In-silico Predictive Mutagenicity Model Generation Using Supervised Learning Approaches

Anurag Passi[#], Abhik Seal[$], OSDD Consortium[*], UC Jaleel[+]

Council of Scientific and Industrial Research (CSIR), India. Email: anuragpassibioinfo@gmail.com
$ Doeacc Society Kolkata (JU Campus Kolkata-700032) West Bengal, India. Emal: abhik1368@gmail.com
* Open Source Drug Discovery, CSIR, India. Email: info@osdd.net
+ Malabar Christian College, Calicut, Kerala, India. Email: jaleel.uc@gmail.com

AIM

To build in-silico predictive mutagenicity model using data mining and machine learning approaches

BACKGROUND

With the advent of High Throughput Screening techniques, it is feasible to filter possible leads from a mammoth chemical space that can act against a particular target and inhibit its action. Virtual screening complements the in-vitro assays which are costly and time consuming. This process is used to sort biologically active molecules by utilizing the structural and chemical information of the compounds and the target proteins in order to screen potential hits. Various data mining and machine learning tools utilize Molecular Descriptors through the knowledge discovery process using classifier algorithms that classify the potentially active hits for the drug development process.

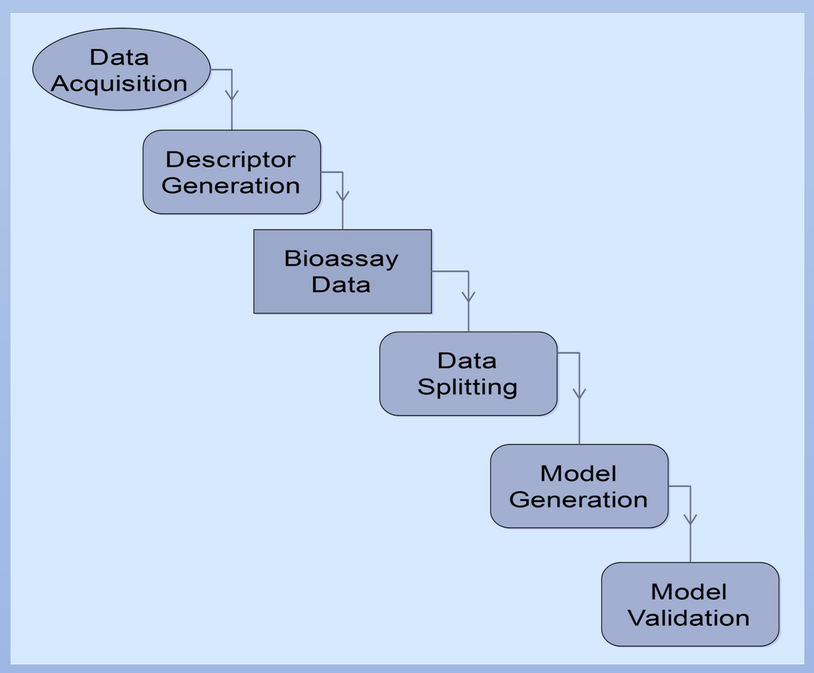
METHODS

Datasets: Bursi Mutagenicity Dataset, Benchmark dataset, AID1189, AID1194

Molecular Descriptors: 8 Properties Descriptors, 147 Pharmacophore Fingerprints, and 24 Weighted Burden Descriptors

Classifier Algorithms: Naïve Bayes, Random Forest, J48, SMO

Descriptor Calculation Tool: PowerMV
Data Mining and Classification Tool: WEKA

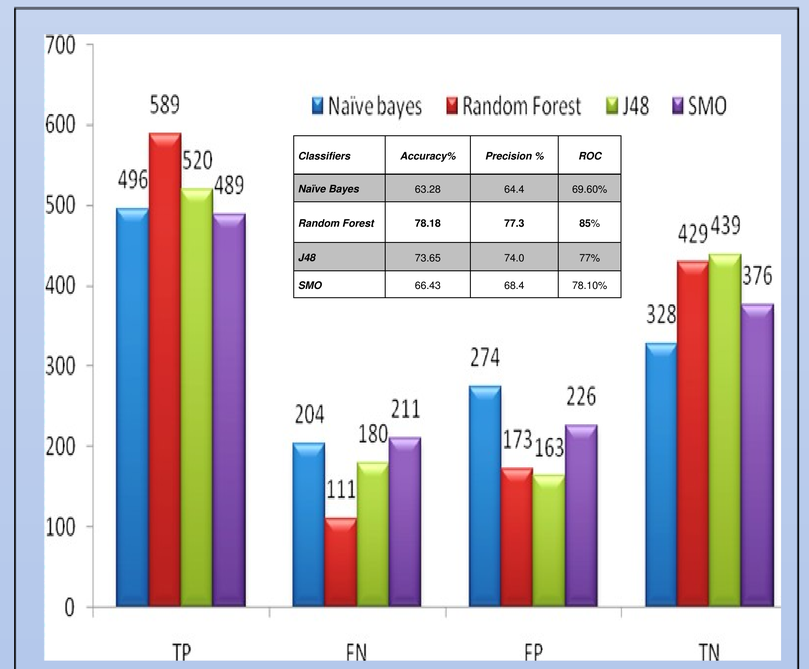


RESULTS

Set 1

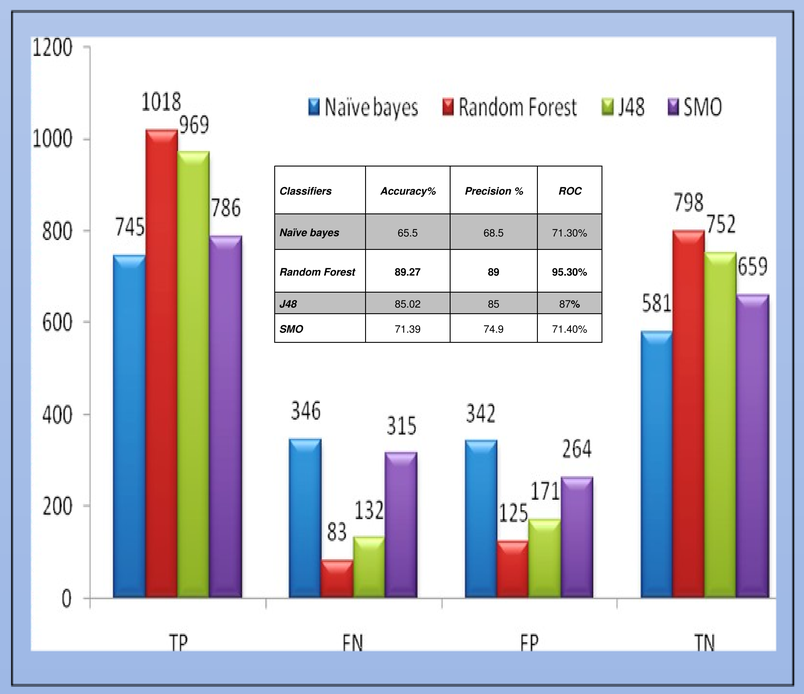
Classifiers	Accuracy%	Precision %	ROC
Naïve Bayes	64.59	67.8	71.90%
Random Forest	79.81	79.5	86.80%
J48	76.35	78.8	77.20%
SMO	70.24	73.1	77.10%

Set 2



Classifiers	Accuracy%	Precision %	ROC
Naïve Bayes	63.28	64.4	69.60%
Random Forest	78.18	77.3	85%
J48	73.65	74.0	77%
SMO	66.43	68.4	78.10%

Set 3



Classifiers	Accuracy%	Precision %	ROC
Naïve bayes	65.5	68.5	71.30%
Random Forest	89.27	89	95.30%
J48	85.02	85	87%
SMO	71.39	74.9	71.40%

Table 1. It shows the result of PubChem dataset, AID 1189 taken for External Validation

Classifiers	Datasets	Acc %	Precision %	Recall%	ROC
NaiveBayes	Set 1	49.08	53.3	36.80203	50.30%
	Set 2	49.28	53.7	36.29442	50.60%
	Set 3	49.08	53.7	33.37563	50.70%
Random Forest	Set 1	61.54	65.4	59.26396	**66.30%**
	Set 2	61.61	66.6	56.21827	**64.50%**
	Set 3	62.35	67.6	56.59898	**65.60%**
J48	Set 1	63.16	68.6	57.1066	64.60%
	Set 2	60.39	66	53.04569	62.50%
	Set 3	62.01	67.5	55.58376	61.40%
SMO	Set 1	50.57	55.3	38.57868	55.90%
	Set 2	57.14	63.2	46.95431	57.90%
	Set 3	57.81	64.7	46.06599	58.70%

Table2. It shows the result of PubChem, AID 1194 taken for External Validation

Classifiers	Datasets	Acc %	Precision %	Recall%	ROC
NaiveBayes	Set 1	55.76	54.3	42.78	57.50%
	Set 2	55.88	54.6	42.27	58.00%
	Set 3	57.21	56.9	40.50	58.40%
Random Forest	Set 1	81.12	78.3	83.29	**90.60%**
	Set 2	87.86	87.7	86.58	**94.30%**
	Set 3	91.94	91.8	91.13	**96.30%**
J48	Set 1	80.88	79.0	80.50	84.20%
	Set 2	84.37	85.7	80.50	86.20%
	Set 3	87.74	88.1	85.82	**90.30%**
SMO	Set 1	62.01	62.6	49.62	67.60%
	Set 2	69.23	71.8	57.97	68.70%
	Set 3	68.87	72.2	55.94	68.20%

CONCLUSION

PubChem is a vast repository of compounds containing many mutagenic molecules that can be taken up for in-silico predictive modeling. In our work we have created a new mutagenicity dataset containing more than 8000 compounds. From this work we recognized that based on the type of datasets and the descriptors used, Random Forest was the best classifier among the four classifiers to distinguish mutagens from non-mutagens.

REFERENCES

1) Amanda C Schierz, Virtual screening of bioassay data, J Cheminform., 2009, 1 (21).

2) J.Kazius, R.McGuire and R.Bursi, Derivation and Validation of Toxicophores for Mutagenicity Prediction, *J. Med. Chem.*, 2005, *48* (1).

For further details, please visit:
http://c2d.osdd.net/home/cheminformatics

Gordon Research Conference: Tuberculosis Drug Development, 3-8 July'2011, IL Ciocco Hotel and Resort, Lucca (Barga), Italy