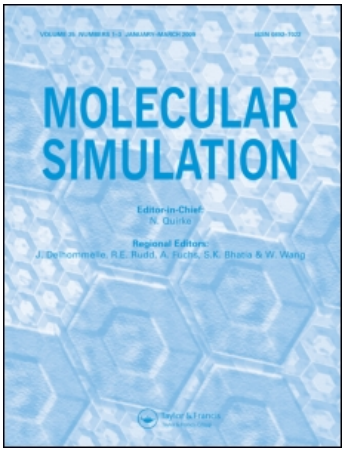


Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Comparative QSAR studies of nitrofuranyl amide derivatives using theoretical structural properties

Payel Ghosh[a]; M. C. Bagchi[a]
[a] Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology, Calcutta, India

To cite this Article Ghosh, Payel and Bagchi, M. C.(2009) 'Comparative QSAR studies of nitrofuranyl amide derivatives using theoretical structural properties', Molecular Simulation, 35: 14, 1185 — 1200
To link to this Article: DOI: 10.1080/08927020903033141
URL: http://dx.doi.org/10.1080/08927020903033141

Molecular Simulation
Vol. 35, No. 14, December 2009, 1185–1200

Taylor & Francis
Taylor & Francis Group

Comparative QSAR studies of nitrofuranyl amide derivatives using theoretical structural properties

Payel Ghosh and M.C. Bagchi*

Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology, 4 Raja S.C. Mullick Road, Jadavpur, Calcutta 700032, India

(*Received 14 April 2009; final version received 6 May 2009*)

The global increase in multidrug-resistant *Mycobacterium tuberculosis* strains and intolerance of first-line anti-tuberculosis drugs may cause major health problems and necessitate modification of the structural therapy regimen. In an ongoing effort to develop new and potent anti-tuberculosis agents, a series of nitrofuranyl amides were subjected to quantitative structure–activity relationship (QSAR) analysis using various feature selection methods. Nitrofuranyl amide derivatives with good therapeutic indices are known to inhibit an enzyme responsible for bacterial cell-wall synthesis and act as novel mycobacterial inhibitors. Successful implementation of a predictive QSAR model largely depends on the selection of a preferred set of molecular descriptors that can signify the chemical–biological interaction. Genetic algorithm (GA), simulated annealing and stepwise regression are applied as variable selection methods for an effective comparison and model development. The results of two-dimensional QSAR showed that a combination of topological indices, hydrophobic properties and autocorrelation descriptors of different atomic properties could be explored to design potent anti-tubercular inhibitors. Further analysis using three-dimensional QSAR technique identifies a suitable model obtained by GA-partial least square method leading to anti-tubercular activity prediction. The influences of steric and electrostatic field effects generated by the contribution plot are analysed and discussed. Both two- and three-dimensional QSAR analyses of such derivatives provide important structural insights for designing potent anti-tuberculosis drugs.

Keywords: nitrofuranyl amide; quantitative structure–activity relationship (QSAR); genetic algorithm; simulated annealing; partial least squares (PLS)

1. Introduction

Tuberculosis (TB) continues to be a major public health threat despite the availability of curative chemotherapy and a vaccine. The World Health Organisation estimates that, in 2005, 1.6 million people died from TB, which is caused by the micro-organism *Mycobacterium tuberculosis* (Mtb) [1]. Current therapy for TB lasts for several months and is called directly observed therapy short-course (DOTS). This regimen calls for an intensive phase of chemotherapy using four drugs for 2 months, followed by a continuation phase using two drugs for 4–6 months [2]. Besides this long treatment regimen, the effectiveness of the currently available agents used in standard TB treatment regimens is severely limited by the emergence of resistance, particularly multidrug-resistant tuberculosis (MDRTB) and extensively drug-resistant tuberculosis (XDR-TB). To overcome this problem, increasing clinical importance has directed additional urgency to researchers to identify new and effective anti-mycobacterial compounds with novel mechanisms of action that can bypass the resistance mechanism and treat the latent phase of infection, shortening the TB treatment [3,4]. Literature survey reveals that nitrofuranyl amide compounds show very good activity against H37Rv strain of Mtb [5–11]. Compounds in this series were easily synthesised, and exhibited good therapeutic indices. Importantly, these compounds exhibit activity against both actively growing and latent bacilli, which is believed to be a beneficial attribute of potential new anti-tuberculosis agents [10]. Although the actual mechanism of action of such derivatives is not clear, it has been experimentally confirmed that they inhibit an enzyme UDP-galactose mutase (Glf), important for synthesising UDP galactofuranose, an essential component of bacterial cell wall [5]. Glf can therefore be considered to be an attractive target for anti-tubercular drug design, rendering the structural studies of nitrofuranyl amide derivatives extremely important.

Understanding the effect of structural features on the activity would help the researchers to design new molecules that may come up as potential new lead molecules. Quantitative structure–activity relationship (QSAR) approach is certainly useful for designing newer drugs when the target is not known or if there are multiple targets. Since physicochemical data are not always available to develop the predictive models; the only alternative is to utilise the theoretical molecular descriptors for such modelling purposes that can be derived solely from the chemical structures of compounds. QSAR

*Corresponding author. Email: mcbagchi@iicb.res.in

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020903033141
http://www.informaworld.com

models can be described as a mathematical relationship between the structural features of a set of chemicals and the activity associated with them. This paradigm leads us to the proper choice of the structural descriptors for predicting the biological response of molecules with the development of optimal QSAR models through some variable selection methodologies. This implies that only a subset of the available descriptors of chemical structures, which are most meaningful and statistically significant in terms of correlation with biological activity, is selected. Among many other techniques, the optimum selection of variables can be achieved by combining statistical and stochastic search methods in conjunction with partial least square (PLS) analysis for model development [12,13]. In the present investigation, three widely used techniques, viz. stepwise, genetic algorithm (GA) and simulated annealing (SA) have been applied for descriptor optimisation. The present paper is an attempt in this direction seeking for the development and comparison of QSAR models of nitrofuranyl amides by different feature selection methods, which ultimately establishes the superiority of the GA-based models. The models developed in such procedures are discussed by means of the selected descriptors from the structural point of view and some interesting results are obtained.

Beside this, three-dimensional QSAR modelling has also been carried out in the present investigation to give some structural insight and suggestions for synthesising more active anti-tubercular agents. Three-dimensional QSAR refers to the application of force-field calculations requiring three-dimensional structures based on protein crystallography or molecule superposition. It uses computed potentials, e.g. the Lennard-Jones or Columbic potential, rather than experimental constants and is concerned with the overall molecule rather than a single substituent. It examines the steric fields that are responsible for the shape of the molecule and the electrostatic fields. The created data space is then usually reduced by a feature extraction. In the present study, PLS technique has been applied in two- as well as three-dimensional modelling after undergoing the feature extraction procedures, and some interesting results are discussed in the following sections.

2. Materials and methods

2.1 Biological dataset and descriptor calculation

Anti-tuberculosis inhibitory concentration data of nitrofuranyl amide compounds (Table 1) have been collected from the literature [5–11] for the present QSAR study. These derivatives show very interesting biological properties and potent anti-tubercular activity, good pharmacokinetics and low-toxicity profile. Importantly, they are not cross-resistant with any currently used TB drugs and are significant for both the active and dormant bacilli. The minimum inhibitory concentration data [MIC (μg/ml)] were converted to logarithmic scale [pMIC (μM/ml)] and then used for subsequent QSAR analyses as the dependent variable. Molecular structure was built in the two-dimensional builder module and then the structure was converted to three-dimensional space for further analysis. For all molecules, energy was minimised using the Merck molecular force field (MMFF) considering a dielectric constant of 1.0 and the convergence criterion or root-mean-square (RMS) gradient at 0.0001 using MDS software of VLife [14]. The energy-minimised geometry was used for the calculation of the molecular descriptors.

For the present QSAR study, topological, shape and geometrical, electrostatic, quantum chemical, physicochemical parameters such as lipophilicity ($\log P$), volume, smr, etc. were used as predictor variables, as they were found to be appropriate for the development of models. A considerable number of the 184 theoretical molecular descriptors comprising of 16 semi-empirical quantum chemical calculations were done using the MDS package of VLife Software.

In addition to the VLife descriptors, 89 descriptors were calculated using the ADRIANA.Code [15] software suite for the molecules that do not possess any ionisable groups. These include: topological polar surface area (TPSA) [16], aqueous solubility ($\log S$) [17], two-dimensional molecular autocorrelation vectors, etc. In the autocorrelation vectors calculated by ADRIANA.Code, the hydrogen atoms were included. Two-dimensional molecular autocorrelation vectors [18] for physicochemical atomic properties were calculated for each molecule by using the following equation:

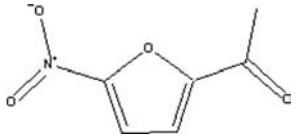
$$A(d) = \sum_{ij} p_i p_j \quad (d = d_j - d_i),$$

where $A(d)$ is the topological autocorrelation coefficient referring to atom pairs i and j which are separated by distance d. p_i is an atomic property, e.g. the σ charge on atom i. Thus, for each compound, a series of coefficients for different topological distances d, a so-called autocorrelation vector, are obtained; seven distances from a distance of $d = 0$ to $d = 6$ were considered. Seven atomic properties are represented by p_i: σ charge (SigChg), π charge (PiChg), total charges (TotChg), σ electronegativity (SigEN), π electronegativity (PiEN), lone-pair electronegativity (LpEN) and atomic polarisability (Apolariz).

Both sets of descriptors calculated by MDS and ADRIANA.Code were used for building models for 82 neutral nitrofuranyl compounds, whereas descriptors calculated by MDS were taken into account for the rest 17 derivatives with ionisable group for predicting anti-tuberculosis activity.

Table 1. The structures of nitrofuranyl amide derivatives with their anti-tuberculosis activities.

R =



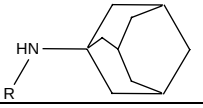
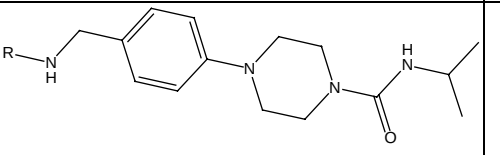
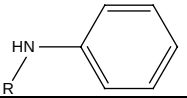
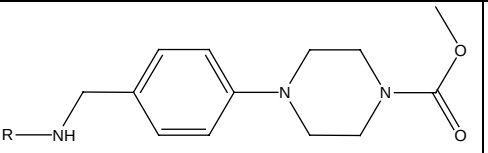
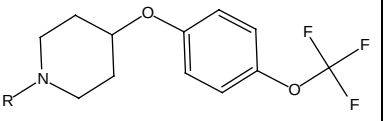
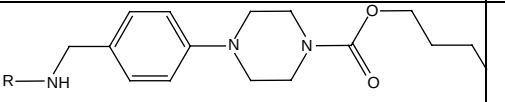
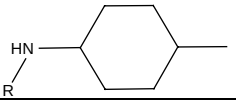
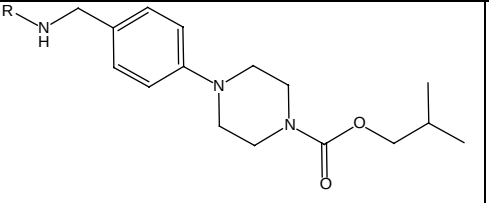
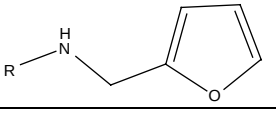
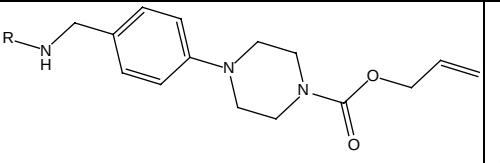
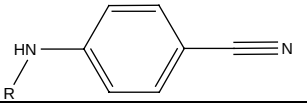
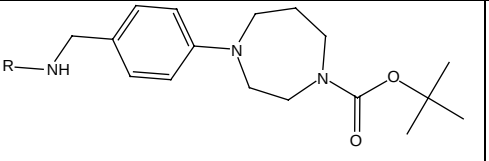
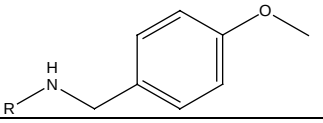
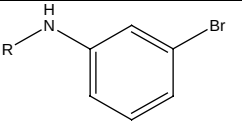
ID	Structures	P[MIC]	ID	Structures	P[MIC]
N-1		1.971514	N-51[b]		2.715428
N-2		2.462764	N-52[c]		3.890287
N-3		5.699308	N-53		4.235994
N-4		1.910504	N-54[b/c]		4.235994
N-5		1.57737	N-55		3.918469
N-6		2.50719	N-56		3.64786
N-7[b]		3.4413	N-57		2.288768

Table 1 – *continued*

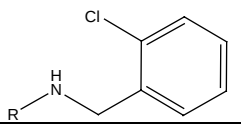
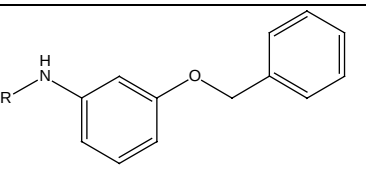
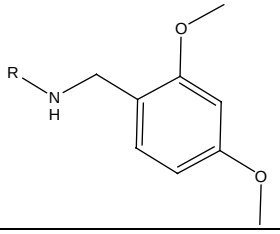
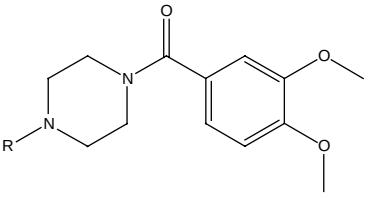
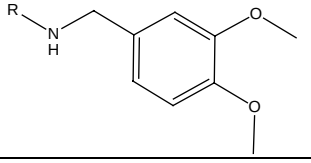
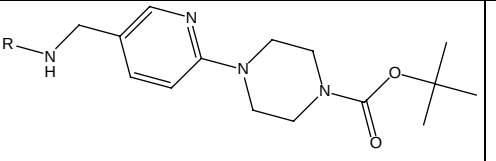
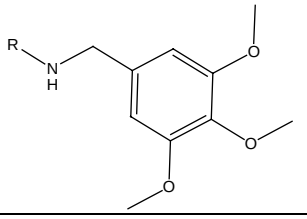
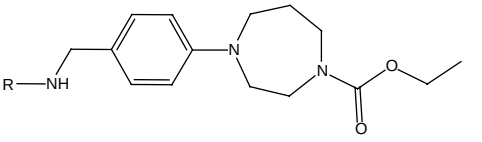
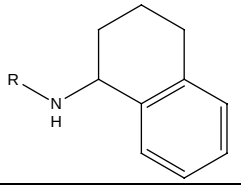
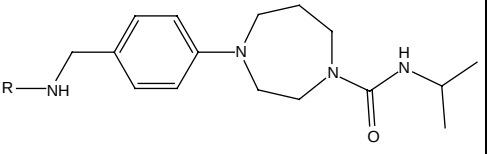
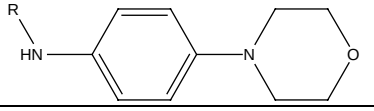
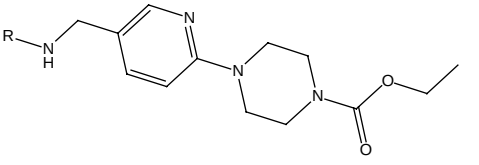
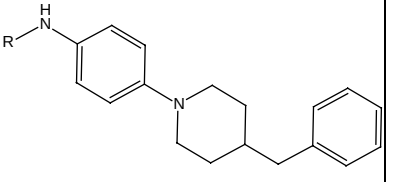
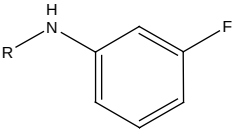
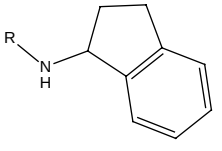
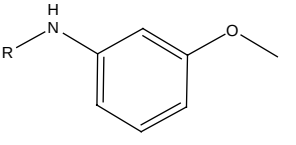
N-8		2.244072	N-58		1.432417
N-9		2.884052	N-59		1.794477
N-10		3.185082	N-60		5.731839
N-11		2.623639	N-61[b]		3.619546
N-12		1.96544	N-62		3.030879
N-13		2.005928	N-63		3.90676
N-14[b]		2.112397	N-64[b]		2.495173
N-15		1.943623	N-65[b/c]		2.515579

Table 1 – *continued*

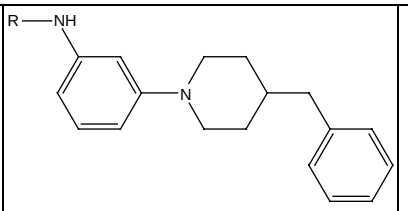
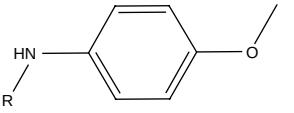
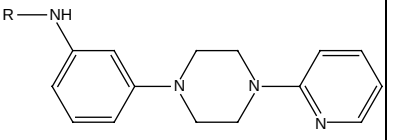
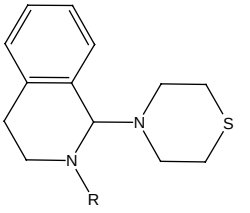
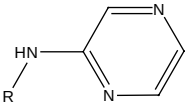
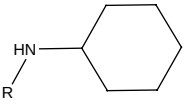
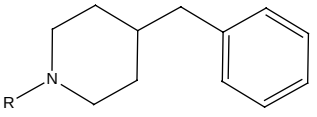
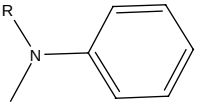
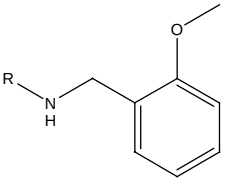
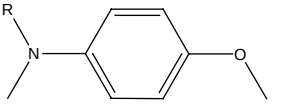
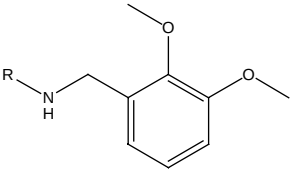
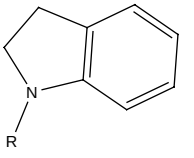
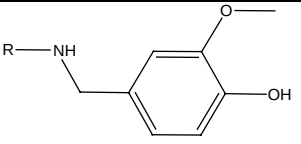
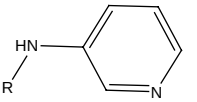
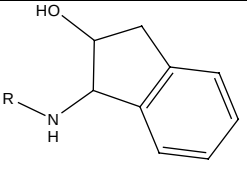
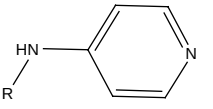
N-16		1.511031	N-66		2.816609
N-17[c]		1.798957	N-67		2.970152
N-18		1.573653	N-68[b]		1.885659
N-19		2.895341	N-69		1.896478
N-20		2.23718	N-70		1.64542
N-21		2.40693	N-71		2.508923
N-22		2.874687	N-72[b]		1.571818
N-23		1.487581	N-73		1.872848

Table 1 – *continued*

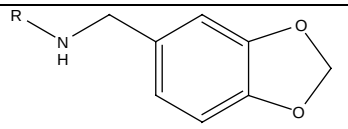
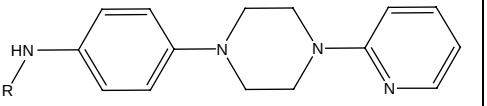
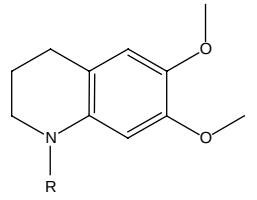
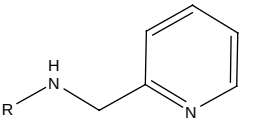
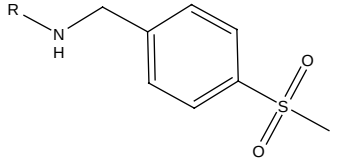
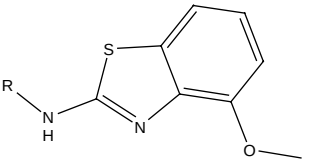
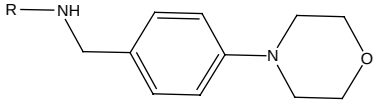
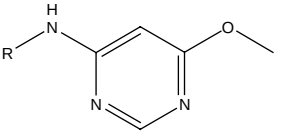
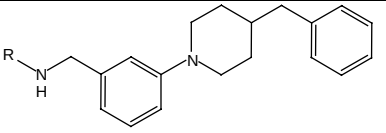
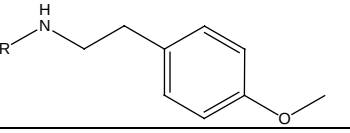
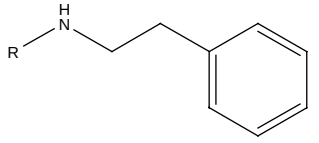
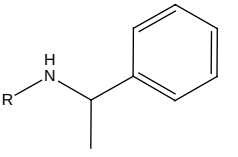
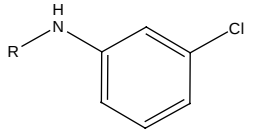
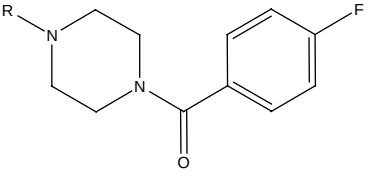
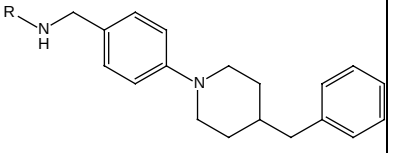
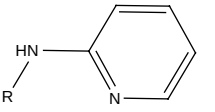
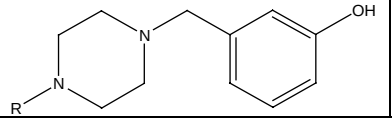
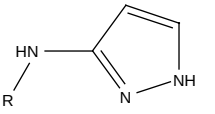
N-24		3.286654	N-74		2.992777
N-25		3.521547	N-75[c]		2.489977
N-26		3.280517	N-76[c]		2.300076
N-27		2.918198	N-77[c]		2.217809
N-28		2.429587	N-78[c]		2.559721
N-29[b]		2.21127	N-79[c]		2.21127
N-30		2.522836	N-80[c]		2.347584
N-31		2.719622	N-81[b/c]		1.872848
N-32		2.327134	N-82[c]		1.550786

Table 1 – *continued*

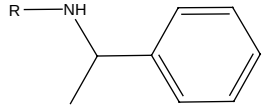
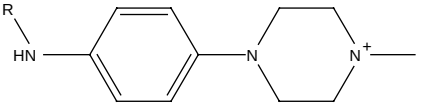
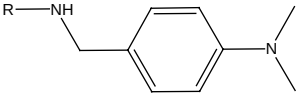
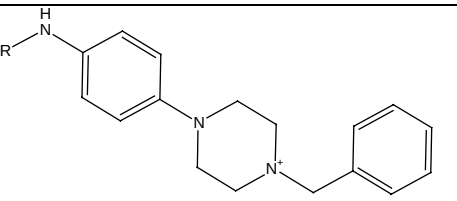
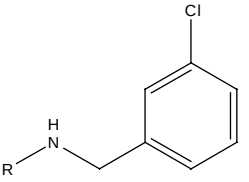
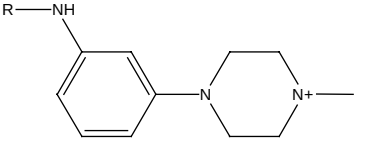
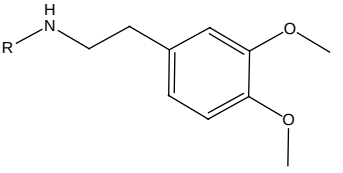
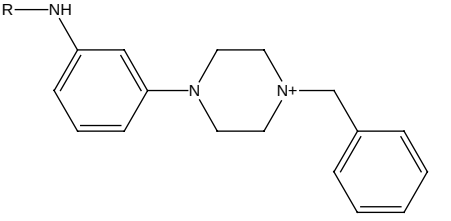
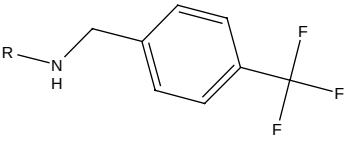
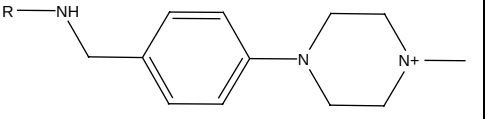
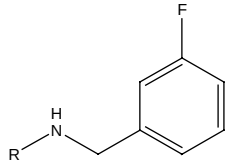
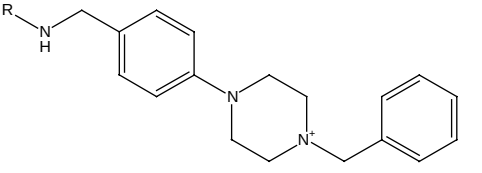
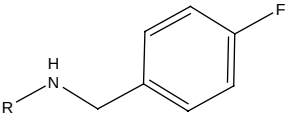
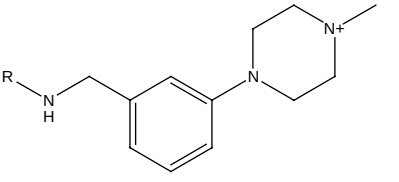
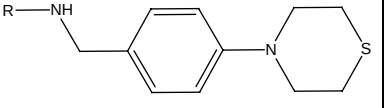
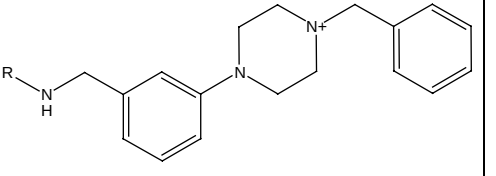
N-33[b]		1.924028	N-83		1.422056
N-34		2.859275	N-84[c]		2.705908
N-35[b]		3.147162	N-85		1.422056
N-36		2.9035	N-86		1.512088
N-37		3.798265	N-87		2.934966
N-38[b]		2.228829	N-88		4.526823
N-39[b]		2.518864	N-89		1.741146
N-40		2.347699	N-90		3.924763

Table 1 – *continued*

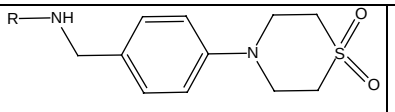
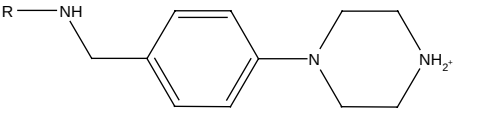
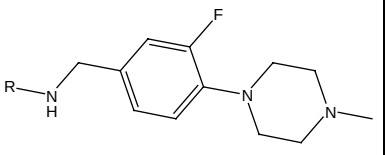
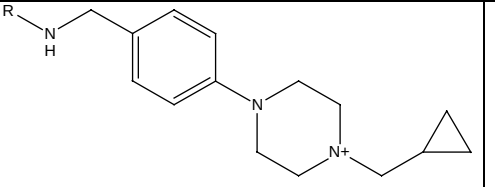
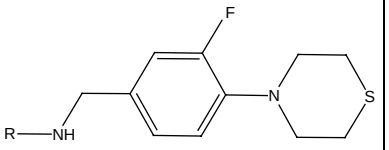
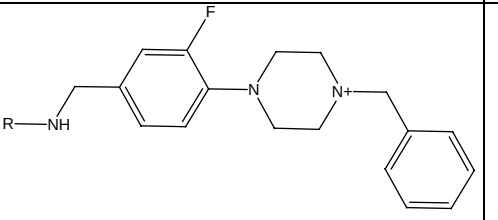
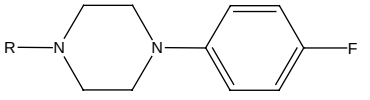
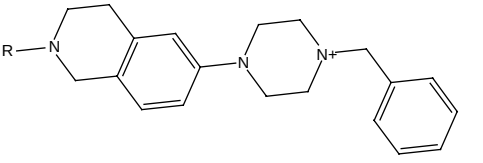
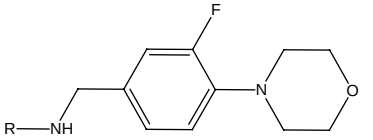
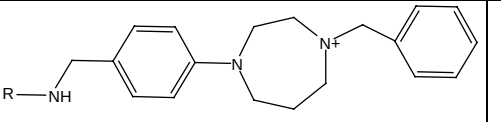
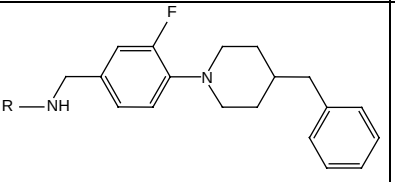
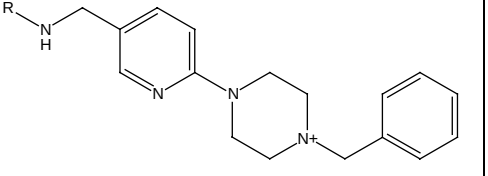
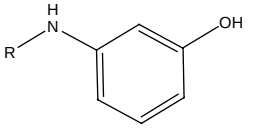
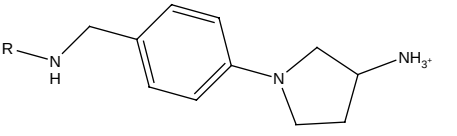
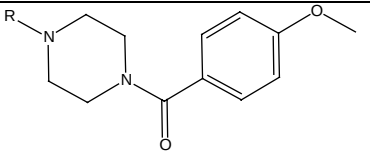
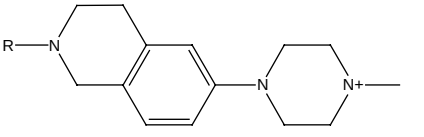
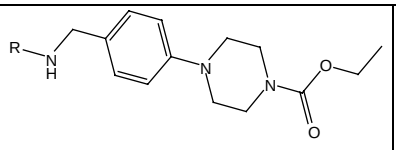
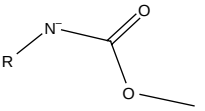
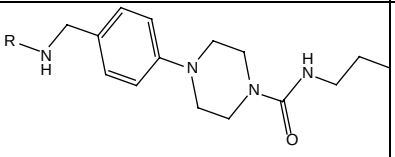
N-41		0.88012	N-91		2.615876
N-42		2.366017	N-92		3.885853
N-43[b]		2.369626	N-93		3.641929
N-44		3.203159	N-94		4.552917
N-45[b/c]		2.350097	N-95		2.734895
N-46		3.038889	N-96		4.83236
N-47		3.254474	N-97[c]		2.617199
N-48[b]		2.190673	N-98		3.869711

Table 1 – *continued*

N-49[b]		4.812274	N-99[c]		2.728627
N-50		3.317488			

[b]Indicates the compounds considered in the test set for two-dimensional QSAR study.
[c]Indicates the compounds considered in test set for three-dimensional QSAR study.

2.2 Feature selection and model development

The problem that is faced frequently by a researcher in developing QSAR models is that of a small number of observations (molecules) compared with a large number of molecular parameters in the descriptor pool. An integral aspect of any model-building exercise is the selection of an appropriate set of features with low complexity and good predictive accuracy. This process forms the basis of a technique known as feature selection [19] or variable selection. Among several search algorithms, stepwise method (SW) [20], GA [21] and SA [22]-based feature selection procedures are most appropriate for building predictive models considering either linear or nonlinear algorithms and can explain the situation more effectively.

In stepwise algorithm, the search procedure begins with developing a trial model step by step with a single independent variable and to each step; independent variables are added one at a time, examining the fit of the model by using the PLS cross-validation procedure. Thus, the model is repeatedly altered from the previous one by adding or removing a predictor variable in accordance with the 'stepping criteria' (in this case $F = 4$ for inclusion; $F = 3.5$ for exclusion for the forward–backward selection method). The method continues until there is no more significant variable remaining outside the model.

GA described by Holland is one of the most popular stochastic optimisation techniques that mimics natural evolution and selection [23]. It is a class of algorithms inspired by the process of natural evolution in which species having a high fitness under some conditions can prevail and survive to the next generation; the best species can be adapted by crossover and/or mutation in the search for better individuals. In this method, a chromosome and its fitness in the species represent a set of molecular descriptors and the cross-validated predictive accuracy of the derived QSAR model, respectively. The reproductive population and new population are combined and mutated with a predefined mutation rate. The best chromosome in the reproductive population is kept from the mutation process. This process is repeated over a number of 'generations' until the population converges to a small set of solutions. The cycle is repeated until the number of generations reaches a given maximum. The final model obtained is further refined by removing descriptors that do not affect predictive accuracy significantly. In this GA search, a gene pool of 100 has been used with a crossover rate (probability of mating) of 0.5 or 50% and a mutation rate (probability of random bit mutation) of 0.01 or 1%. The GA program returns only the optimum solution so that suboptimal solutions are not available for this report.

Another widely applied variable selection method in the QSAR studies is based on SA [22]. SA is a multivariate optimisation technique based on the Metropolis Monte Carlo algorithm for examining the equations of state and frozen states of n-body systems [24,25]. The concept is based on the manner in which liquids freeze or metals recrystalise in the process of annealing. In SA, the process starts from an initial state of very high temperature and introduces perturbations, or random moves, which create a new state. At high temperature, the system can overcome high-energy barriers due to the presence of temperature in the denominator of the Boltzmann factor; this enables it to explore the search space widely. As the temperature falls, the lower energy states become more probable. In a variable selection procedure, a move typically consists of adding or deleting a single variable from the set under consideration. The difference in fitness between the initial and perturbed states defines whether the move is accepted. If the fitness function is better, leading to an improved model (in terms of fitness function like cross-validated q^2), then the new subset is accepted for the next iteration. If the new subset is less predictive, then the probability, p that it is accepted depends on the Metropolis condition,

$\exp[-\Delta E/k_B T]$, where ΔE is the difference between the current and previous fitness values, k_B is the Boltzmann constant and T is the temperature. This offers the algorithm the possibility of jumping out of the local optimum [26]. The process continues for a fixed number of iterations or until no further improvement is observed in the model. In the SA variant, the temperature of the system is gradually reduced, hence increasing the chance of finding the globally optimal solution. The maximum and minimum temperatures in this analysis are set as 100,000 and 1 K, respectively, and the temperature is decreased by 5 U with a single iteration at that particular temperature.

The feature selection method can be used together with PLS regression analysis for constructing a QSAR model. PLS is a generalisation of regression, which can handle data with strongly correlated and/or noisy or numerous independent variables [27]. The linear PLS model finds new variables (latent variables or X scores) that are linear combinations of the original variables. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross-validation is a practical and reliable method of testing this significance [27].

2.3 Three-dimensional molecular modelling and alignment

The most crucial input for the three-dimensional QSAR modelling is the alignment of the molecules. Energy minimisations have been carried out using the MMFF and MMFF charge [28] and considering a dielectric constant of 1.0 and a convergence criterion of 0.01 kcal/mol. In the present study, the molecules of the dataset are aligned by a template-based method [29], where a template structure is defined and used as a basis for alignment of a set of molecules, and a reference molecule is chosen on which the other molecules of the dataset get aligned considering the chosen template. The template structure, i.e. nitrofuranyl amide ring, is used for alignment by considering the common elements of the series as shown in Figure 1. A potent synthetic analogue N-96 served as the reference structure and all other derivatives were aligned to it using the template alignment method in MDS, VLife Sciences (Figure 2). N-96, i.e. 5-nitro-furan-2-carboxylic acid [6-(4-benzyl-piperazin-1-yl)-pyridin-3-ylmethyl]-amide, shows very high-activity value against *M. tuberculosis* and has been considered as a potential lead molecule among these nitrofuranyl amide derivatives.

Molecular field analysis (MFA) model is predictive and sufficiently reliable to guide the chemist in designing novel compounds. This approach is effective for the analysis of data sets, where activity information is available but the structure of the receptor site is unknown. It attempts to postulate and represent the essential features of a receptor site from the aligned common features of the molecules that bind to it. The MFA calculates probe interaction energies on a rectangular grid around a bundle of active molecules. Atomic coordinates are used to compute field values on each point of a three-dimensional grid. Fields of molecules are represented using grids and the energy associated with each grid point can serve as input for the calculation of a QSAR. The molecular field is created using methyl group as a probe, which represents steric, electrostatic and hydrophobic fields. The major steps of MFA are: (1) generating conformers and energy minimisation; (2) matching atoms with the template structure and aligning molecules against the reference molecule using default option; (3) setting MFA preferences (rectangular grid with 2 Å step size, charges by MMFF, CH_3 as probe); (4) creating the field.

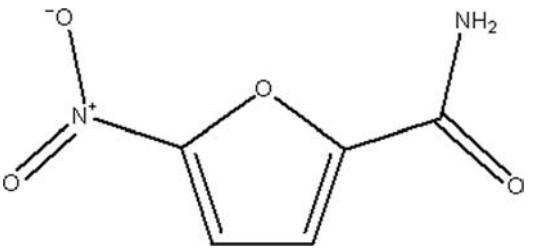


Figure 1. Nitrofuranyl amide ring (template structure).

2.4 Model quality and validation

The goodness-of-fit is expressed in terms of the squared correlation coefficient (r^2) and the root-mean-square error (RMSE); the robustness of the model, by the cross-validation coefficient (q^2) and the RMSE of cross-validated result ($RMSE_{CV}$); and the predictivity, by the external validation coefficient (pred_r^2) and the RMSE of prediction for the external validation set ($RMSE_P$). The coefficients are estimated as, $r^2 = (1 - \text{SSE/TSS})$, $q^2 = (1 - \text{PRESS/SSTotal})$, F ratio $= [(n-2)r^2/(1-r^2)]$ and RMSE = Square root of (SSE/n), where n is the number of observations, SSTotal is the total sum of squares and SSE is the sum of squares, PRESS is the predicted sum of squares based on leave-one-out (LOO) method [30]. The RMSE (also known as the SE of the estimate) is an indication of the quality of the fit. It is the SD of the data about the regression line, rather than about the sample mean. A large F indicates that the model fit is not a chance occurrence. If r^2 and q^2 are above a value of 0.6, it indicates a good model fit.

2.4.1 Internal and external validation

For all models, cross-validation was performed by the LOO procedures approach, in which each compound is removed, in turn, from the data set and the regression is fitted based on the remaining $(n-1)$ compounds. As it is based on the compounds that are external to the

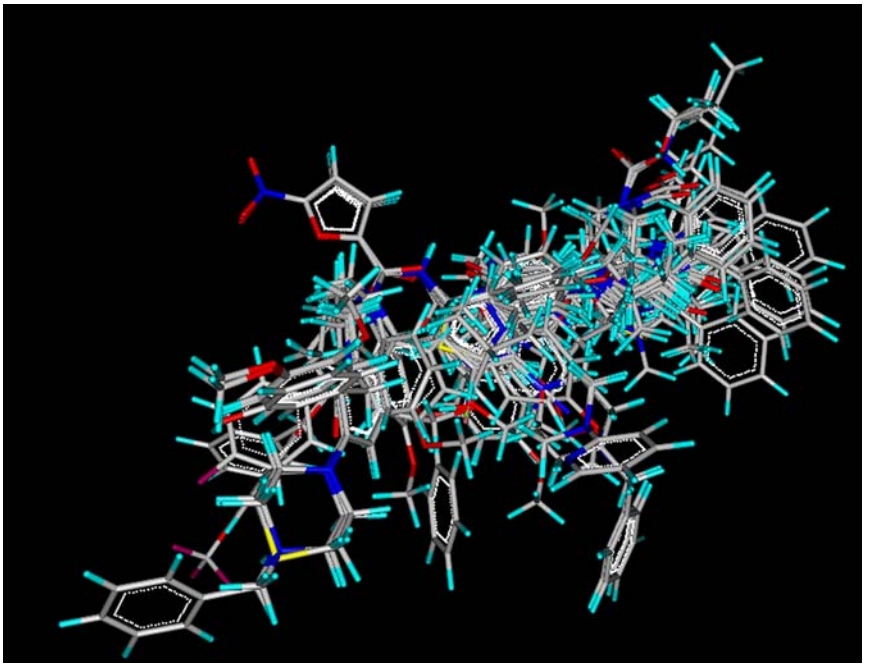


Figure 2. Alignment of 99 nitrofuranyl amide derivatives.

fitted regression, similar in this respect to using an external test set, it is a reliable measure of model predictability and this approach is preferred over the test-set approach when the sample size is small [31]. Unlike r^2, the cross-validated r^2, denoted by q^2, may be negative, signifying very poor model quality. In the case of nitrofuranyl amide derivatives without having an ionisable group and for three-dimensional QSAR model with all derivatives, the external predictivity was assessed by test-set compounds that are not considered in the model development. An external validation of models permits us to examine the power of the models to reproduce known experimental activities. The data sets were subdivided into training and test sets using the sphere exclusion method [32]. Thus, the above modelling procedure guarantees that the QSAR study develops a technically sound model having a good q^2 value and is simultaneously externally validated.

2.4.2 *Validation by y-randomisation*

Y-randomisation is another widely used technique to ensure the robustness of a QSAR model [33]. In this test, the dependent-variable vector, Y-vector, is randomly shuffled, and new QSAR models are developed using the original independent-variable matrix. This process is repeated several (typically, 10) times. It is expected that the resulting QSAR models from the randomised sets should generally have low training set fitness values r^2(random) and q^2(random). Sometimes, though infrequently, high training set q^2(random) may be obtained due to a chance correlation or structural redundancy of the training set [34]. If all QSAR models obtained in the Y-randomisation test exhibit relatively high training set q^2(random) values, it implies that an acceptable QSAR model cannot be built for the given dataset by the current modelling method.

3. Results and discussion

For the development of two-dimensional QSAR models for nitrofuranyl compounds, PLS methodology was used based on various feature selection methods viz. stepwise, GA and SA. A total number of 99 nitrofuranyl amide derivatives have been considered for the QSAR study, and among these, 17 compounds contain ionisable functional groups that would be expected to carry a charge at physiological pH. In order to account for this and to investigate the effect of protonation or deprotonation of these functional groups on model predictivity, separate models were built considering only the ionised compounds. Due to small data size (17 compounds), models are validated by the LOO cross-validation technique. For the rest 82 compounds, which are neutral in nature, the dataset is divided into training and test sets for an effective QSAR modelling with nitrofuranyl amide derivatives. The training- and test-set molecules for this group of compounds are selected by sphere exclusion method [32] and the models are validated by both internal and external validation procedures.

For three-dimensional QSAR, all the 99 molecules were taken into account considering 83 in the training set and remaining 16 in the test set. For three-dimensional QSAR, PLS model based on GA was built by considering the combination of both the fields (S and E). This model seemed to be the most predictive among others developed in the course of the study dealing with either a steric or electrostatic group of descriptors or applying other variable selection or model development methods. It was considered for further analysis because of its ability for providing the most descriptive information about such group of compounds.

3.1 *Modelling with two-dimensional descriptors*

With regard to QSAR modelling, our first goal was to establish a predictive model with a reasonable number of input features to ensure good generalisation performance. While correlating various descriptors with biological activity is the most important means to study structure–activity relationships, the interest lies in deciding when to stop adding a new descriptor to the model. Thus, the optimal model should use the minimum number of descriptors to obtain the best fit. To achieve this, a well-accepted method is to find out the saturation point, a point beyond which there is no considerable improvement in the regression coefficient (r^2 and q^2) values even if a new descriptor is added. SW, GA and SA were used in this study for selecting a significant set of descriptors in order to build the models by linear PLS technique. In this section, the prediction performances of the method proposed by three different models (SW-based PLS, GA-based PLS and SA-based PLS) were evaluated.

To ensure a fair comparison, the same training and test sets were used for each model's development. The GA–PLS, SA–PLS and SW–PLS models predicted the training data with an r^2 of 0.7012, 0.7045 and 0.7674 together with RMSE estimating to 0.5062, 0.5121 and 0.4504, respectively. However, the respective prediction result (pred_r^2) for the test set amounts to 0.7314, 0.6836 and 0.6838. The plots of calculated versus observed values of p[MIC] are shown in Figure 3 in the case of all three models. Equation (1) appears to be the best QSAR model obtained by the GA with 50,000 iterations.

$$\begin{aligned} \text{p[MIC]} = {} & 4.25664 - 4.9880(\text{2DACorr\_SigChg\_2}) \\ & + 0.174761(\text{SaaScount}) \\ & - 0.224219(\text{SssNHEindex}) \\ & - 0.0282489(\text{2DACorr\_LpEN\_4}) \\ & + 0.2067(\text{k3alpha}) \\ & - 3.40385(\text{XAMostHydrophobic}). \end{aligned} \tag{1}$$

$N_{\text{traing}} = 63$, $N_{\text{test}} = 19$, Optimum Components = 3, $r^2 = 0.7012$, $q^2 = 0.6337$, $F_{\text{test}} = 46.1482$, RMSE = 0.5062, $\text{RMSE}_{\text{CV}} = 0.5604$, pred_$r^2 = 0.7314$, $\text{RMSE}_{\text{P}} = 0.440$, $r^2(\text{random}) = 0.31744$, $q^2(\text{random}) = 0.03732$.

According to Tong et al. [35], the nature of the descriptors used, and more specifically the effectiveness in which descriptors encode the structural features of the molecules related to the activity, is far more important than the specific method employed. Although the physical interpretations of all descriptors are not always straightforward, it is still possible to gain some understanding of the molecular mechanisms from them.

Equation (1) shows the positive contribution of SaaScount indicating the total number of sulphur atoms connected with two aromatic bonds and another descriptor k3alpha, signifying third alpha-modified shape index. SssNHEindex, i.e. electrotopological state indices for the number of —NH groups connected with two single bonds and XAMostHydrophobic, which denotes the most hydrophobic value on the vdW surface, have a detrimental effect on the mycobacterial inhibition. Two autocorrelation descriptors 2DACorr_SigChg_2 and 2DACorr_LpEN_4 also have a negative influence on the activity.

The models obtained by SA/PLS and stepwise/PLS methodology also give a good prediction in terms of the pred_r^2 value and is denoted by Equations (2) and (3),

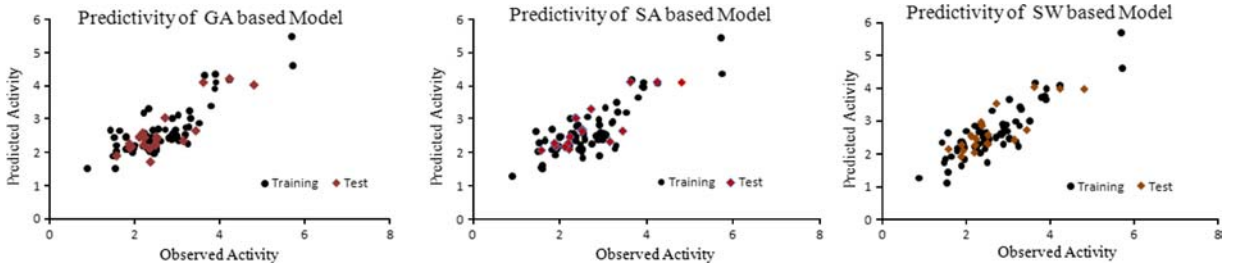


Figure 3. Prediction of two-dimensional QSAR models given by Equations (1)–(3).

respectively.

$$\begin{aligned} \text{p[MIC]} = & -0.00429553(\text{XKHydrophilicArea}) \\ & + 0.0106781(\text{2DACorr\_LpEN\_7}) \\ & - 0.0285051(\text{TPSA}) \\ & - 4.1251(\text{XAMostHydrophobic}) \\ & - 7.55757(\text{2DACorr\_SigChg\_2}) \\ & - 1.37465(\text{SaaOcount}). \end{aligned} \quad (2)$$

$N_{\text{traing}} = 63$, $N_{\text{test}} = 19$, Optimum Components = 5, $r^2 = 0.7045$, $q^2 = 0.5915$, $F_{\text{test}} = 27.1831$, RMSE = 0.5121, $\text{RMSE}_{\text{CV}} = 0.6021$, pred_$r^2 = 0.6836$, $\text{RMSE}_{\text{P}} = 0.4777$, $r^2(\text{random}) = 0.36217$, $q^2(\text{random}) = 0.14165$.

$$\begin{aligned} p\text{[MIC]} = & -0.452746 \\ & - 2.62291(\text{2DACorr\_SigChg\_2}) \\ & + 0.0125636(\text{2DACorr\_LpEN\_3}) \\ & - 0.0121868(\text{XKHydrophilicArea}) \\ & - 4.02764(\text{XAMostHydrophilic}) \\ & - 3.56897(\text{2DACorr\_PiChg\_9}) \\ & + 4.14804(\text{2DACorr\_PiChg\_11}). \end{aligned} \quad (3)$$

$N_{\text{traing}} = 63$, $N_{\text{test}} = 19$, Optimum Components = 4, $r^2 = 0.7674$, $q^2 = 0.7026$, $F_{\text{test}} = 47.8373$, RMSE = 0.4504, $\text{RMSE}_{\text{CV}} = 0.5093$, pred_$r^2 = 0.6838$, $\text{RMSE}_{\text{P}} = 0.4776$, $r^2(\text{random}) = 0.39342$, $q^2(\text{random}) = 0.19896$.

In the case of nitrofuranyl amide compounds with ionisable groups, the GA-based PLS model gives the best fit and is represented by Equation (4).

$$\begin{aligned} p\text{[MIC]} = & -3.28281 \\ & + 0.00757593(\text{ZZPolarisability}) \\ & - 0.374279(\text{HOMOEnergy}) \\ & + 0.0449251(\text{XKHydrophilicArea}) \\ & + 0.0493889(\text{RadiusOfGyration}). \end{aligned} \quad (4)$$

$N = 17$, Optimum Components = 1, $r^2 = 0.6869$, $q^2 = 0.6720$, $F_{\text{test}} = 32.9034$, RMSE = 0.6492, $\text{RMSE}_{\text{CV}} = 0.6644$, $r^2(\text{random}) = 0.57515$, $q^2(\text{random}) = 0.13030$.

In the above equation, ZZPolarisability (induced polarisability along ZZ axis), XKHydrophilicArea (vdW surface descriptor showing hydrophilic surface area) and RadiusOfGyration (size descriptor for the distribution of atomic masses in a molecule) have a positive influence on the anti-tuberculosis activity values of the compounds, whereas HOMOEnergy signifying the energy of the highest occupied molecular orbital has a detrimental effect on the inhibitory activity. The HOMO energy reflects the compounds' ability to donate electrons in a charge-transfer type of interaction, possibly charge-transfer complex forms between the electron-rich aromatic ring at the secondary amine position of nitrofuranyl amide molecules and the electron-deficient component of the receptor active site.

The models obtained on the same nitrofuranyl amides with ionisable groups by other feature selection methods viz. SA and stepwise, are given below as Equations (5) and (6), respectively, along with the corresponding statistical parameters.

$$\begin{aligned} p\text{[MIC]} = & -3.23572 \\ & + 1.0496(\text{RotatableBondCount}) \\ & - 0.520285(\text{SssNHEindex}) \\ & + 0.0620296(\text{XKHydrophilicArea}) \\ & - 0.191292(\text{chiV3Cluster}). \end{aligned} \quad (5)$$

Optimum Components = 3, $n = 17$, $r^2 = 0.7583$, $q^2 = 0.6234$, $F_{\text{test}} = 13.5916$, RMSE = 0.6128, $\text{RMSE}_{\text{CV}} = 0.7648$, $r^2(\text{random}) = 0.57266$, $q^2(\text{random}) = 0.14637$.

$$\begin{aligned} p\text{[MIC]} = & 2.71015 + 0.05168(\text{YZPolarisability}) \\ & + 1.43193\text{e} - 007(\text{HosoyaIndex}) \\ & - 0.00560712(\text{SKHydrophilicArea}) \\ & + 1.4061(\text{SaaNcount}). \end{aligned} \quad (6)$$

Optimum Components = 1, $n = 17$, $r^2 = 0.7556$, $q^2 = 0.5697$, $F_{\text{test}} = 46.3682$, RMSE = 0.5736, $\text{RMSE}_{\text{CV}} = 0.7611$, $r^2(\text{random}) = 0.55567$, q^2 (random) = 0.33016.

The model represented by Equation (4), where the GA method is considered for variable selection, gives the best q^2 value compared with other methods. Moreover, the r^2 and q^2 values are comparable for the prediction of activity for such compounds, but in the case of Equations (5) and (6) cross-validated prediction is much poor than the corresponding r^2 value. Thus, it is evident that Equation (4) appears to be a more reliable model for the prediction of anti-tubercular activity of the nitrofuranyl compounds having an ionisable group.

Comparing the above models viz. Equations (1)–(6) developed by various feature selection methods, it is seen

that GA offers a better solution than SW or SA when applied to such nitrofuranyl amide derivatives. The GA algorithm approach has a number of important advantages over other techniques, such as it builds multiple models rather than a single model and it automatically selects features that can be used in the models. The above QSAR models are not based on properties that may be influenced by three-dimensional arrangement of the nitrofuranyl derivatives. In fact, the variables appearing in the model are two-dimensional descriptors, independent of the conformations used during calculations. For a better understanding of the QSAR models of such nitrofuranyl compounds, an attempt to generate a three-dimensional QSAR model based on MFA for the same set of compounds has also been made.

3.2 *Modelling with three-dimensional descriptors*

Three-dimensional QSAR studies of nitrofuranyl amide derivatives having inhibitory activities against *M. tuberculosis* have been performed using GA-based variable selection methods for developing PLS models. A total of 99 nitrofuranyl compounds were taken into account for the three-dimensional QSAR modelling and among these, 83 derivatives constitute the training set for the model development and the remaining 16 compounds were chosen as the test set for validation purpose. The training and test sets for such a study were selected by the sphere exclusion method. The model quality for the training set was determined by calculating r^2 and r^2_{cv}, whereas external validation was determined by predictive r^2 (pred_r^2) for the test-set compounds. The best model in terms of the associated statistical parameters for 83 training-set molecules was developed and represented by Equation (7). In this equation, the steric (S) and electrostatic (E) descriptors specify the regions, where variations in the structural features (steric or electrostatic) of different compounds in the training set lead to increase or decrease in activities. The number accompanied by the descriptors represents its position in the three-dimensional MFA grid. GA/PLS was carried out over 50,000 generations with a population size of 100. The optimal number of components considered for the model was four. An energy cut-off of 30 and 10 kcal/mol was set for steric and electrostatic contributions, respectively. The PLS model for steric and electrostatic fields obtained with GA-based descriptor selection methodology is given by

$$\begin{aligned} p[\mathrm{MIC}] = {} & 2.1256 - 27.3373(\mathrm{S\_3006}) \\ & - 290.595(\mathrm{S\_301}) + 0.031719(\mathrm{S\_2002}) \\ & - 0.0265192(\mathrm{S\_2516}) \\ & + 0.060298(\mathrm{E\_1704}) \\ & + 0.0291196(\mathrm{S\_2551}). \end{aligned} \tag{7}$$

$n = 83$, $r^2 = 0.6919$, $q^2 = 0.6471$, $F_{\mathrm{test}} = 43.7875$, RMSE $= 0.5555$, $\mathrm{RMSE_{CV}} = 0.5945$, pred_$r^2 = 0.6849$, $\mathrm{RMSE_P} = 0.4079$, r^2(random) $= 0.25640$, q^2(random) $=$ 0.05786.

The best LOO cross-validated value (q^2) of PLS analysis was found to be 0.6471, suggesting that the model could be useful for predicting TB inhibitory activity for such nitrofuranyl amides. The correlation coefficient between the calculated and experimental activities amounts to a value of 0.6919 with an SE of 0.555. The respective relative contributions of steric and electrostatic fields were 0.906 and 0.094, indicating that steric field is more predominant. The graph of observed activity versus predicted activities of the training-set molecules from MFA analysis is illustrated in Figure 4. The three-dimensional QSAR model was validated using 16 test-set compounds (indicated in Table 1) that were not included in the development of the model. Figure 4 also represents the graph of the actual versus predicted p(MIC) values of the test-set molecules for this model.

The contribution plot representations of the three-dimensional QSAR results for TB inhibitors are presented in Figure 5. To aid the visualisation, the most potent compound N-96 is overlaid on the map. The green-coloured balls specify the positions of the steric descriptors and the descriptors with positive coefficients indicate the areas where the steric bulky group enhances

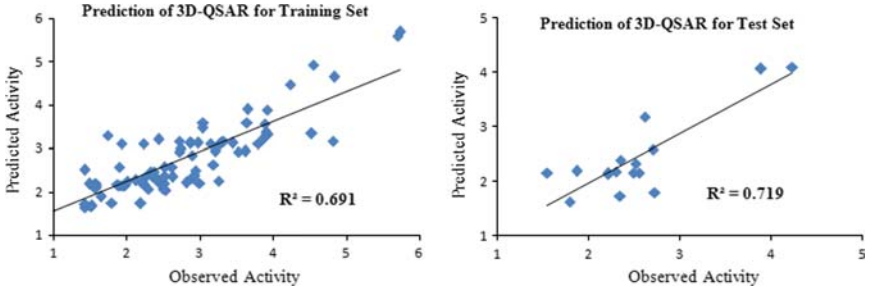


Figure 4. Prediction of three-dimensional QSAR model for training and test sets given by Equation (7).

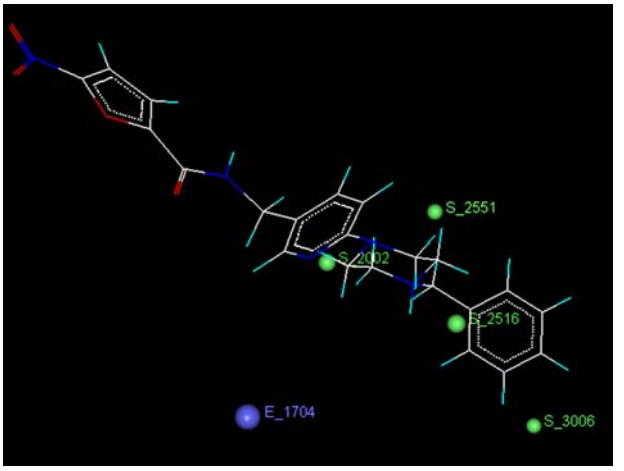


Figure 5. Contribution plot for steric and electrostatic interactions considering the lead molecule (N-96).

TB inhibitory activity, while the descriptors with negative coefficient point out regions where the steric bulky group is detrimental for the biological activity. In the present study, steric three-dimensional descriptors like S_2002 (21.8%) and S_2551 (9.89%) with a positive coefficient present around the benzamide group indicate favourable substitutions of a bulky group in the para position of the benzyl ring. Most of the compounds (like N-53 to N-56, N-60 to N-63, etc) with higher activity having bulky substitution at the para or meta position of the benzylamide group strongly support the above comment. Other steric descriptors S_2516 (−16.8%), S_301 (−20.76%) and S_3006 (−21.35%) with negative coefficients indicate the regions, where less bulky group substitutions are preferable for better anti-tubercular activity. Only one electrostatic interaction, E_1704 (9.41%), with a positive coefficient has been considered in the model, represented by a blue-coloured ball, showing the area where electropositive charged groups enhance TB inhibitory activity.

In the randomisation study, none of the identified models have shown any chance correlation. We have obtained consistently much higher r^2 and q^2 values for the actual data set compared to those from Y-randomisation r^2(random) and q^2(random). Three models viz. Equation (1) for nitrofuranyl amides that do not have ionisable groups, Equation (4) for nitrofuranyl amides having ionisable groups and Equation (7) representing the three-dimensional QSAR model, seem to perform well. The q^2 values in case of Equation (1), Equations (4) and (7) for the real data set are 0.6337, 0.6720 and 0.64710, while they were 0.03732, 0.13030 and 0.05786, respectively, for the random data sets. This demonstrated that the high q^2 values of the models for the real data sets were not due to chance correlations.

4. Conclusion

A number of feature selection methods have been applied to the same group of descriptors for a set of nitrofuranyl amides with a view to developing QSAR models. A total number of 99 derivatives were selected for this study from various publications of Lee and others [5–11] and with such a complex data set, an objective selection viz. SW, SA and the GA derived from a chemometric study can lead to better predictive models. For such nitrofuranyl amide derivatives considered in the present study, multiple linear regression does not produce good quality models after applying different feature selections; however, the model quality improves significantly when the data are treated with PLS regression in conjunction with a descriptor selection method. Among various combinations, GA-based partial least-squares method provides the best results for training and test sets. The GA/PLS model is further validated by a randomisation test, which confirms that the QSAR is not a result of chance correlation. The overall results of the conducted comparative QSAR analysis bring more insight into the nature and structural dominants of the studied classes of nitrofuranyl amides and, if necessary, can help rationalising the design and discovery of novel antimicrobials and human therapeutics in future.

Acknowledgements

Payel Ghosh thanks the Council of Scientific and Industrial Research, New Delhi 110001, India, for the grant of a Senior Research Fellowship to her. MCB acknowledges the Department of Biotechnology, New Delhi, India, for the grant of a project.

References

[1] WHO International Fact Sheet on *Tuberculosis* No. 104, 2007. Available at http://www.who.int/mediacentre/factsheets/fs104/en/index.html.
[2] *Treatment of Tuberculosis: Guidelines for National Programmes* WHO/CDS/TB 2003.313, 3rd ed., WHO, Geneva, 2003.
[3] A.M. Ginsberg and M. Spigelman, *Challenges in tuberculosis drug research and development*, Nat. Med. 13 (2007), pp. 290–294.
[4] J.C. Sacchettini, E.J. Rubin, and J.S. Freundlich, *Drugs versus bugs: in pursuit of the persistent predator Mycobacterium tuberculosis*, Nat. Rev. Microbiol. 6 (2008), pp. 41–52.
[5] R.P. Tangallapally, R. Yendapally, R.E. Lee, K. Hevener, V.C. Jones, A.J. Lenaerts, M.R. McNeil, Y. Wang, S. Franzblau, and R.E. Lee, *Synthesis and evaluation of nitrofuranylamides as novel antituberculosis agent*, J. Med. Chem. 47 (2004), pp. 5276–5283.
[6] R.P. Tangallapally, R. Yendapally, R.E. Lee, and A.J. Lenaerts, *Synthesis and evaluation of cyclic secondary amine substituted phenyl and benzyl nitrofuranyl amides as novel antituberculosis agents*, J. Med. Chem. 48 (2005), pp. 8261–8269.
[7] R.P. Tangallapally, R.E. Lee, A.J. Lenaerts, and R.E. Lee, *Synthesis of new and potent analogues of anti-tuberculosis agent 5-nitro-furan-2-carboxylic acid 4-(4-benzyl-piperazin-1-yl)-benzylamide with improved bioavailability*, Bioorg. Med. Chem. Lett. 16 (2006), pp. 2584–2589.
[8] R.P. Tangallapally, R. Yendapally, A.J. Daniels, R.E. Lee, and R.E. Lee, *Nitrofurans as novel anti-tuberculosis agents: identification, development and evaluation*, Curr. Top. Med. Chem. 7 (2007), pp. 509–526.

[9] N.R. Budha, N. Mehrotra, R. Tangallapally, Rakesh A. Daniels, R.E. Lee, and B. Meibohm, *Pharmacokinetically-guided lead optimization of nitrofuranylamide anti-tuberculosis agents*, AAPS J. 10 (2008), pp. 157–165.

[10] J.G. Hurdle, R.B. Lee, N.R. Budha, E.I. Carson, J. Qi, M.S. Scherman, S.-H. Cho, M.R. McNeil, A.J. Lenaerts, S.G. Franzblau, et al., *A microbiological assessment of novel nitrofuranylamides as anti-tuberculosis agents*, J. Antimicrob. Chemother. 62 (2008), pp. 1037–1045.

[11] N.R. Budha, R.E. Lee, and B. Meibohm, *Biopharmaceutics, pharmacokinetics and pharmacodynamics of antituberculosis drugs*, Curr. Med. Chem. 15 (2008), pp. 809–825.

[12] J.M. Sutter, S.L. Dixon, and P.C. Jurs, *Automated descriptor selection for quantitative structure–activity relationships using generalized simulated annealing*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 77–84.

[13] D. Rogers and A.J. Hopfinger, *Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure-property relationships*, J. Chem. Inf. Comput. Sci. 34 (1994), pp. 854–866.

[14] Molecular Design Suite 3.5, VLife Technologies, Pune, India. Available at www.vlifesciences.com

[15] ADRIANA.code, Available at http://www.molecular-networks.com/ (accessed March 2008).

[16] P. Ertl, B. Rohde, and P. Selzer, *Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties*, J. Med. Chem. 43 (2000), pp. 3714–3717.

[17] A. Yan, J. Gasteiger, M. Krug, and S. Anzali, *Linear and nonlinear functions on modeling the aqueous solubility of organic compounds by two structure representation methods*, J. Comput.-Aided Mol. Des. 18 (2004), pp. 75–87.

[18] M. Wagener, J. Sadowski, and J. Gasteiger, *Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks*, J. Am. Chem. Soc. 117 (1995), pp. 7769–7775.

[19] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, J. Mach. Learn. Res. 3 (2003), pp. 1157–1182.

[20] R.B. Darlington, *Regression and Linear Models*, McGraw-Hill, New York, 1990.

[21] K. Hasegawa, T. Kimura, and K. Funatsu, *GA strategy for variable selection in QSAR studies: enhancement of comparative molecular binding energy analysis by GA-based PLS method*, Quant. Struct. Act. Relat. 18 (1999), pp. 262–272.

[22] W. Zheng and A. Tropsha, *Novel variable selection quantitative structure-property relationship approach based on the knearest-neighbor principle*, J. Chem. Inf. Comput. Sci. 40 (2000), pp. 185–194.

[23] J.H. Holland, *Genetic algorithms*, Sci. Am. 267 (1992), pp. 66–72.

[24] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, and A.H. Teller, *Equation of state calculations by fast computing machines*, J. Chem. Phys. 21 (1953), pp. 1087–1092.

[25] S. Kirkpatrick, C.D. Gelatt, Jr, and M.P. Vecchi, *Optimization by simulated annealing*, Science 220 (1983), pp. 671–680.

[26] S.S. Kulkarni and V.M. Kulkarni, *Three-dimensional quantitative structure-activity relationship of interleukin 1-beta converting enzyme inhibitors: a comparative molecular field analysis study*, J. Med. Chem. 42 (1999), pp. 373–380.

[27] S. Wold, *PLS for multivariate linear modeling*, in *Chemometric Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 195–218.

[28] T.A. Halgren, *Merck molecular force field. III. Molecular geometries and vibrational frequencies*, J. Comp. Chem. 17 (1996), pp. 553–586.

[29] S. Ajmani, K. Jadhav, and S.A. Kulkarni, *Three-dimensional QSAR using the k-Nearest Neighbor method and its interpretation*, J. Chem. Inf. Model. 46 (2006), pp. 24–31.

[30] A.R. Leach, *Molecular Modelling Principles and Applications*, 2nd ed., Pearson Education EMA, UK, 2001, pp. 698–702.

[31] D.M. Hawkins, S.C. Basak, and D. Mills, *Assessing model fit by cross-validation*, J. Chem. Inf. Comput. Sci. 43 (2003), pp. 579–586.

[32] A. Golbraikh and A. Tropsha, *Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection*, J. Comput.-Aided Mol. Des. 16 (2002), pp. 357–369.

[33] S. Wold and L. Eriksson, *Statistical validation of QSAR results*, in *Chemometrics Methods in Molecular Design*, H. van de Waterbeemd, ed., VCH, Weinheim, 1995, pp. 309–318.

[34] A. Tropsha, P. Gramatica, and V.K. Gombar, *The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models*, QSAR Comb. Sci. 22 (2003), pp. 69–77.

[35] W. Tong, H. Fang, H. Hong, Q. Xie, R. Perkins, J. Anson, and D.M. Sheehan, *Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: a perspective*, Pure Appl. Chem. 75 (2003), pp. 2375–2388.