 11          | R2c                                                      | 4.5380             |
| 4     | HPO derv 12          | R2d                                                      | 4.4560             |
| 5     | HPO derv 14          | R2e                                                      | 4.6680             |
| 6     | HPO derv 15          | R2f                                                      | 5.0000             |
| 7     | HPO derv 16, 19      | R2g                                                      | 4.7220             |
while the lowest contribution was attributed to the last input layer, i.e. F4 (R1 network: .7181; R2 network: 1.9730) in both the networks. To prevent overfitting of the QSAR model, the training error was plotted with the number of learning cycles (Supplementary Figure S1). It was noted that the learning rate was optimized when the maximum error (R1 network: .030212; R2 network: .043367) trended linearly with the average error (R1 network: .009987; R2 network: .009886) indicating the reliability of the optimization scheme adopted in respective network modelling. The accuracy of the networks was then assessed by scatter plot analysis.

3.2630), while the lowest contribution was attributed to the last input layer, i.e. F4 (R1 network: .7181; R2 network: 1.9730) in both the networks. To prevent overfitting of the QSAR model, the training error was plotted with the number of learning cycles (Supplementary Figure S1). It was noted that the learning rate was optimized when the maximum error (R1 network: .030212; R2 network: .043367) trended linearly with the average error (R1 network: .009987; R2 network: .009886) indicating the reliability of the optimization scheme adopted in respective network modelling. The accuracy of the networks was then assessed by scatter plot analysis.

Scatter plots of biological activities prediction
The scatter plots were developed to understand the extent of residuals between the scaled biological activities and its predicted biological activities provided by R1- and R2-specific neural networks, respectively (Figure 6). The regression coefficient ($R^2$) of .895 and .907 for R1 and R2 fragment training molecules showed its statistical significance by obtaining trend line of best fit (Table 7).

The PS-QSAR models were validated by test set molecules in which the factor scores computed through
R1- and R2-specific PCAs were defined to predict its biological activities which are then compared with the scaled biological activities calculated (described in biological activity assignment section) to arrive at its residuals. For example, the test fragment molecule, HPO derivative 6-R1 resemble the training fragment molecule R1d whose R1 site constitute propylpiperidine moiety. Similarly, the predicted biological activity of R2-specific test molecules was compared with the representative fragment present in the R2-specific training fragment molecules. The scatter plot provides a graphical view of the best line of fit along with its data points. The plot suggests significant predictive ability of the developed PS-QSAR neural networks. In addition, the model predictive ability was evaluated using external test set molecules (5 CP compounds).

The network predicted biological activities of R1- and R2-specific test and external test molecules with scaled and actual biological activities yielded significant regression coefficients ($R^2_{\text{test}}$ and $R^2_{\text{ext.test}}$). The $R^2_{\text{test}}$ for R1- and R2-specific test set molecules were .9994 and .9971 indicating the significance of PS-QSAR models. Beside, the $R^2_{\text{ext.test}}$ for R1- and R2-specific external test set molecules were .9599 and .9640, respectively, which shows that the built QSAR model can be extended to predict biological activities of HPO- and CP-based iron

|                | F1   | F2   | F3   | F4   | F5   | F6   |
|----------------|------|------|------|------|------|------|
| R1 fragment training molecules |      |      |      |      |      |      |
| Eigen value    | 6.315| 3.239| 1.267| .942 | .175 | .062 |
| Variability (%)| 52.621| 26.990| 10.562| 7.852| 1.456| .518 |
| Cumulative (%)  | 52.621| 79.611| 90.173| 98.025| 99.482|100.000 |
| R2 fragment training molecules |      |      |      |      |      |      |
| Eigen value    | 7.382| 4.263| 2.623| .326 | .248 | .159 |
| Variability (%)| 49.211| 28.417| 17.485| 2.175| 1.652| 1.059 |
| Cumulative (%)  | 49.211| 77.628| 95.114| 97.289| 98.941|100.000 |

Figure 4. Graphical illustration of the Eigen values and cumulative variability showing its variability across factors from R1 (a) and R2 (b) descriptor sets and its space over F1 and F2 (c, d).
chelators. It should be noted that the residuals estimated for the entire fragment dataset is ~1 numerical value which suggests that the developed PS-QSAR model on HPO dataset may predict the antimalarial response meaningfully. In addition, \( \Delta r^2_m \) value proposed by Ojha, Mitra, Das, and Roy (2011) was taken to measure the predictive accuracy of the PS-QSAR models. The \( \Delta r^2_m \) values for both R1- and R2-specific training and test sets were found to be <.2 (cutoff value; Ojha et al., 2011) which suggests the better predictive quality of the developed models (Supplementary Figure S2).

**Contribution of various fragments toward overall biological activities**

To recognize the fragments contributing toward enhancement or detrimental effects on the whole molecule biological activity (experimental pIC\(_{50}\) values of the original HPO dataset), the Z-score statistic and percentile metrics were applied (Figure 7). Z-scores explain how outstanding an individual is relative to the population mean using the SD for that population. Z-score was considered as an effective way to compare individuals in different populations. On the other hand, percentiles are very useful for estimating the relative standing of an individual in a population and essentially represent the rank position of an individual (Cohen & Holliday, 1996). The terminology ‘whole molecule’ corresponds to the original HPO derivative molecules designed by Dehkordi et al. (2008) in order to distinguish site-specific fragment molecules. Accordingly, the term ‘whole molecule biological activity’ represents the experimental pIC\(_{50}\) values studied by Dehkordi et al. (2008).

The Z score for whole molecule biological activity analysis:

\[
Z \text{ score} = \frac{\text{experimental pIC}_{50} - \text{mean of the experimental pIC}_{50} \text{ from the training set}}{\text{SD of the experimental pIC}_{50} \text{ from the training set}}
\]

The Z score for fragment molecules (both R1 and R2) biological activity analysis:

\[
Z \text{ score} = \frac{\text{predicted biological activity} - \text{mean of the predicted biological activity from the training set}}{\text{SD of the predicted biological activity from the training set}}
\]
Table 7. Scaled and predicted R1- and R2-specific activities for training, test and external set molecules.

| Compounds | Predicted site-specific activities | Scaled whole molecule activities |
|-----------|------------------------------------|---------------------------------|
| Training set |                                    |                                 |
| R1a       | 4.189                             | 4.129                           |
| R1b       | 4.266                             | 4.084                           |
| R1c       | 4.310                             | 4.301                           |
| R1d       | 4.831                             | 4.842                           |
| R1e       | 4.183                             | 4.301                           |
| R1f       | 4.349                             | 4.388                           |
| R1g       | 4.221                             | 4.303                           |
| R2a       | 4.368                             | 4.188                           |
| R2b       | 4.156                             | 4.280                           |
| R2c       | 4.549                             | 4.538                           |
| R2d       | 4.679                             | 4.456                           |
| R2e       | 4.554                             | 4.668                           |
| R2f       | 4.796                             | 5.000                           |
| R2g       | 4.751                             | 4.722                           |
| Test set |                                    |                                 |
| Comp6     | 4.801                             | 4.244                           |
| Comp9     | 4.288                             | 4.478                           |
| Comp10    | 4.288                             | 4.766                           |
| Comp13    | 4.288                             | 4.390                           |
| External test set |                                    |                                 |
| CP38      | 4.179                             | 4.244                           |
| CP40      | 4.288                             | 4.273                           |
| CP51      | 4.279                             | 4.330                           |
| CP96      | 4.186                             | 4.330                           |
| CP110     | 4.182                             | 4.228                           |

Z score = (PS-QSAR predicted pIC50 – mean of the PS-QSAR predicted pIC50 from the training set) /SD of the PS-QSAR predicted pIC50 from the training set

Similarly, the Z score can be calculated for the test and external set molecules. Compounds such as HPO derivatives 6, 9, 10, and 13 were taken as test set molecules and its respective experimental biological activities were used for comparison, to study the contribution of various fragments promoting the biological activities in question. The predictions were then cross-validated by applying external test set (5 CP compounds viz. CP38, CP40, CP51, CP96, and CP110). The calculated Z score and percentiles for R1 and R2 fragment training, test and external test set molecules are provided in Supplementary Table S7.

Based on the percentile scored by each of the test set molecules, the contribution of R-group-specific activity (in %) was calibrated (Figure 8). In R1 substitution site, an aliphatic linker with piperidine moiety (HPO derv 6) strongly influences the biological activity of the whole molecule and contributes 99.49% of the activity at R1 site. In addition, merely placing a methyl group would have only 46% of contribution toward R1 site activity. In R2 substitution site, ethylpiperidine (HPO derv 9) and ethylmorpholine (HPO derv 10) strongly influences the R2 site-specific activity and contributes 75% and 99.06%, respectively. Both explicit hydrogen and a methyl group would have less influential (24% and 56%) property for enhancing the activity at R2 site. Thus, the site-specific preference of moieties along with the contribution rate will facilitate the medicinal chemist to intelligently substitute chemical groups at its respective sites in order to enhance the activity of the molecule as a whole. In addition, the R6 site which was not explored due to less chemical complexities in the HPO derivatives can be explained by activity distribution of the compounds. For example, Dehkordi et al. (2008) reported 19 active HPO molecules having antimalarial activities among which compounds, 3-hydroxy-2-(hydroxymethyl)-1-[2-(piperidin-1-yl)ethyl]-1,4-dihydropyridin-4-one (18a, pIC50 = 5.0) and 3-hydroxy-2-(1-hydroxyethyl)-1-[2-(piperidin-1-yl)ethyl]-1,4-dihydropyridin-4-one (18b, pIC50 = 5.0969) constituted H at R6 site. Hence, it is evident that R6 substitution site prefers less chemical complexities.

The site-specific activity contribution chart for external test set molecules (Figure 9) showed that the butanal and ethyl groups at R1 site of CP38 and CP96 molecules have 25% influential property in activity determination which is threefold less than the piperidine moiety (HPO derv 6). The coordination mode of HPO derivatives in contact with Fe3+ atom indicated that oxygen groups available in the HPO scaffold (nonsubstitution sites) are directly involved in iron-scavenging activities, while R1 and R2 sites may be involved in developing bis- and tris-chelating podants to establish interaction with iron atom (Santos, Marques, & Chaves, 2012). Hence, appropriate bulk group can be substituted at R1 site. In accordance with the proposed coordination model, the presence of H bond donor in the aliphatic chain at R2 site (CP40, 29%) do not contribute to activity enhancement, while H bond acceptor is preferred at R2 site as suggested by compounds CP51 and CP96 (42%), respectively. Further, R2 site-specific activity was found to be improved by substitution with a steric group. For example, ethylpiperidine (HPO derv 9, 75%), ethylmorpholine (HPO derv 10, 99.06%), aliphatic chain with more than 4 interconnected atoms (CP51, CP96, 42%). Thus, site-specific contribution charts are very useful in understanding the role of various chemical groups at different substitution sites. The molecular dataset should reflect higher order of chemical complexities to efficiently encode various pharmacophoric...
descriptors which will be helpful to prioritize chemical groups at respective substitution sites to develop lead molecules.

The major advantage of this proposed PS-QSAR method is that given an analogous molecular dataset, the group-based activity prediction can be efficiently scrutinized provided all the molecules possess identical template with variable substitution sites. The limitations of this approach is the requirement of analogous molecular dataset with predefined substitution sites (R-groups) which may in turn encode similar topological pharmacophoric descriptors by its respective fragment molecules so as to ensure appropriate PCA calculations. If the chemical space is populated by diverse molecules constituting variable R-groups, the identification of R-groups based on common template alignment will be laborious.

Figure 7. Plots of Z scores based on scaled and predicted biological activities for training (R1-(a), R2-(b)) and test and external test (R1-(c), R2-(d)) molecules.

Figure 8. Group-based activity contribution chart for the test set molecules.
Conclusion

Fragment-based QSAR and group-based QSAR methodologies are useful for fragment-based strategy to predict biological activities based on the physico-chemical properties, fragment interaction descriptors, etc., and the predicted activities will help the medicinal chemist to understand the fragment contribution with the anticipation of enhancement and/or detrimental effects. A new methodology of pharmacophore/similarity-based QSAR (PS-QSAR) was developed which utilizes topological pharmacophoric atom-pair-based descriptors at substitution sites and makes extensive usage of PCAs for factor score calculation. The factor scores were subsequently employed as independent variables to develop a substitution site-specific neural network to estimate site-specific biological activities. Fragment contribution toward enhancement or detrimental effects with respect to whole molecule activities were then analyzed by employing Z-statistic and percentile metrics. This PS-QSAR approach may highlight the preference of pharmacophoric features encoded by various chemical moieties at respective substitution sites in its scaffold. This approach was validated using 3-hydroxypyridinones antimalarial compounds.

Acknowledgments

Sivakumar Prasanth Kumar acknowledges support from Department of Science and Technology, New Delhi, India as INSPIRE Fellowship.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2013.849618.

References

Ajmani, S., Jadhav, K., & Kulkarni, S. A. (2009). Group-based QSAR (G-QSAR): Mitigating interpretation challenges in QSAR. QSAR & Combinatorial Science, 28, 36–51.

ChemAxon LLC. (2012). Marvin Sketch 5.11.4. www.chemaxon.com.

Cohen, L., & Holliday, M. (1996). Practical statistics for students: An introductory text. New Delhi: Sage.

da Cunha, E. F. F., Mancini, D. T., & Ramalho, T. C. (2012). Molecular modeling of the Toxoplasma gondii adenosine kinase inhibitors. Medicinal Chemistry Research, 21, 590–600.

Devinyak, O., Zimenkovsky, B., & Lesyk, R. (2012). Biologically active 4-thiazolidinones: a review of QSAR studies and QSAR modeling of antitumor activity. Current Topics in Medicinal Chemistry, 12, 2763–2784.
Dehkordi, L. S., Liu, Z. D., & Hider, R. C. (2008). Basic 3-hydroxypyridin-4-ones: Potential antimalarial agents. *European Journal of Medicinal Chemistry, 43*, 1035–1047.

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., ... Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry, 24*, 1999–2012.

Gasteiger, J., & Engel, T. (2003). *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH.

Güner, O. F. (2000). *Pharmacophore – Perception, development, and use in drug design*. La Jolla, CA: International University Line.

Hansch, C., Kurup, A., Garg, R., & Gao, H. (2001). Chemoinformatics and QSAR: A review of QSAR lacking positive hydrophobic terms. *Chemical Reviews, 101*, 619–672.

Hershko, C., Theanacho, E. N., Spira, D. T., Peter, H. H., Dobbin, P., & Hider, R. C. (1991). The effect of N-alkyl modification on the antimalarial activity of 3-hydroxypyridin-4-one oral iron chelators. *Blood, 77*, 637–643.

JustNN (4.0). Neural Planner Software. Cheshire: Neural Planner Software Ltd.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics, 11*, 137–149.

Krieger, E., Darden, T., Nabuurs, S. B., Finkelstein, A., & Friend, G. (2004). Making optimal use of empirical energy functions: Force-field parameterization in crystal space. *Proteins Structure Function and Bioinformatics, 57*, 678–683.

Kubinyi, H. (1993). *QSAR: Hansch analysis and related approaches*. Weinheim: Wiley-VCH.

Kumar, S. P., Jasrai, Y. T., Pandya, H. A., George, L. B., & Patel, S. K. (2013). Structural insights into the theoretical model of *Plasmodium falciparum* NADH dehydrogenase and its interaction with artemisinin and derivatives: Towards global health therapeutics. *OMICS: A Journal of Integrative Biology, 17*, 231–241.

Kumar, S. P., Patel, S. K., Kapopara, R. G., Jasrai, Y. T., & Pandya, H. A. (2012). Evolutionary and molecular aspects of Indian tomato leaf curl virus coat protein. *International Journal of Plant Genomics, 417935*, doi:10.1155/2012/417935

Langer, T., & Hoffmann, R. D. (2006). *Pharmacophores and pharmacophore searches*. Weinheim: Wiley-VCH.

McGregor, M. J., & Muskal, S. M. (1999). Pharmacophore fingerprinting. I. Application to QSAR and focused library design. *Journal of Chemical Information and Modeling, 39*, 569–574.

Modh, R. P., Kumar, S. P., Jasrai, Y. T., & Chikhalia, K. H. (2013). Design, synthesis, biological evaluation and molecular modeling of coumarin–piperazine derivatives as acetylcholinesterase inhibitors. *Archiv der Pharmazie, 346*, doi:10.1002/ardp.201300242

Myint, K. Z., Ma, C., Wang, L., & Xie, X. Q. (2011). Fragment-similarity-based QSAR (FS-QSAR) algorithm for ligand biological activity predictions. *SAR and QSAR in Environmental Research, 22*, 385–410.

Ojha, P. K., Mitra, I., Das, R., & Roy, K. (2011). Further exploring rm2 metrics for validation of QSRR models. *Chemosometrics and Intelligent Laboratory Systems, 107*, 194–205.

Pattanapanyasat, K., Kotipun, K., Yongvanitchit, K., Hider, R. C., Kyle, D. E., Heppner, D. G., & Walsh, D. S. (2001). Effects of hydroxypyridinone iron chelators in combination with antimalarial drugs on the *in vitro* growth of *Plasmodium falciparum*. *The Southeast Asian Journal of Tropical Medicine and Public Health, 32*, 64–69.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, 2*, 559–572.

Saghaie, L., Sakhi, H., Sabzyan, H., Shahlaei, M., & Shamshiran, D. (2013). Stepwise MLR and PCR QSAR study of the pharmaceutical activities of antimalarial 3-hydroxypyridinone agents using B3LYP/6-311+G** descriptors. *Medicinal Chemistry Research, 22*, 1679–1688.

Santos, M. A., Marques, S. M., & Chaves, S. (2012). Hydroxypyridinones as “privileged” chelating structures for the design of medicinal drugs. *Coordination Chemistry Reviews, 256*, 240–259.

Shahlaei, M., Fassihi, A., & Nezami, A. (2009). QSAR Study of some 5-methyl/trifluoromethoxy-1H-indole-2,3-dione-3-thiosemicarbazone derivatives as anti-tubercular agents. *Research in Pharmaceutical Sciences, 4*, 123–131.

Shaw, P. J. A. (2003). *Multivariate statistics for the environmental sciences*. Chichester: Wiley-VCH.

SPSS Inc. (2007). *SPSS for Windows (16.0)*. Chicago: IBM Corporation.

Sud, M. (March, 2012). *MayaChemTools: An open source package for computational discovery*, 243rd ACS National Meeting & Exposition, San Diego, CA.

VLifeMDS: Molecular Design Suite. (2010). Pune: VLife Sciences Technologies Pvt. Ltd. www.vlifesciences.com.