AVCpred: An integrated web server for prediction and design of antiviral compounds

(Running Title: Antiviral compound prediction)

Abid Qureshi¹, Gazaldeep Kaur² and Manoj Kumar³,*

*Correspondence to: Manoj Kumar (E-mail: manojk@imtech.res.in)

¹Abid Qureshi
Bioinformatics Centre, Institute of Microbial Technology, Council of Scientific and Industrial Research, Sector 39-A, Chandigarh-160036, India
E-mail: qureshisaab@imtech.res.in

²Gazaldeep Kaur
Bioinformatics Centre, Institute of Microbial Technology, Council of Scientific and Industrial Research, Sector 39-A, Chandigarh-160036, India
E-mail: gazaldeep@imtech.res.in

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/cbdd.12834
This article is protected by copyright. All rights reserved.
Abstract

Viral infections constantly jeopardize the global public health due to lack of effective antiviral therapeutics. Therefore, there is an imperative need to speed up the drug discovery process to identify novel and efficient drug candidates. In the present study, we have developed Quantitative structure-activity relationship (QSAR) based models for predicting Antiviral compounds (AVCs) against deadly viruses like Human immunodeficiency virus (HIV), Hepatitis C virus (HCV), Hepatitis B virus (HBV), Human herpesvirus (HHV) and 26 others using publicly available experimental data from the ChEMBL bioactivity database. Support Vector Machine (SVM) models achieved a maximum Pearson Correlation Coefficient of 0.72, 0.74, 0.66, 0.68 and 0.71 in regression mode and a maximum Matthew's correlation coefficient 0.91, 0.93, 0.70, 0.89 and 0.71 respectively in classification mode during 10-fold cross-validation. Furthermore, similar performance was observed on the independent validation sets. We have integrated these models in the AVCpred web server, freely available at http://crdd.osdd.net/servers/avcpred. In addition, the datasets are provided in a searchable format. We hope this web server will assist researchers in the identification of potential antiviral agents. It would also save time and cost by prioritizing new drugs against viruses before their synthesis and experimental testing.

This article is protected by copyright. All rights reserved.
Keywords: antiviral compounds, drug design, QSAR, inhibition, prediction, algorithm

Introduction

Antiviral compounds (AVCs) inhibit the development of viruses in the host cell and are relatively harmless to the host (1). They can be natural, e.g., antivirals found in turmeric (2) and eucalyptus oil (3) or synthetic, e.g., zidovudine (a nucleoside analog) (4) and tamiflu (neuraminidase inhibitor) (5). Many compounds and drugs have also been tested and found to be useful in restricting the growth of certain viruses (6, 7). Scientists are endeavoring to broaden the range of antivirals to other families of viruses (8).

However, designing safe and effective antiviral drugs is a difficult task due to the high genetic diversity and consequent drug resistance in viruses (9). Initially, antivirals were discovered using traditional trial-and-error methods (10). However, it was a very lengthy process for discovering effective antivirals (10, 11). Later, research on virology helped to identify many target pathways to block viral multiplication (12, 13). Scientists are now using rational drug design strategies for developing antivirals that target the viruses at different stages of their life cycles (14). During the past decade many new drugs have been successfully identified in controlling the viral replication in host cells e.g., maraviroc (inhibits Human immunodeficiency virus or HIV entry), pleconaril (inhibits picornavirus uncoating), acyclovir (inhibits herpesvirus replication), oseltamivir (inhibits influenza release) etc. (9, 15).

To save time and money for discovering a new drug, researchers have widely used various computational methods to screen virtual libraries of compounds before the synthesis and animal testing of chemicals. Among the different approaches, Quantitative structure-activity relationship
QSAR is mostly used (16-18). In this approach, relationships connecting molecular descriptors and activity are used to predict the property of other molecules (19). Molecular descriptors transform the chemical information (structure and linking of groups) of a molecule into simple numbers. (20). QSAR based virtual screening is an effective computational technique leading towards identification and design of novel antiviral agents (21).

Lately, many dedicated bioinformatics resources have been developed for antivirals. For example, in the area of RNA interference resources published are VIRsiRNAdb -antiviral siRNAs resource for about 42 disease causing viruses (22), HIVsirDB -anti HIV siRNAs database (23), VIRsiRNAPred -antiviral siRNA inhibition efficacy predictor (24) and VIRmiRNA -database of virus encoded miRNAs including antiviral miRNAs (25). Similarly, for peptides based antivirals, a few web servers have also been created like AVPdb -collection of antiviral peptides targeting more than 60 medically important viruses (26), HIPdb -HIV inhibiting peptides repository (27) and AVPpred -predictor of antiviral activity of peptides (28).

Many general depositories provide information of antiviral molecules. For example, ChEMBL (29), PubChem - a database of molecules and their activities (30), ZINC -database of commercial compounds for virtual screening (31), DrugBank - a knowledgebase for drugs and drug targets (32) etc. In addition there are a few QSAR studies targeting specific viral proteins (33-41). However, till date there is no web server/software, which can regressively predict the percentage inhibition value of a compound against different human viruses under a single platform.

To cater this need, we developed AVCpred, a web server for prediction and design of antiviral compounds. In this method, we used previously known AVCs against HIV, Hepatitis C virus (HCV), Hepatitis B virus (HBV), Human herpesvirus (HHV) and 26 other viruses with
experimentally validated percentage inhibition from ChEMBL, a large-scale bioactivity database for drug discovery (29). This was followed by descriptor calculation and selection of best performing molecular descriptors. The latter were then used as input for Support Vector Machine (in regression mode) to develop QSAR models for different viruses as well as a general model for other viruses. We have integrated these models in the AVCPred web server, which will be helpful for virtual screening of AVCs and designing new compounds to target the viruses.

Methods and Materials

Datasets

In this study, we have used different datasets of AVCs having experimentally verified percent inhibition values against HIV, HCV, HHV, HBV and a general dataset having AVCs against 26 human viruses. The data was obtained from the ChEMBL resource (https://www.ebi.ac.uk/chembl/). The desired data was fetched using target browser (taxonomy tree) as well as target search using keywords like HIV, HCV, HBV, HHV, virus, viral, viruses etc. Initially, among the AVCs the majority of data belonged to HIV (1383 compounds), HCV (803 compounds), HHV (473 compounds), HBV (416 compounds) and other viruses (1635 compounds). After filtering entries with desired information and removing redundant entries, we were left with 389 compounds for HIV, 467 in case of HCV, 124 for HHV, 112 against HBV and 1391 AVCs targeting the 26 viruses (Table 1 and Supplementary Table S1). These datasets were used for descriptor selection and model development. The datasets are available along with references on the web server and can be downloaded from this url: http://crdd.osdd.net/servers/avcpred/datasets.php

This article is protected by copyright. All rights reserved.
**Descriptor calculation**

To develop virus specific as well as general QSAR models, we computed about 18000 chemical descriptors (1D, 2D, and 3D), including geometric, constitutional, electrostatic, topological, hydrophobic, binary fingerprints etc. using PaDEL, an open source software to calculate molecular descriptors and fingerprints. (42).

**Feature selection**

To improve the speed of calculation we selected the most essential descriptors using 'RemoveUseless' filter followed by ClassifierSubsetEval (attribute evaluator) with BestFirst (search method) module available in Weka package (43). ClassifierSubsetEval evaluates attribute subsets on training/testing data using a classifier to estimate the merit of a set of attributes (44, 45). The selected descriptors were then used to develop the QSAR models (Supplementary Table S3).

**Machine learning**

We developed individual QSAR models for each of the 4 viruses (HIV, HCV, HHV and HBV) as well as a general model comprising 26 different viruses using SMOreg algorithm (46) in Weka machine learning software (43) freely available at http://www.cs.waikato.ac.nz/ml/weka. SMOreg implements the support vector machine in regression mode. In SMOreg, Pearson VII function-based universal kernel (Puk) and RegSMOImproved optimizer were used along with parameters such as i) the regularization constant/complexity value (c) that allows trade-off between training error and margin, ii) the omega exponent value (ω) which controls peak half-width and iii) the sigma bandwidth value (σ) that controls peak tailing factor (47, 48). Simultaneously, software SVMlight (freely available at http://svmlight.joachims.org) was
employed for machine learning in classification mode. In SVM\textsuperscript{light}, Radial basis function (RBF) kernel was used with parameters i) gamma (g) which defines how far the influence of a single training example reaches and ii) complexity constant (c) that allows trade-off amid training error and margin (49). Selected molecular descriptors and fingerprints were used as input features for the development of QSAR models.

Evaluation

In order to evaluate performance of our models, we employed a number of statistical parameters including Pearson’s correlation coefficient, Coefficient of Determination, Mean absolute error Root-mean-square error, Sensitivity, Specificity, Accuracy and Mathew’s correlation coefficient as briefly described below.

The **Pearson’s correlation coefficient (R)** is a measure of correlation between two variables.

$$R = \frac{n \sum_{i=1}^{n} E_i^{\text{act}} E_i^{\text{pred}} - \sum_{i=1}^{n} E_i^{\text{act}} \sum_{i=1}^{n} E_i^{\text{pred}}}{\sqrt{n \sum_{i=1}^{n} (E_i^{\text{act}})^2 - (\sum_{i=1}^{n} E_i^{\text{act}})^2} \sqrt{n \sum_{i=1}^{n} (E_i^{\text{pred}})^2 - (\sum_{i=1}^{n} E_i^{\text{pred}})^2}}$$

where \( n \) is the size of test set, \( E_i^{\text{pred}} \) and \( E_i^{\text{act}} \) is the predicted and actual efficacy of AVCs respectively.

A value of 1 denotes total positive correlation, 0 is no correlation and −1 is total negative correlation.

The **Coefficient of Determination (R}^2\)** indicates how well data fit a statistical model. An R}^2 of 1 indicates that the model perfectly fits the data, while an R}^2 of 0 means that the model does not fit the data at all.

The **Mean absolute error (MAE)** measure indicates how close the predictions are to the eventual outcomes.

This article is protected by copyright. All rights reserved.
\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |E_{i}^{\text{pred}} - E_{i}^{\text{act}}| \]  

(2)

where \( E_{i}^{\text{pred}} \) is the prediction, \( E_{i}^{\text{act}} \) the true value and \( |E_{i}^{\text{pred}} - E_{i}^{\text{act}}| \) the absolute error.

MAEs are negatively-oriented scores, that is, lower values are better.

The **Root-mean-square error (RMSE)** measures the average magnitude of the error.

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (E_{i}^{\text{pred}} - E_{i}^{\text{act}})^2} \]  

(3)

RMSEs are also negatively-oriented scores, that is, lower values are better.

**Sensitivity** \((S_n)\) or the true positive rate measures the percentage of correctly identified positives.

\[ S_n = \frac{TP}{(TP+FN)} \times 100 \]  

(4)

An ideal predictor would be expressed as 100% sensitive.

**Specificity** \((S_p)\) or the true negative rate measures the percentage of correctly identified negatives

\[ S_p = \frac{TN}{(TN+FP)} \times 100 \]  

(5)

An ideal predictor would be expressed as 100% specific.

**Accuracy** \((A_c)\) is the percentage of correct results (i.e. both true positives as well as true negatives) among the total number of cases.

\[ A_c = \frac{TP+TN}{(TP+FP+TN+FN)} \times 100 \]  

(6)

An ideal predictor would be expressed as 100% accurate.
The **Matthew’s correlation coefficient (MCC)** is used in machine learning to evaluate the performance of binary classifications.

\[
\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\] (7)

In the above (4-7) equations TP, FP, TN and FN represent the true positives, false positives, true negatives and false negatives respectively.

Its value ranges from −1 to 1 and a value close to 1 means a better prediction.

**Results**

**Performance of QSAR models**

In order to identify the most effective features or descriptors of antiviral drugs, we computed the correlation between selected chemical features of antiviral drugs and their percent inhibition using comprehensive pharmacological screening datasets from ChEMBL (29) (Figure 1).

After attribute selection, the relevant descriptors were 45 for HIV, 52 for HCV, 15 for HBV, 20 for HHV and 65 for rest of the viruses. A combination of selected chemical descriptors like partial charge, atom type electrotopological state, extended topochemical atom, chi cluster, weighted path and fingerprints based on substructure, graph, path and extended features including pubchem and Klekota-Roth were found to be useful in prediction. The selected descriptors were then used to develop the QSAR models (Supplementary Table S3). During tenfold cross validation, we achieved maximum Pearson’s correlation coefficient (PCC) of 0.72 in case of HIV; 0.74 in case of HCV; 0.66 in case of HBV; 0.68 in case of HHV and 0.71 in case of rest of the viruses. Also during validation on independent dataset, we achieved a maximum PCC of 0.63, 0.65, 0.61, 0.64 and 0.67 respectively (Table 2). Other statistical parameters used

This article is protected by copyright. All rights reserved.
in the development of QSAR models are depicted in Supplementary Table S2. A scatter plot between actual and predicted efficacy in each case is shown in Figure 2.

In addition, we also checked the performance of our models developed using classification mode of machine learning. During tenfold cross validation, we achieved maximum Matthew's correlation coefficient (MCC) of 0.91 in case of HIV; 0.93 in case of HCV; 0.70 in case of HBV; 0.89 in case of HHV and 0.71 in case of rest of the viruses. Similar performance was shown on the independent datasets (Table 3). The machine learning parameters used for model development are shown in Table 4. Receiver Operating Characteristic (ROC) plots illustrating the performance of the QSAR models are shown in Figure 3.

Web server

The QSAR models have been integrated into a freely available and easy to use web server, 'AVCpred', where users can predict the antiviral potential of their query molecules against the different viruses in terms of percent inhibition value. AVCpred web server includes the following modules:

Submission

This allows users to submit on or more molecules at a time. Users have to choose the viruses on which they want to test their query chemical compounds. On submission, it returns with percent inhibition values against the selected viruses. Also users can view the different properties of the query molecule such as structure, charge, molecular weight, logP value, hydrogen and Lipinski bond donors/acceptors, rigid and rotatable bonds etc. to identify drug-like molecular structures (Figure 4).

This article is protected by copyright. All rights reserved.
Design analogs

It has been found that analogs of known chemical compounds are sometimes more effective than the parent molecule (50). In order to identify potent analogs of an existing AVC we have included the 'Design analogs' tool, where user can design analogs based on given building blocks and predict their inhibition on the viruses.

Draw structure

Using the 'Draw tool' one can sketch the structure of the query molecule using Marvin editor (Figure 5). This tool also gives the predicted percent inhibition values against the different viruses. In addition, one can view the various properties of the query structure.

Search

AVCpred also provides the users a search tool to browse the compounds used in our datasets. In this module, different compounds targeting the viruses are stored in a database. The records can be readily searched, filtered/sorted and downloaded via the web-interface.

Implementation

AVCpred has been developed using the open source LAMP (Linux-Apache-MySQL-PHP) system. The prediction software runs on Red Hat Enterprise Linux 5 environment using Apache httpd server.

Discussion

To inhibit viral growth the antiviral molecules or drugs target different phases of viral life cycle like fusion, integration, replication, maturation etc. and should be relatively non-toxic to the host organism (51, 52). Each stage can be targeted using AVCs that can for example inhibit entry
receptors (CD4, CCR5) or viral enzymes (protease, neuraminidase) etc. (53-55). Various AVCs are currently in medical use and new ones are in clinical trials (56, 57).

Finding new and improved viral inhibitors is a major concern in the treatment of deadly human viruses (58, 59). However, discovery of novel AVCs is a tedious process (60). To speed up the identification of new AVCs, a computational approach using QSAR method is a rational strategy to decrease cost and time efforts in the wet lab (20). QSAR techniques have been widely used in drug designing and further identification of lead molecules (17).

Although there are many QSAR studies pertaining to different types of viral protein inhibitors, however they are very specific in their approach and deal with a particular class of inhibitors such as endonuclease inhibitors (33) in which 40 compounds were used and reached a correlation of 0.76, thiourea derivatives (34) where 85 compounds had a correlation of 0.92, protease inhibitors (39) in which 170 compounds had a correlation of 0.60-0.83 and flavonoid inhibitors (38) where 20 compounds had a correlation of 0.75-0.97 etc. (Table 5). In most of the cases, the studies are carried out on a limited number of inhibitors. Due to this reason they predict the inhibitors that are similar to the compound type with a high correlation, but do not work on other dissimilar inhibitors for the same target virus. To address these limitations, AVCpred models have been developed using diverse and large number of inhibitors. In the current algorithm, we have employed antiviral compound datasets from different studies due to which the overall correlation is less than above studies, yet the models are comparatively more robust to predict different classes of inhibitors. However, as new high-throughput screening data tested under homogeneous conditions on antiviral drugs becomes available, performance of the QSAR method can be improved.
In this study, we developed virus specific as well as general prediction models to identify the likelihood of a compound being antiviral using selected chemical attributes of experimentally validated AVCs. PaDEL, an open source software was used to calculate molecular descriptors and fingerprints. However the software calculates a large number of descriptors and hence we used attribute selection approach to reduce their number by eliminating unrelated and extraneous descriptors to get a highly correlated descriptor set. Our analysis revealed that several chemical descriptors are important in predicting the compound inhibition activity e.g. partial charge, atom type electrotopological state, extended topochemical atom, chi cluster, weighted path, fingerprints etc.

We employed machine learning to train the QSAR models on different sets of experimentally validated data. These models were validated on independent datasets, not used during training and were found to have satisfactory performance. We used the pharmacological data from the ChEMBL resource for training/testing the models developed for general as well as specific viruses. These models were integrated in an open source web server for evaluation and screening of antiviral compounds.

The applicability domain of the QSAR models was demonstrated using Williams plot (Figure 6) in which standardized residuals are plotted against leverages (61). If the standardized residual of a compound is greater than three times standard deviation units (±3σ), the compound is treated as an outlier. The warning value of leverage (h*) is considered as 3p/n, where p is the number of model descriptors plus one and n is the number of training compounds (62, 63). If the leverage of a compound exceeds h*, it is regarded as dissident. The plots demonstrate that the leverages of majority of the compounds do not surpass the critical value (h*) in the regression models and
hence the compounds are within the chemical domain, implying that the predictivity of the models is reliable.

The web server also provides useful services like designing analogs based on given building blocks and drawing structure to sketch novel compounds and predict their inhibition potential against multiple viruses. The AVCpred algorithm is hoped to assist the researchers in discovering novel antiviral compounds as well as virtually check the effect of modifications on existing drugs.

Conclusions

AVCpred is the first web based algorithm for prediction of AVCs based on experimentally validated data sets. Five prediction models pertaining to HIV, HCV, HHV, HBV and a general one were implemented in the web server to make comprehensive predictions. In addition tools for drug design, virtual screening and collection of existing AVCs have also been integrated. This web server would be helpful for researchers working for the development of antiviral therapeutics.

Competing interests

The authors declare that they have no competing interests.
Figure Legends and Tables

Figure 1. Schematic diagram demonstrating workflow of AVCpred.

Figure 2. Scatter plot between actual and predicted percentage inhibition on independent validation datasets of a) HIV, b) HCV, c) HBV, d) HHV, and e) General (26 viruses).

Figure 3. ROC curves depicting performance of QSAR models for a) HIV, b) HCV, c) HBV, d) HHV and e) General (26 viruses).

Figure 4. AVCpred submission form with output

Figure 5. Web interface of 'AVCpred Draw' tool.

Figure 6. Applicability domain plots of the QSAR models for a) HIV, b) HCV, c) HBV, d) HHV and e) General (26 viruses).

Table 1. Creation of datasets for the development of prediction models

| S. No. | Virus Description | Overall data | Data filter | Percent inhibition | Reference | Non redundant |
|--------|-------------------|--------------|-------------|-------------------|-----------|--------------|
| 1      | Human immunodeficiency virus (HIV) | 1383 | 594 | 535 | 389 |
| 2      | Hepatitis C virus (HCV) | 803 | 648 | 618 | 467 |
| 3      | Hepatitis B virus (HBV) | 416 | 284 | 283 | 112 |
| 4      | Human herpesvirus (HHV) | 473 | 312 | 278 | 124 |
| 5      | General (26 viruses) | 5684 | 2662 | 1635 | 1391 |

*Data from ChEMBL was filtered and only compounds with [a] percent inhibition, [b] reference and [c] non redundant SMILES were considered.

# The general dataset is comprised of below viruses with unique number of AVCs in brackets:
- Dengue virus 1 (1), Dengue virus 2 (16), Enterovirus (30), Human adenovirus 5 (41), Human cox B1 (4), Human cox B5 (21), Human echovirus 13 (3), Human echovirus 9 (2), Human enterovirus 71 (19), Human

This article is protected by copyright. All rights reserved.
enterovirus C (1), Human polio virus 1 (4), Human rhinovirus (1), Human rhinovirus 14 (29), Human rhinovirus 1B (18), Human rhinovirus 2 (2), Human T lymphotrophic virus (42), Influenza A (36), Influenza A (H1N1) (16), Influenza B (1), Monkeypox virus (1), Respiratory syncitial virus (4), Rift valley fever virus (cercopithecidae) (1), Sandfly fever sicilian virus (2), SARS Coronavirus (23), Simian virus 40 (45), Sindbis virus (4), Vaccinia virus (12), Vaccinia virus WR (22), Variola virus (1), Vescicular stomatitis virus (63), West Nile Virus (17), Yellow fever virus (51)

Table 2. Pearson correlation values obtained for each viral dataset on their respective QSAR models.

| S. No. | Virus                                           | Antiviral compounds | No. of selected descriptors | Pearson's correlation coefficient (PCC) |
|--------|-------------------------------------------------|---------------------|-----------------------------|----------------------------------------|
| 1      | Human immunodeficiency virus (HIV)              | 389                 | 351 38 45                  | 0.72 0.63                              |
| 2      | Hepatitis C virus (HCV)                         | 467                 | 421 46 52                  | 0.74 0.65                              |
| 3      | Hepatitis B virus (HBV)                         | 112                 | 101 11 15                  | 0.66 0.61                              |
| 4      | Human herpesvirus (HHV)                         | 124                 | 112 12 20                  | 0.68 0.64                              |
| 5      | General (26 viruses)                            | 1391                | 1252 139 65                | 0.71 0.67                              |

[a] Ten-fold cross validation

[b] The general dataset is comprised of below viruses with unique number of AVCs in brackets:
Dengue virus 1 (1), Dengue virus 2 (16), Enterovirus (30), Human adenovirus 5 (41), Human cox B1 (4), Human cox B5 (21), Human echovirus 13 (3), Human echovirus 9 (2), Human enterovirus 71 (19), Human enterovirus C (1), Human polio virus 1 (4), Human rhinovirus (1), Human rhinovirus 14 (29), Human rhinovirus 1B (18), Human rhinovirus 2 (2), Human T lymphotrophic virus (42), Influenza A (36), Influenza A (H1N1) (16), Influenza B (1), Monkeypox virus (1), Respiratory syncitial virus (4), Rift valley fever virus (cercopithecidae) (1), Sandfly fever sicilian virus (2), SARS Coronavirus (23), Simian virus 40 (45), Sindbis virus (4), Vaccinia virus (12), Vaccinia virus WR (22), Variola virus (1), Vescicular stomatitis virus (63), West Nile Virus (17), Yellow fever virus (51)
Table 3. Performance of QSAR models obtained for each viral dataset using classification mode of machine learning.

| S. No. | Virus    | Training/Testing (10-fold) | Validation |
|--------|----------|-----------------------------|------------|
|        |          | Sensitivity | Specificity | Accuracy | MCC | Sensitivity | Specificity | Accuracy | MCC |
| 1      | HIV      | 94.30       | 96.40       | 95.10    | 0.91 | 88.10       | 82.30       | 86.10    | 0.70 |
| 2      | HCV      | 96.90       | 96.40       | 96.60    | 0.93 | 86.61       | 87.20       | 86.80    | 0.73 |
| 3      | HBV      | 87.10       | 81.60       | 85.80    | 0.70 | 87.20       | 80.40       | 84.30    | 0.69 |
| 4      | HHV      | 93.40       | 92.30       | 93.50    | 0.89 | 87.10       | 91.30       | 88.6     | 0.77 |
| 5      | General  | 88.30       | 82.20       | 85.70    | 0.71 | 81.70       | 82.10       | 81.90    | 0.64 |

Table 4. Machine learning parameters selected for the development of the QSAR models.

| S. No. | Model                             | Parameters     |
|--------|-----------------------------------|----------------|
|        |                                   | SMOreg         | SVM*light   |
|        |                                   | Kernel | Optimizer | c | ω | σ | Kernel | g | c |
| 1      | HIV                               | Puk     | RegSMOImproved | 4 | 2 | 3 | RBF    | 0.02 | 200 |
| 2      | HCV                               | Puk     | RegSMOImproved | 5 | 5 | 5 | RBF    | 0.02 | 50  |
| 3      | HBV                               | Puk     | RegSMOImproved | 0.1| 0.3| 0.3 | RBF    | 0.001| 300 |
| 4      | HHV                               | Puk     | RegSMOImproved | 3 | 3 | 3 | RBF    | 0.1  | 100 |
| 5      | General (26 viruses)              | Puk     | RegSMOImproved | 3 | 3 | 5 | RBF    | 0.01 | 50  |

Abbreviations: Puk: Pearson VII function-based universal kernel. RegSMOImproved: Optimizer for algorithm speed improvement. c: regularization constant/complexity parameter allows trade-off between training error and margin. ω: omega exponent value (controls half-width of the peak) σ: sigma bandwidth value (controls tailing factor of the peak). RBF: Radial basis function g: parameter gamma in RBF kernel.
Table 5. Existing QSAR studies pertaining to antiviral compounds.

| S. No. | Compound type                                      | No. of compounds | Correlation | Target virus | Web server /Software | Year | Reference |
|--------|---------------------------------------------------|------------------|-------------|--------------|----------------------|------|-----------|
| 1      | PA endonuclease inhibitors                        | 40               | 0.76        | INFV         | No                   | 2014 | (33)          |
| 2      | Thiourea derivatives                              | 85               | 0.92        | HCV          | No                   | 2013 | (34)          |
| 3      | Integrase inhibitors                              | 77               | 0.98        | HIV          | No                   | 2012 | (35)          |
| 4      | Three different series of HBV inhibitors          | 30               | 0.92        | HBV          | No                   | 2010 | (36)          |
| 5      | HIV-1 entry inhibitors                            | 36               | 0.72        | HIV          | No                   | 2010 | (37)          |
| 6      | Neuraminidase flavonoid inhibitors                | 20               | 0.75-0.97   | H1N1         | No                   | 2010 | (38)          |
| 7      | Protease inhibitors                               | 170              | 0.6-0.83    | HIV          | No                   | 2010 | (39)          |
| 8      | HPV6-E1 helicase ATPase inhibitors                | Full text not available | 0.92 | HPV          | No                   | 2010 | (40)          |
| 9      | Thymidine kinase N2-phenylguanine inhibitors      | 20               | 0.85-0.98   | HSV          | No                   | 2000 | (41)          |

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Table S1 Details of the dataset used for the development of general prediction model.

Table S2 Details of statistical parameters used for the development of the QSAR models.

Table S3 Details of chemical descriptors used in the development of QSAR models.
Acknowledgments

Authors are thankful to Council of Scientific and Industrial Research (CSIR), (GENESIS-BSC0121), Department of Biotechnology (GAP001) and CSIR-Institute of Microbial Technology for providing infrastructure and financial support.

References

1. Prusoff WH, Lin TS, August EM, Wood TG, Marongiu ME (1989) Approaches to antiviral drug development. Yale J Biol Med; 62: 215-25.
2. Moghadamtousi SZ, Kadir HA, Hassandarvish P, Tajik H, Abubakar S, Zandi K (2014) A review on antibacterial, antiviral, and antifungal activity of curcumin. Biomed Res Int; 2014: 186864.
3. Schnitzler P, Schon K, Reichling J (2001) Antiviral activity of Australian tea tree oil and eucalyptus oil against herpes simplex virus in cell culture. Pharmazie; 56: 343-7.
4. Wright K (1986) AIDS therapy. First tentative signs of therapeutic promise. Nature; 323: 283.
5. Jefferson T, Jones M, Doshi P, Spencer EA, Onakpoya I, Heneghan CJ (2014) Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments. BMJ; 348: g2545.
6. Liu S, Wolfe MS, Borchardt RT (1992) Rational approaches to the design of antiviral agents based on S-adenosyl-L-homocysteine hydrolase as a molecular target. Antiviral Res; 19: 247-65.
7. Arakawa T, Yamasaki H, Ikeda K, Ejima D, Naito T, Koyama AH (2009) Antiviral and virucidal activities of natural products. Curr Med Chem; 16: 2485-97.
8. McKinlay MA, Rossmann MG (1989) Rational design of antiviral agents. Annu Rev Pharmacol Toxicol; 29: 111-22.
9. Razonable RR (2011) Antiviral drugs for viruses other than human immunodeficiency virus. Mayo Clin Proc; 86: 1009-26.
10. Bauer DJ (1985) A history of the discovery and clinical application of antiviral drugs. Br Med Bull; 41: 309-14.
11. Hirschman SZ (1971) Approaches to antiviral chemotherapy. Am J Med; 51: 699-703.
12. Capobianchi MR, Giombini E, Rozera G (2013) Next-generation sequencing technology in clinical virology. Clin Microbiol Infect; 19: 15-22.
13. Ru J, Li P, Wang J, Zhou W, Li B, Huang C, et al. (2014) TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminform; 6: 13.
14. Muller B, Krausslich HG (2009) Antiviral strategies. Handb Exp Pharmacol; 1-24.
15. De Clercq E (2002) Strategies in the design of antiviral drugs. Nat Rev Drug Discov; 1: 13-25.
16. Pisurunenkar RR, Khedkar VM, Iyer RP, Coutinho EC (2011) Ensemble QSAR: a QSAR method based on conformational ensembles and metric descriptors. J Comput Chem; 32: 2204-18.
17. Demchuk E, Ruiz P, Chou S, Fowler BA (2011) SAR/QSAR methods in public health practice. Toxicol Appl Pharmacol; 254: 192-7.
18. Ruusmann V, Sild S, Maran U (2014) QSAR DataBank - an approach for the digital organization and archiving of QSAR model information. J Cheminform; 6: 25.
19. Gonzalez-Diaz H, Romaris F, Duardo-Sanchez A, Perez-Montoto LG, Prado-Prado F, Patlewicz G, et al. (2010) Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretical backgrounds, applications, and legal issues. Curr Pharm Des; 16: 2737-64.
20. Cruz-Monteagudo M, Cordeiro MN, Tejera E, Domínguez ER, Borges F (2012) Desirability-based multi-objective QSAR in drug discovery. *Mini Rev Med Chem;12*: 920-35.
21. Prado-Prado FJ, Borges F, Uriarte E, Perez-Montoto LG, Gonzalez-Diaz H (2009) Multi-target spectral moment: QSAR for antiviral drugs vs. different viral species. *Anal Chim Acta;651*: 159-64.
22. Thakur N, Qureshi A, Kumar M (2012) ViRSiRNAdb: a curated database of experimentally validated viral siRNA/shRNA. *Nucleic Acids Res;40*: D230-6.
23. Tyagi A, Ahmed F, Thakur N, Sharma A, Raghava GP, Kumar M (2011) HIVsirDB: a database of HIV inhibiting siRNAs. *PLoS One;6*: e25917.
24. Qureshi A, Thakur N, Kumar M (2013) ViRSiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses. *J Transl Med;11*: 305.
25. Qureshi A, Thakur N, Monga I, Thakur A, Kumar M (2014) ViRmiRNA: a comprehensive resource for experimentally validated viral miRNAs and their targets. *Database (Oxford);2014*: 
26. Qureshi A, Thakur N, Tandon H, Kumar M (2014) AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res;42*: D1147-53.
27. Qureshi A, Thakur N, Kumar M (2013) HIPdb: a database of experimentally validated HIV inhibiting peptides. *PLoS One;8*: e54908.
28. Thakur N, Qureshi A, Kumar M (2012) AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res;40*: W199-204.
29. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, et al. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res;42*: D1083-90.
30. Wang Y, Suzek T, Zhang J, Wang J, He S, Cheng T, et al. (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res;42*: D1075-82.
31. Irwin JJ, Shoichet BK (2005) ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model;45*: 177-82.
32. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res;36*: D901-6.
33. Yan Z, Zhang L, Fu H, Wang Z, Lin J (2014) Design of the influenza virus inhibitors targeting the PA endonuclease using 3D-QSAR modeling, side-chain hopping, and docking. *Bioorg Med Chem Lett;24*: 539-47.
34. Sharma A, Gupta SP, Siddiqui AA (2013) A QSAR study on a series of thiourea derivatives acting as anti-hepatitis C virus agents. *Indian J Biochem Biophys;50*: 278-83.
35. Sun XH, Guan JQ, Tan JJ, Liu C, Wang CX (2012) 3D-QSAR studies of quinoline ring derivatives as HIV-1 integrase inhibitors. *SAR QSAR Environ Res;23*: 683-703.
36. Arora PK, Patil VM, Gupta SP (2010) A QSAR study on some series of anti-hepatitis B virus (HBV) agents. *Bioinformation;4*: 417-20.
37. Pirhadi S, Ghasemi JB (2010) 3D-QSAR analysis of human immunodeficiency virus entry-1 inhibitors by CoMFA and CoMSIA. *Eur J Med Chem;45*: 4897-903.
38. Mercader AG, Pomilio AB (2010) QSAR study of flavonoids and biflavonoids as influenza H1N1 virus neuraminidase inhibitors. *Eur J Med Chem;45*: 1724-30.
39. Basak SC, Mills D, Garg R, Bhattachari B (2010) A QSAR study of HIV protease inhibitors using theoretical descriptors. *Curr Comput Aided Drug Des;6*: 269-82.
40. Pillai GG, Sikk L, Tamm T, Karelson M, Burk P, Tamm K (2014) Theoretical modeling of HIV: QSAR and novodesign with fragment approach. *Curr Comput Aided Drug Des;10*: 303-14.
41. Gaudio AC, Richards WG, Takahata Y (2000) QSAR and molecular graphics analysis of N2-phenylguanines as inhibitors of herpes simplex virus thymidine kinases. *J Mol Graph Model;18*: 33-41.
42. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem;32*: 1466-74.

This article is protected by copyright. All rights reserved.
43. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics; 20: 2479-81.
44. Kumar R, Chaudhary K, Singla D, Gautam A, Raghava GP (2014) Designing of promiscuous inhibitors against pancreatic cancer cell lines. Sci Rep; 4: 4668.
45. Chauhan JS, Dhanda SK, Singla D, Agarwal SM, Raghava GP (2014) QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. PLoS One; 9: e101079.
46. Li C, Jiang L. (2006) Using locally weighted learning to improve SMOreg for regression. Using locally weighted learning to improve SMOreg for regression. Springer; p. 375-84.
47. Üstün B, Melssen WJ, Buydens LMC (2006) Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel. Chemometrics and Intelligent Laboratory Systems; 81: 29-40.
48. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KK (2000) Improvements to the SMO algorithm for SVM regression. IEEE Trans Neural Netw; 11: 1188-93.
49. Joachims T (2002) Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers.
50. Kumar A, Khan IA, Koul S, Koul JL, Taneja SC, Ali I, et al. (2008) Novel structural analogues of piperine as inhibitors of the NorA efflux pump of Staphylococcus aureus. J Antimicrob Chemother; 61: 1270-6.
51. Littler E, Oberg B (2005) Achievements and challenges in antiviral drug discovery. Antivir Chem Chemother; 16: 155-68.
52. Reusser P (2000) Antiviral therapy: current options and challenges. Schweiz Med Wochenschr; 130: 101-12.
53. De Clercq E (2009) Anti-HIV drugs: 25 compounds approved within 25 years after the discovery of HIV. Int J Antimicrob Agents; 33: 307-20.
54. Arts EJ, Hazuda DJ (2012) HIV-1 antiretroviral drug therapy. Cold Spring Harb Perspect Med; 2: a007161.
55. Barik S (2012) New treatments for influenza. BMC Med; 10: 104.
56. De Clercq E (2004) Antiviral drugs in current clinical use. J Clin Virol; 30: 115-33.
57. De Clercq E (2008) Emerging antiviral drugs. Expert Opin Emerg Drugs; 13: 393-416.
58. Martinez JP, Sasse F, Bronstrup M, Diez J, Meyerhans A (2015) Antiviral drug discovery: broad-spectrum drugs from nature. Nat Prod Rep; 32: 29-48.
59. Niu B, Lu L, Liu L, Gu TH, Feng KY, Lu WC, et al. (2009) HIV-1 protease cleavage site prediction based on amino acid property. J Comput Chem; 30: 33-9.
60. Minovski N, Zuperl S, Dragan V, Novic M (2013) Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum euclidean distance space analysis: a case study. Anal Chim Acta; 759: 28-42.
61. Fatemi MH, Heidari A, Gharaghan S (2015) QSAR prediction of HIV-1 protease inhibitory activities using docking derived molecular descriptors. J Theor Biol; 369: 13-22.
62. Zhao Q, Yang K, Li W, Xing B (2014) Concentration-dependent polyparameter linear free energy relationships to predict organic compound sorption on carbon nanotubes. Sci Rep; 4: 3888.
