RSC Advances

ROYAL SOCIETY OF CHEMISTRY

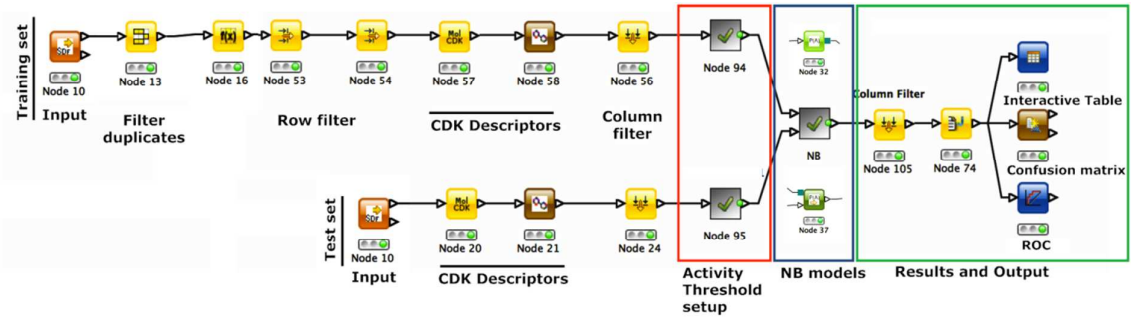
www.rsc.org/advances

Graphical Abstract

The KNIME Based Classification Models for Yellow Fever Virus Inhibition

N.S. Hari Narayana Moorthy [§*] and Vasanthanathan Poongavanam[¶]

The Naïve Bayes method as implemented in KNIME platform was used for the classification analysis of yellow fever inhibitors obtained from ChEMBL database. The best classification model is able to discriminate > 90% of inhibitors from non-inhibitors with an overall accuracy of >90%.



The KNIME Based Classification Models for Yellow Fever Virus Inhibition

N.S. Hari Narayana Moorthy [§*] and Vasanthanathan Poongavanam[¶]

[§]Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, s/n, Rua do Campo Alegre, 4169-007 Porto, Portugal.

[¶]Department of Physics, Chemistry, Pharmacy, University of Southern Denmark, Campusvej 55, DK-5230, Odense M, Denmark.

*For correspondence

Email: hari.nmoorthy@gmail.com, hari.moorthy@fc.up.pt

Abstract:

Yellow fever is one of the virus-infected diseases spreading through mosquitos and kills more than thirty thousand people every year. Although there are a large number of compounds have been reported, none of the drugs have yet been approved for the clinical use. In the process of drug development against yellow fever virus (YFV), in the present investigation, we have developed efficient classification models based on a large dataset (309 compounds) compiled from the ChEMBL database. The Naïve Bayes method as implemented in KNIME platform was used for the classification analysis. The best models obtained using the combined dataset showing accuracy of >90% on the test set prediction (Matthew's correlation coefficients of >0.7). All the models developed in this study could be applicable for virtual screening of yellow fever virus inhibition.

Keywords: KNIME, classification, yellow fever virus, QSAR, virtual screening, ChEMBL.

Introduction

The yellow fever virus (YFV) is a member of Flaviviridae family and this family contains hepatitis C virus (HCV), Dengue virus (DENV), West Nile virus (WNV), Japanese encephalitis virus (JEV), Tick-borne encephalitis virus (TBEV) and bovine viral diarrhoea virus (BVDV)[1,2]. These viruses are classified into three genera: flavivirus, hepacivirus, and pestivirus. YFV is one of the mosquito-borne flavivirus causes the acute viral infection called yellow fever (YF). Flaviviruses are small, enveloped RNA viruses responsible for the above mentioned clinical diseases in humans. These viruses share similar genomic organization and replication strategy and those are transmitted by arthropods and mosquitos to humans and birds[2-4]. YFV majorly affects the tropical areas of South America and Africa and this virus causes nearly 200000 new infections and 30000 deaths every year[5,6]. Since 1980, number of cases for YFV infection has been increased, due to the frequent migration of people, less immune, most people living in cities and climate changes. According to WHO, severe cases of this infection may cause fatality which is more than 50%. Eventhough the YFV rate has been increased over the last 10 years, due to the reasons stated above[1,7]. YFV is an enveloped virus with polyprotein of over 350 kDa, encoded by a single and positive stranded RNA genome. The non-structural NS3 serine protease (non-structural protein part in flavivirade family) present in the virus is essentially acts for viral replication (maturation of the viral polyprotein) and an attractive target for antiviral drug discovery[1,2,8,9].

Presently, only common antiviral drugs are being used for the treatment of YFV infection and no specific chemotherapeutic agents are available for any of these flaviviral infections

including yellow fever. Still, there is single drug has not yet been approved for effective against YFV treatment; however anti-YFV vaccine (17D) is available to prevent this disease. However, this live-attenuated 17D vaccine has shown to cause wild-type disease and systemic infections in a subset of patients[1,10]. It reveals that the antiviral chemotherapeutics would be inexpensive, stable, safe and would have efficient when administered before and after virus infection and can be broadly active against a range of viruses[1].

In order to discover novel molecules, virtual screening of large database or knowledge based drug design like computational methods are appropriate. From extensive literature analysis on this target, revealed that there are only a limited number of *in silico* studies have been reported on flaviviridae family viruses (including YFV). A computational based screening analysis on NCI library has been reported to identify novel flaviviral inhibitors using *n*-octyl-β-D-glucoside (β-OG) binding pocket of dengue E protein. The β-OG pocket is an ideal target for structure based design of potential antiviral agents, because the ligand complex could change the conformational equilibrium associated with the hinge angle (interferes with the fusion of the viral envelope with the host cell membrane) and inhibit virus maturation. They reported three compounds as significant hit for YFV inhibitory activity through structure based virtual screening and cell based assay methods. On account of these templates, a series of molecules were constructed by them as YFV inhibitors through structural modification[11-13]. The computational studies such as molecular docking and 3D-comparative molecular similarity analysis-quantitative structure activity analysis (3D-CoMSIA-QSAR) were applied on the ChEMBL database to identify significant N-substituted indole based HCV replication inhibitors. The relative field contributions of 27.6%, 42.1% and 30.3% for steric, hydrophobic and H-bond acceptor fields, respectively were applied to the CoMSIA analysis. The QSAR model validation statistics such as R^2_{test} and R^2_m exhibited a value of 0.727 and 0.635 respectively. The docking studies of the molecules on HCV revealed that the indole moiety in the active compounds oriented to the binding site responsible for group of water molecules located (inactive compounds have different orientation). Interestingly, the active and the inactive compounds reported to possess equal docking scores[14]. The molecular modelling (homology modelling and docking) study was reported on the development of fusion inhibitors on ectodomain of TBEV E protein in virus. The β-OG pocket of the homology model (open state) was used for the virtual screening, which identified 89 compounds as hit from substituted 1,4-dihydropyridines and pyrido[2,1-*b*][1,3,5]thiadiazines containing data set. Experimental results on the identified hits showed that 17 compounds had significant inhibition against different viruses (TBEV, Powassan

virus, or Omsk haemorrhagic fever virus)[15]. Docking and pharmacophore studies were reported for some flavivirus inhibitors by Tonelli *et al.* against BVD virus. The pharmacophore results showed that 98% chance for the best pharmacophore hypothesis to represent a true correlation in the training set activity. Docking and multiple alignments of RNA virus proteins showed that the active compounds target effectively the BVDV RNA-dependent RNA-polymerase (RdRp), which shares some structural similarity with HCV RdRp[16-18]. On account of the above statements, in the present investigation, we have used a set of literature compounds that inhibit the YFV to develop the classification models using the Naïve Bayes method as implemented in the KNIME platform[19-21].

Computational Methods and Materials

A data set of 379 YFV inhibitors was retrieved from the ChEMBL database (https://www.ebi.ac.uk/chembl/) (composed mainly from six journal literatures)[11,16-18,22,23]. Each dataset has different parent structures, which are provided in **Figure 1**. Before the datasets used for the classification study, each dataset was manually checked and curated, which includes removal of salts, generation of 3D structures, energy minimization using OPLS2005 force field. Subsequently, thirty 2D physicochemical descriptors of the compounds were calculated using the CDK tool as it is implemented in the KNIME an open source data analyzer and integrator[19-21]. Classification models were developed using the Weka data mining software[24] as implemented in KNIME. The Weka provides a large collection of supervised and unsupervised machine learning algorithms, attribute selection and visualization methods[24,25]. The dataset was characterized using the SIMCA-P software (Version 10.5. Umetrics, Umea, Sweden).

Naïve Bayesian theory

Naïve Bayesian classification is a probabilistic supervised learning method utilizes the Bayes theorem to calculate how the degree of belief in a proposition changes in accordance to evidence. Briefly, the Bayesian learning works as follows: before any data has been observed, the expectation as to what the true relationship between those data can be expressed in a probability distribution over the assumptions that define this relationship. For example, a fruit may be considered to be an apple if it is red, round and about 3" in diameter. A Naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features[21,26,27].

The probability model (a conditional model) for a Naïve Bayesian classifier is

$$p(C \mid F_{1,}.....,F_n) \quad \ldots (1)$$

C, a dependent class variable with a small number of outcomes or classes, conditional on several feature variables F_1 through F_n. The conditional distribution over the class variable *C* under the independence assumption is:

$$p(C \mid F_{1,}....,F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i \mid C) \quad \ldots (2)$$

Where the evidence $Z = p(F_1, ..., F_n)$ is a scaling factor dependent only on $F_1, ..., F_n$, that is, a constant if the values of the feature variables are known.

The Gaussian Naïve Bayes classifier is applied for a class of continuous data which are distributed according to a Gaussian distribution. When the training data contains a continuous attributes, *x* and μ_c, the mean of the values in *x* associated with class *c,* and let σ_c^2 be the variance of the values in *x* associated with class *c*. The probability density of some value given a class, $P(x = v/c)$, can be computed by plugging *v* into the equation for a normal distribution parameterized by μ_c, and σ_c^2. That is,

$$p(x = v \mid c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}} \quad \ldots (3)$$

Classification model development

In total 16 classification models were developed, models from 1 to 12 represent individual datasets and models 13-16 were developed from the combined dataset (309 compounds). The workflows were constructed with KNIME platform containing the CDK nodes. The workflow charts used for the analysis are provided in supplementary material (Figure S1).

Before the models development, the whole dataset was divided into a training (65%) and a test set (35%) according to the stratified sampling method, which divides the inhibitors and non-inhibitors equally in the test and training sets. The inhibitors and the non-inhibitors were defined according to different activity thresholds (e.g. IC_{50} <=10 μM is inhibitors; IC_{50} >10 μM is non-inhibitors). In this study, we explored the quality of models from different activity thresholds, e.g. 10, 30, 50 and 100 μM. However, models developed from the activity thresholds 10 and 100 μM were not discussed due to unbalanced distribution of inhibitors and non-inhibitors in the datasets, which leads to insignificant predictions (results are provided in

the supplementary materials (Table S1)). Therefore, models from 30 and 50 μM are presented and discussed.

Assessment of classification models

Confusion matrix from each classification model was used to calculate various statistical parameters to assess the quality of models. Statistical parameters used in this study are sensitivity (true positive rate), specificity (true negative rate), G-mean, Matthew's correlation coefficient (MCC) and overall accuracy.

$$Sensitivity = \frac{TP}{TP + FN} \quad \ldots (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad \ldots (5)$$

$$G - mean = \sqrt{Sensitivity \times Specificity} \quad \ldots (6)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \ldots (7)$$

$$Accuracy = \frac{(TP + TN)}{(T + N)} \quad \ldots (8)$$

$$F - measure = \frac{\Pr ecision \times Sensitivity}{\Pr ecision + Sensitivity} \quad \ldots (9)$$

Here, TP, TN, FP and FN denote true positive, true negative, false positive and false negative respectively. The accuracy is the proportion of correctly predicted positives and negatives. The F-measure is a measure of a test's accuracy. Sensitivity and specificity measure the proportion of actual positives and negatives which are correctly identified, respectively. The geometric mean (G-mean) evaluates the degree of inductive bias in terms of a ratio of positive accuracy and negative accuracy and this term is used to check how well the model is able to predict two classes. Matthew's correlation coefficient (MCC) indicates the degree of the correlation between the actual and predicted classes. It ranges from -1 to +1 and is generally regarded as a good measure of the quality of the binary classification[28-30].

Results and Discussion

Characterization of dataset

A principal component analysis (PCA) was performed to check possible presence of clusters, outliers, similarities or dissimilarities, distribution of inhibitors and non-inhibitors in the training and test set in the physicochemical space. First two principal components from 24 CDK 2D-descriptors explain 79% of variance in the data set. The score plot from PCA shows (**Figure 2a**) that the diversity of dataset is satisfactorily reflected in the training set and there are no distinct clusters in the dataset. There are some distinct outliers were observed (distance to model plot is provided in **Figure 2b**), however there is no structural similarity within this class of compounds. It was observed from the loading plot (**Supplementary Figure S2**) that most of the inhibitors are highly influenced by the topological polar surface area (TPSA) and polar bonds. This reveals that non-inhibitors are relatively more hydrophobic than inhibitors.

Construction of classification models

Naïve Bayes based classification models for YFV inhibition was developed using a set of 24 descriptors. Descriptors used for the model development were selected from the BestFirst attribute selection method provided in the Weka software. The quality of the models for each dataset was compared in terms of MCC of test set. Overall KNIME based workflow is provided in **Figure 3**. In general, all 6 datasets (activity threshold 30 μM) perform equally well and an overall accuracy of the test set is > 75%.

The quality of model was better for all the dataset in terms of MCC which found to be larger than 0.6, except for dataset 2, which performs very poor (-0.11). This poor performance was due to the fact that there were only two non-inhibitors, which were predicted as inhibitors because of not only sharing of common scaffolds as positive but also shares similar structural patterns. This is exemplified in the **Figure 4**. In addition, we developed models with activity threshold of 50 μM and this leads to overall improvements in the quality of the model in all datasets. Summary of the model quality is provided in the **Table 1.** Models derived from all the datasets are statistically significant (MCC and G-mean values >0.7). In the same way, the F-score or F-measure also describes the significance of the analysed data set. It gives the values >0.8 for all the data set explain that the descriptors used in the models classified the data set significantly. The other statistical parameters such as sensitivity and specificity showed the values >0.75, except for the model 3 and 7 (provided the values little less than 0.75).

Models from the combined dataset

Although the models obtained from different datasets were significantly predictive, future prediction might be insignificant due to low diversity as each dataset contains unique series of compounds. Thus, it would be interesting to see the prediction ability by combining all the datasets into one. Out of 379 compounds, 70 compounds (mainly Krecmerová M *et al.* [22]) was used as a test set and the remaining datasets were combined into one in order to train the model. Overall, the activity threshold 50 μM has yielded a relatively good accuracy. The models developed from other activity thresholds (10 and 30 μM) were efficient to predict inhibitors (>90%) compared to non-inhibitors (~65%). However, most of the models were suffered from an imbalanced class distribution which was reflected in the quality of the model. It is highly important to have models that are able to predict correctly both classes in a reasonably balanced manner and not only correctly predicts one of the classes with high accuracy.

Therefore, new models were constructed based on a set of 309 compounds (Krecmerová *et al* dataset was excluded), in which 200 compounds were used as a training set and the remaining 109 compounds were used as a test set. The models for each activity thresholds (10, 30, 50 and 100 μM) were developed. These models provided the sensitivity and specificity values of >0.9, however the specificity values have significantly decreased for the models 13 and 14. Overall, the model 15 from activity threshold at 50 μM gives the better performance over the other thresholds. The model (15) predicts 92% of inhibitors and 78% of non-inhibitors correctly with good coefficient (MCC =0.71), and the quality is also reflected in the high G-mean score (0.84) (**Table 2**).

Open source KNIME workflow for YFV

In order to provide the YFV inhibition model to the medicinal chemistry community, model 15 is implemented into KNIME workflow as shown in **Figure 5**. The workflow can be directly downloaded into any workstation having KNIME software package installed. There is no prerequisite before running KNIME, as most of the cheminformatics nodes are already available in the KNIME suite. The KNIME workflow reads the molecules (2D or 3D conformation) in sdf file format (.sdf) and other procedures are automated. Subsequently, it produces output files (xls or csv) containing predictions with statistical significance.

Conclusion

The number of cases of YFV infection has significantly been increased in the recent years; although the vaccines are available for YFV infection, an inexpensive, safe and effective

chemotherapeutic agent is still needed for wide usability. In the present study, the KNIME based classification models were developed using the existing YFV inhibitors from the ChEMBL database. The best classification model is able to discriminate >90% of inhibitors from non-inhibitors with an overall accuracy of >90%. Subsequently, the best model is implemented in the KNIME workflow which could be used as a virtual screening workflow to screen novel molecules for the YFV inhibitory activity.

Acknowledgement

The authors are thankful to ChemAxon, KNIME and OpenEye scientific software for providing a free academic license.

References

1. J. G. Julander, *Antiviral Res.,* 2013, **97,** 169–179.
2. S. Agnihotri, R. Narula, K. Joshi, S. Rana, M. Singh, *Bioinformation* 2012, **8(3),** 123–127.
3. K. Lohr, J. E. Knox, W. Y. Phong, N. L. Ma, Z. Yin, A. Sampath, S. J. Patel, W. L. Wang, W.L., Chan, K. R. Ranga Rao, G. Wang, S. G. Vasudevan, T. H. Keller, S. P. Lim, *J. Gen. Virol.,* 2007, **88,** 2223–2227.
4. C. C. Pacca, A. A. Severino, A. Mondini, P. Rahal, S. G. P. Davila, J. A. Cordeiro, M. C. L. Nogueira, R. V. M. Bronzoni, M. L. Nogueira, *Virus Genes,* 2009, **38,** 224–231.
5. J. Lescar, D. Luo, T. Xu, A. Sampath, S. P. Lim, B. Canard, S. G. Vasudevan, *Antiviral Res.*, 2008, **80,** 94–101.
6. G. Chatelain, Y. Debing, T. D. Burghgraeve, J. Zmurko, M. Saudi, J. Rozenski, J. Neyts, A. V. Aerschot, *Eur. J. Med. Chem.,* 2013, **65,** 249–255.
7. Global alert and response (GAR)-Yellow fever. World Health Organization. http://www.who.int/csr/disease/yellowfev/en/
8. T. J. Chambers, A. D. Droll, Y. Tang, Y. Liang, V. K. Ganesh, K. H. M. Murthy, M. Nickells, *J. Gen. Virol.,* 2005, **86,** 1403–1413.
9. D. A. Droll, H. M. K. Murthy, T. J. Chambers, *Virol.*, 2000, **275,** 335–347.
10. M. G. V. Santana, P. C. C. Neves, J. R. Santos, N. S. Lima, A. A. C. dos Santos, D. I. Watkins, R. Galler, M. C. Bonaldo, *Virol.,* 2014, **452-453,** 202–211.
11. Z. Li, M. Khaliq, Z. Zhou, C. B. Post, R. J. Kuhn, M. Cushman, *J. Med. Chem.,* 2008, **51,** 4660–4671.

12. Y. Modis, S. Ogata, D. Clements, S. C. A. Harrison, *Proc. Natl. Acad. Sci. U.S.A.,* 2003, **100,** 6986–6991.
13. Y. Zhang, W. Zhang, S. Ogata, D. Clements, J. H. Strauss, T. S. Baker, R. J. Kuhn, M. G. Rossmann, *Structure,* 2004, **12,** 1607–1618.
14. E. Vrontaki, G. Melagraki, T. Mavromoustakos, A. Afantitis, *Methods,* 2015, **71,** 4-13.
15. D. I. Osolodkin, L. I. Kozlovskaya, G. G. Karganova, E. V. Dueva, V. A. Palyulin, N. S. Zefirov, V. M. Pentkovski, *J. Cheminf.* 2012, **4(Suppl 1),** P29.
16. M. Tonelli, V. Boido, C. Canu, A. Sparatore, F. Sparatore, M. S. Paneni, M. Fermeglia, S. Pricl, L. Colla, L. Casula, C. Ibba, D. Collu, R. Loddo, *Bioorg. Med. Chem.,* 2008, **16,** 8447–8465.
17. M. Tonelli, I. Vazzana, B. Tasso, V. Boido, F. Sparatore, M. Fermeglia, M. S. Paneni, P. Posocco, S. Pricl, P. L. Colla, C. Ibba, B. Secci, G. Collu, R. Loddo, *Bioorg. Med. Chem.,* 2009, **17,** 4425–4440.
18. M. Tonelli, M. Simone, B. Tasso, F. Novelli, V. Boido, F. Sparatore, G. Paglietti, S. Pricl, G. Giliberti, S. Blois, C. Ibba, G. Sanna, R. Loddo, P. L. Colla, *Bioorg. Med. Chem.,* 2010, **18,** 2937–2953.
19. S. Beisken, T. Meinl, B. Wiswedel, L. F. de Figueiredo, M. Berthold, C. Steinbeck, *BMC Bioinformatics,* 2013, **14,** 257.
20. M. R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel, KNIME: The Konstanz Information Miner. In. *Studies in classification, data analysis, and knowledge organization (GfKL 2007)*. Heidelberg-Berlin, Springer-Verlag, 2007.
21. M. R. Berthold, N. Cebron, F. Dill, G. Di Fatta, T. R. Gabriel, F. Georg, T. Meinl, P Ohl, C. Sieb, B. Wiswedel, KNIME-The Konstanz Information Miner ACM SIGKDD Explorations Newsletter. New York, USA: ACM, 2009, pp. 31.
22. M. Krecmerova, A. Holy, A. Pıskala, M. Masojıdkova, G. Andrei, L. Naesens, J. Neyts, J. Balzarini, E. De Clercq, R. Snoeck, *J. Med. Chem.,* 2007, **50,** 1069–1077.
23. M. Mazzei, E. Nieddu, M. Miele, A. Balbi, M. Ferrone, M. Fermeglia, M. T. Mazzei, S. Pricl, P. L. Colla, F. Marongiu, C. Ibbac, R. Loddo, *Bioorg. Med. Chem.,* 2008, **16,** 2591–2605.
24. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H., Witten, The WEKA data mining software: an update, SIGKDD Explorations Newsletter 2009, 11, 10–18.

25. G. Melagraki, A. Afantitis, *Chemom. Intell. Lab. Sys.*, 2013, **123,** 9–14.
26. A. McCallum, K. A. Nigam, A comparison of event models for Naive Bayes text classification. AAAI-98 workshop on learning for text categorization, Madison, Wisconsin, USA. 26-27 July 1998, 752.
27. K. Chai, H. T. Hn, H. L. Chieu, H.L. Bayesian online classifiers for text classification and filtering. Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval, August 2002, pp 97-104.
28. N. S. H. N. Moorthy, S. F. Sousa, M. J. Ramos, P. A. Fernandes, *RSC Adv.,* 2014, **4(106),** 61624–61630**.**
29. N. S. H. N. Moorthy, N. M. F. S. A. Cerquira, M. J. Ramos, P. A. Fernandes, *Chemom. Intellegent Lab. Sys.,* 2015, **140,** 102-116.
30. K. K. Chohan, S. W. Paine, J. Mistry, P. Barton, A. M. Davis, *J. Med. Chem.,* 2005, **48,** 5154–5161.

Figure 1: Parent structure of the compounds considered for the present investigation

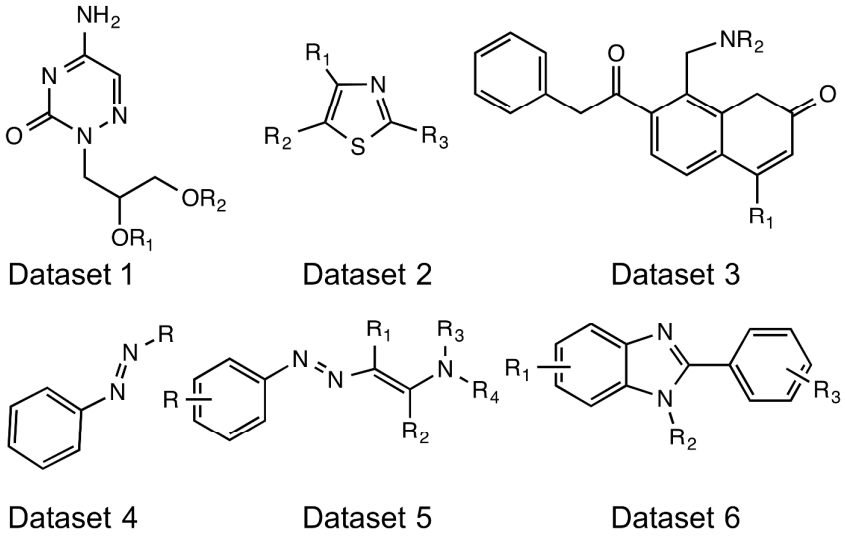


Figure 2: a) Principal component analysis from first 2 PCs, (b) Distance to model plot. The compounds are colored as follows: training set and non-inhibitors as green dots, training set and inhibitors as blue dots, test set and non-inhibitor as brown diamonds, and test set and inhibitor as red circles.

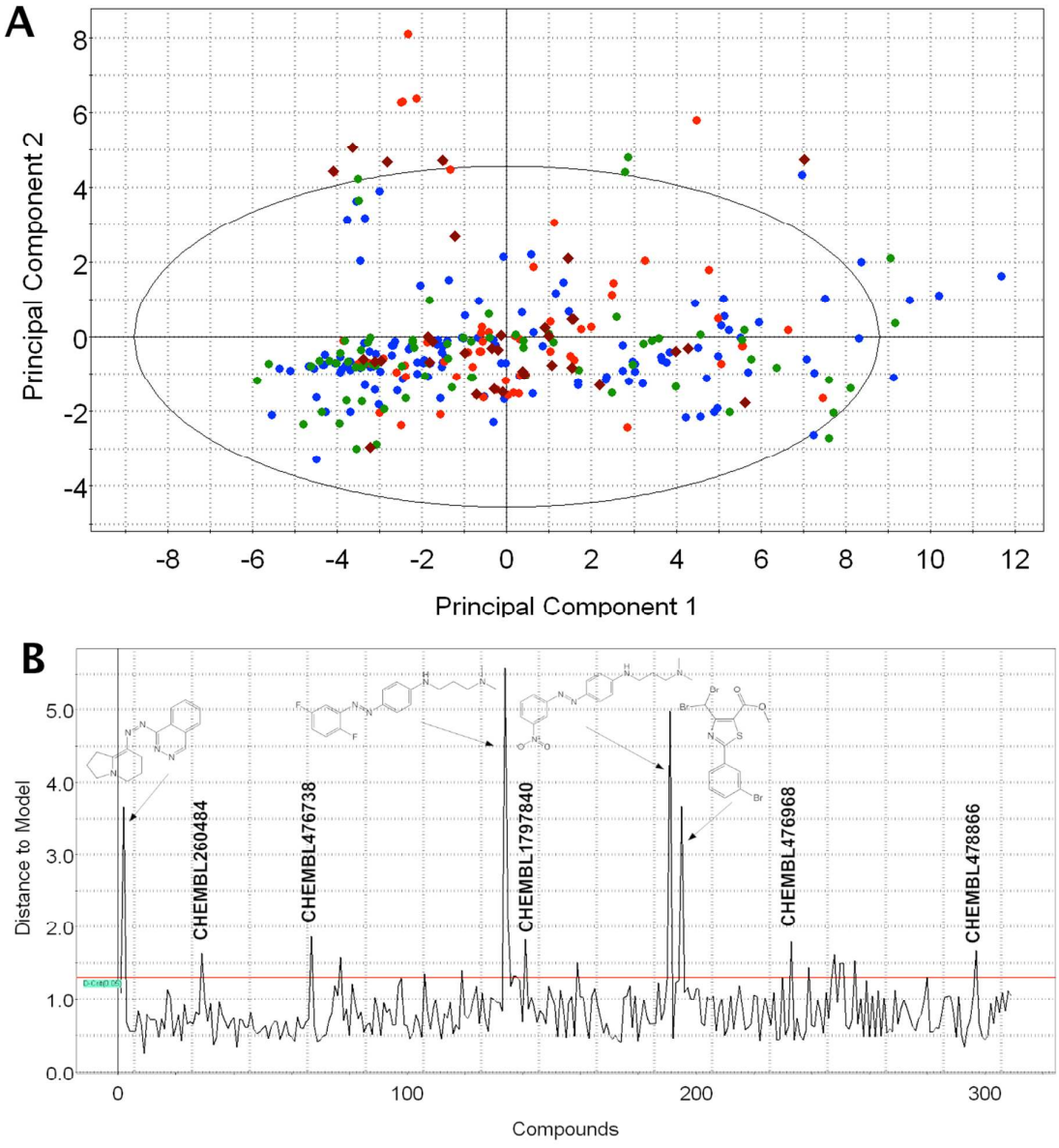


Figure 3: Overall classification workflow is shown and various task nodes are highlighted, including the results output.

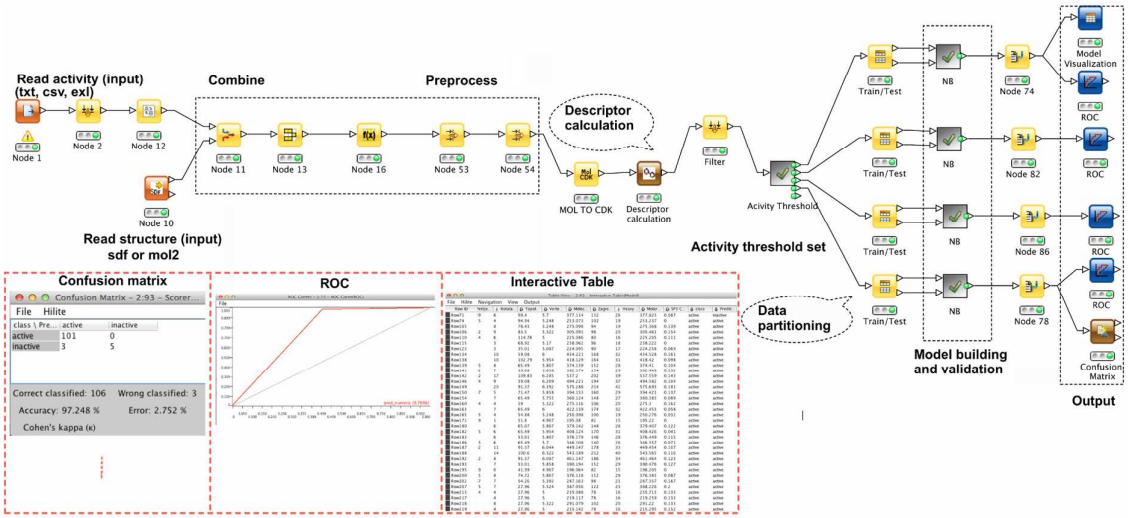


Figure 4: Examples of misclassified compounds in the test set.

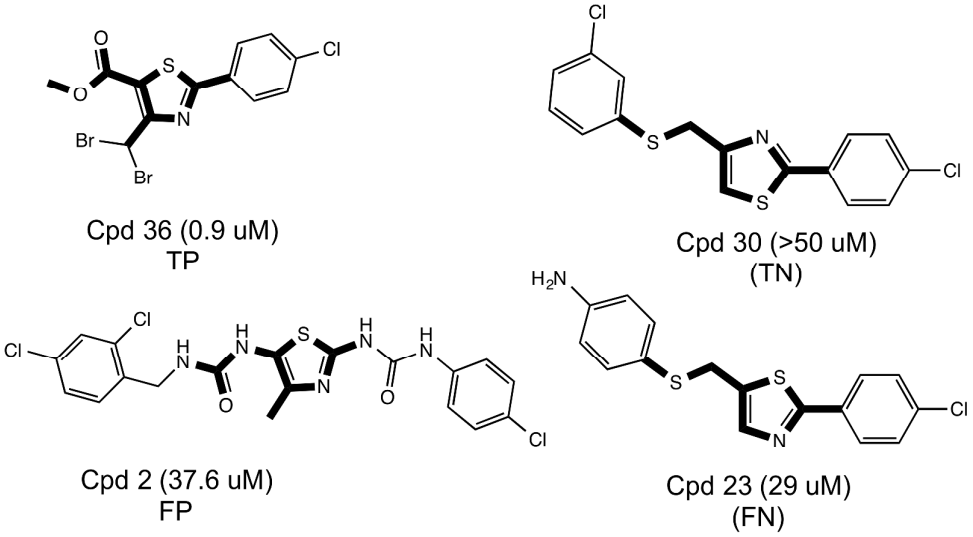


Figure 5: Final KNIME workflow for the classification model of yellow fever inhibition

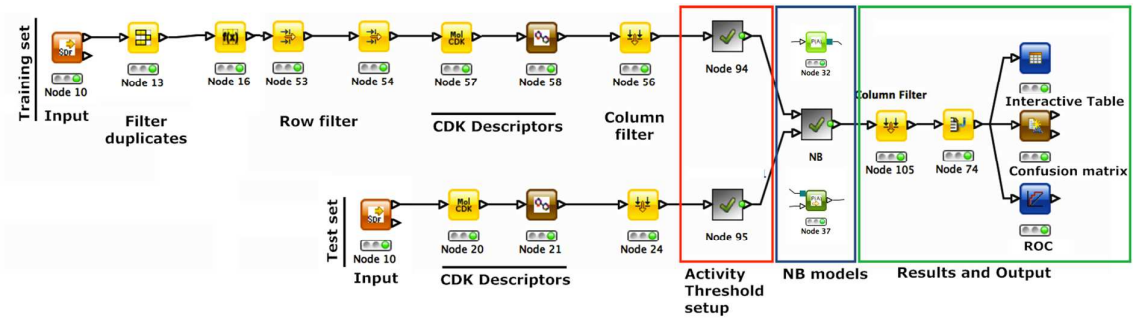


Table 1: Summary of models for individual dataset (Test set prediction)

Dataset	Models	ActThrd	Confusion Matrix				Sens.	Spec.	ROC	G-mean	F-measure	MCC	Accu
			TP	FN	TN	FP							
1	1	30	16	2	6	1	0.89	0.86	0.86	0.87	0.91	0.72	0.88
	2	50	14	1	10	0	0.93	1.00	1.00	0.97	0.97	0.92	0.96
2	3	30	11	1	0	2	0.92	0.00	0.91	0.00	0.88	-0.11	0.79
	4	50	8	1	5	0	0.89	1.00	0.88	0.94	0.94	0.86	0.93
3	5	30	22	0	1	0	1.00	1.00	0.52	1.00	1.00	1.00	1.00
	6	50	19	2	1	1	0.90	0.50	0.92	0.67	0.93	0.34	0.87
4	7	30	19	1	4	1	0.95	0.80	0.52	0.87	0.95	0.75	0.92
	8	50	10	5	10	0	0.67	1.00	0.92	0.82	0.80	0.67	0.80
5	9	30	18	1	4	0	0.95	1.00	0.88	0.97	0.97	0.87	0.96
	10	50	16	1	6	0	0.94	1.00	0.83	0.97	0.97	0.90	0.96
6	11	30	16	2	7	0	0.89	1.00	0.91	0.94	0.94	0.83	0.92
	12	50	15	0	6	4	1.00	0.60	0.88	0.77	0.88	0.69	0.84

ActThrd = Activity threshold; TP = True positive; FN = False negative; TN = True negative; FP = False positive; ROC = Receiver operating curve; Sens. = Sensitivity; Spec. = Specificity; MCC = Matthew's correlation coefficient; Accu. = Accuracy; F = F-measure.

Table 2: Statistical parameters for the combined data set models

Models	Confusion Matrix				Sensitivity	Specificity	ROC	G-mean	F-measure	MCC	Accuracy
	TP	FN	TN	FP							
13[a]	101	0	5	3	1.00	0.63	0.81	0.79	0.97	0.78	0.99
14[b]	85	2	12	10	0.98	0.55	0.76	0.73	0.89	0.63	0.93
15[c]	67	6	28	8	0.92	0.78	0.85	0.84	0.87	0.71	0.91
16[d]	47	1	61	0	0.98	1.00	0.99	0.99	0.99	0.98	0.99

TP = True Positive; FN = False Negative; TN = True Negative; FP = False Positive; ROC = Receiver operating curve; MCC = Matthew's Correlation Coefficient; F = F-measure. [a] model from activity threshold at 10 μM, [b] model from activity threshold at 30 μM, [c] model from activity threshold at 50 μM, [d] model from activity threshold at 100 μM.