A Benchmark Data Set for In Silico Prediction of Ames Mutagenicity

Katja Hansen[1], Sebastian Mika[3], Timon Schroeter[1], Andreas Sutter[2], Antonius Ter Laak[2], Thomas Steger-Hartmann[2], Nikolaus Heinrich[2] and Klaus-Robert Müller[1]

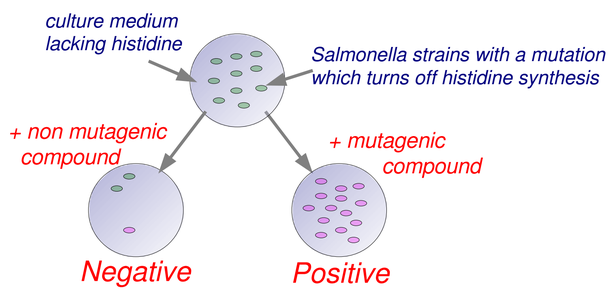
Introduction:

In silico prediction tools for Ames mutagenicity represent a cost-effective high throughput approach for prioritization of compounds before submission to experimental testing. Various modeling approaches have been pursued in the last years. But publicly available data sets are mostly very limited in terms of size and chemical coverage.

Our Approach:

- collect a representative data set
- evaluate different prediction systems on data set
- make data set publicly available for benchmark testing of other methods

Ames Test

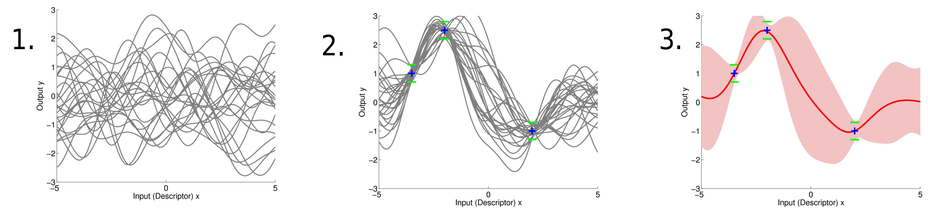


- positive labeled Compounds:
  significantly induces revertant colony growth at least in one strain, either in the presence or absence of a mammalian metabolizing system (S9 mix)
- negative labeled Compounds:
  does not induce revertant colony growth in any strain tested, both in the presence and absence of S9 mix

Machine Learning Methods:

Supervised Machine Learning Methods infer properties of unknown compounds from a training set of compounds.[5]
We considered 3 different Methods:

1. Gaussian Processes[6]



Technique from the field of Bayesian statistics:
(1) Specify a huge number of possible functions; (2) Eliminate those that don't agree with data; (3) Average over what remains: Prediction is a probability distribution

2. Support Vector Machines (SVM)
   Construct a separating hyperplane in a high dimensional feature space.
3. Random Forests
   Combine the predictions of 50 decision trees trained on random chosen features

Each compound is represented as a selected set of 904 Molecular Descriptors from DRAGON-X version 1.2 based on a 3D structure generated by CORINA version 3.4.

Results:

Evaluation of Models:

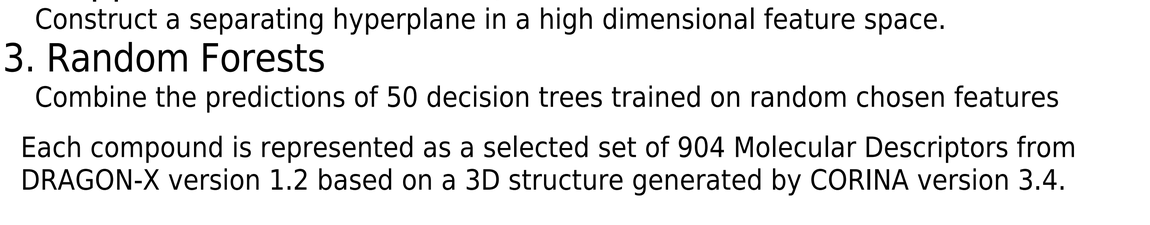
- 10 times 5 fold Cross Validation
- Performance Measures:
  1. Specificity: TN/(TN+FP)
  2. Sensitivity: TP/(TP+FN)
  3. Area under Curve (AUC):
     Plot false positive rate versus true positive rate. The Area under the resulting curve was used as optimization criterion for the learning Algorithms.

	GP	SVM	Forest
Specificity	75 %	75%	75%
Sensitivity	86 %	87%	83%
AUC	0,88	0,89	0,83

Examples of Prediction Results:

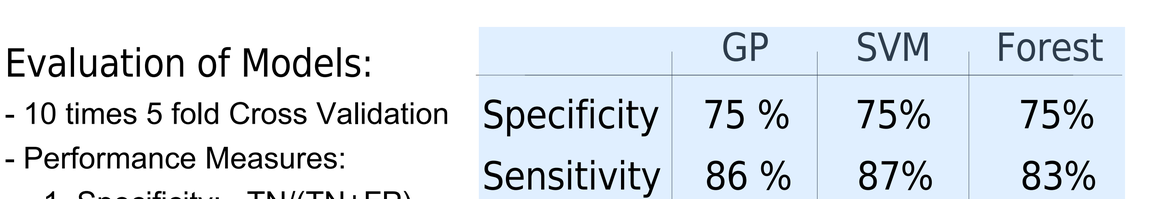
Six true negative oral drugs[4] and six true positive pharmacologically active compounds from the World Drug Index are shown to exemplify the diversity of chemical structures predicted.

Positive compounds:



Nitracrine, Carboquone, Amidox, Stemphyltoxin 3, Aflatoxin M1, Daniquidone

Negative compounds:

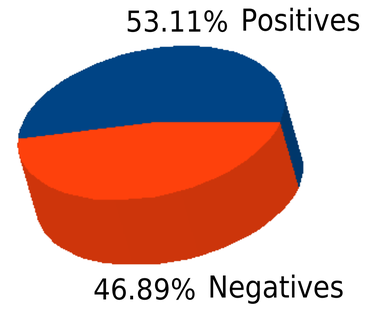


Amphetamine, Metoprolol, Tamoxifen, Chlorpheniramine, Ergotamine, Dapsone

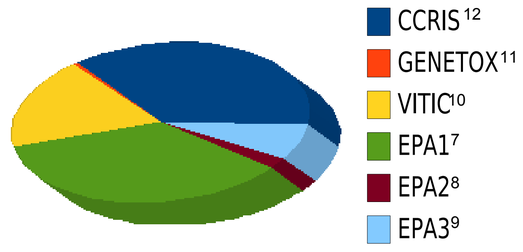
Data Set:

- 7096 compounds together with their activity in Ames mutagenicity test
  (1521 of them listed in World Drug Index)
- public sources (see diagram)
- balanced classes:
  3769 Positives vs 3327 Negatives

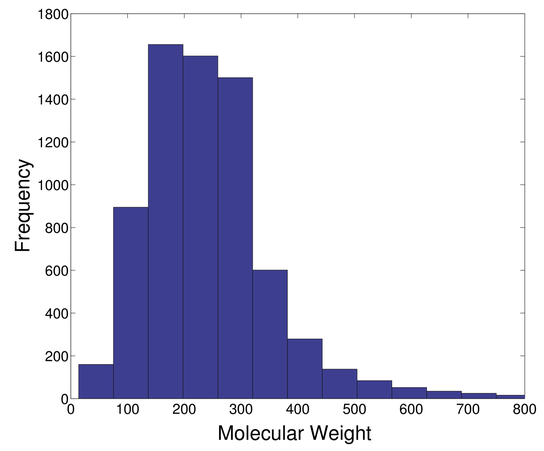
Ames mutagenicity test



Composition of Data Sources



Distribution of Molecular Weights



Public Benchmark Data Set:

http://ml.cs.tu-berlin.de/toxbenchmark

The website offers the structures of the 7096 compounds together with the corresponding Ames test results & references in SMILES and SD-format.
To facilitate comparative evaluation of methods please use the fixed cross validation splits. (10 times 3 folds)

Discussion:

The presented benchmark data set will facilitate the development and analysis of QSAR approaches for Ames mutagenicity.

All three evaluated methods yield satisfactory results on the benchmark data set. The Gaussian Processes and SVMs are superior to the Random Forests.

The evaluation of other prediction methods on the proposed benchmark data set remains an open issue.

References:

[4] Kasim NA et al. Amidon GL.Molecular properties of WHO essential drugs and provisional biopharmaceutical classification.Mol Pharm. 2004 Jan 12;1(1):85-96.

[5] Anton Schwaighofer et al. Accurate solubility prediction with error bars for electrolytes: A machine learning approach. Journal of Chemical Information and Modelling, 47(2):407-424, 2007

[6] Anton Schwaighofer et al. A probabilistic approach to classifying metabolic stability. Journal of Chemical Information and Modelling, 2008

[7] Kazius J et al. Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. 2005, 48, 312-320

[8] C. Helma et al. Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. J. Chem. Inf. Comput. Sci. 2004, 44, 1402-1411

[9] J. Feng et al. Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. J. Chem. Inf. Comput. Sci. 2003, 43, 1463-1470

[10] PN Judsom et al. Towards the Creation of an International Toxicology Information Centre. Toxicology 213(1-2):117-28, 2005

[11] Genetic Toxicity, Reproductive and Development Toxicity, and Carcinogenicity Database. http://www.fda.gov/Cder/Offices/OPS_IO/genrepcar.htm

[12] Chemical Carcinogenesis Research Information System. http://www.cancerinformatics.org.uk/matrix/CCRIS.htm

1 berlin

2 Bayer HealthCare
Bayer Schering Pharma

3 idalab intelligent data analysis