Structure-Activity Modeling - QSAR

Uwe Koch

Lead
Discovery
Center
LDC

QSAR

Lead Discovery Center LDC

Assumption:

QSAR attempts to quantify the relationship between activity and molecular strcucture by correlating descriptors with properties

Biological activity = function (parameters)

The QSAR relationship is derived from a training set of compounds with known activities

Objective:

Predictive and robust QSAR models to predict activity of untested molecules

Extract patterns related to biological activity to better understand the underlying principles of biological activity.

Application: Prioritize large number of compounds → less important to have each prediction correct

QSAR:Application I

Application:

QSAR relationships are most often used to predict the following quantities:

Biological:	**Chemical:**
Target binding	Boiling point
Toxicity	Melting point
Drug metabolism and clearance	Solubility
Mutagenicity	Stability
Permeability	
Protein binding	

REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals)

QSAR: Application II

REACH explicitly expresses the need to use (Q)SARs to reduce the extent of experimental animal testing

140.000 chemicals, cost animal testing per compound $200.000

QSAR models to prioritize compounds:

Only for compound predicted to be toxic animal testes will be performed

Reduce number of false negatives

According to OECD principles the QSAR model should have:

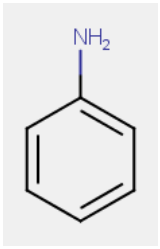
- A defined end-point
- An unambigous algorithm
- Defined domain of applicability
- Measures for goodness-of-fit, robustness, predictivity
- Mechanistic interpretation (if possible)

QSAR: Workflow

Lead Discovery Center LDC

Molecular Structures

Nc1ccccc1 1D



2D

3D

Molecular Descriptors

Constitutional
Electronic
Geometrical
Physicochemical
Steric
Quantum chemical
Topological

Data pre-processing

Normalization
Standardization
Feature selection
Outlier detection

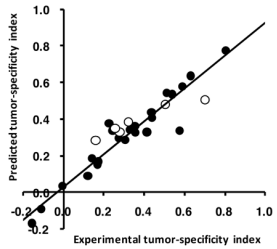
Data processing

Principle component analysis
Multiiple linear regression
Partial least squares
Neural Network
Support vector machine
Random Forest

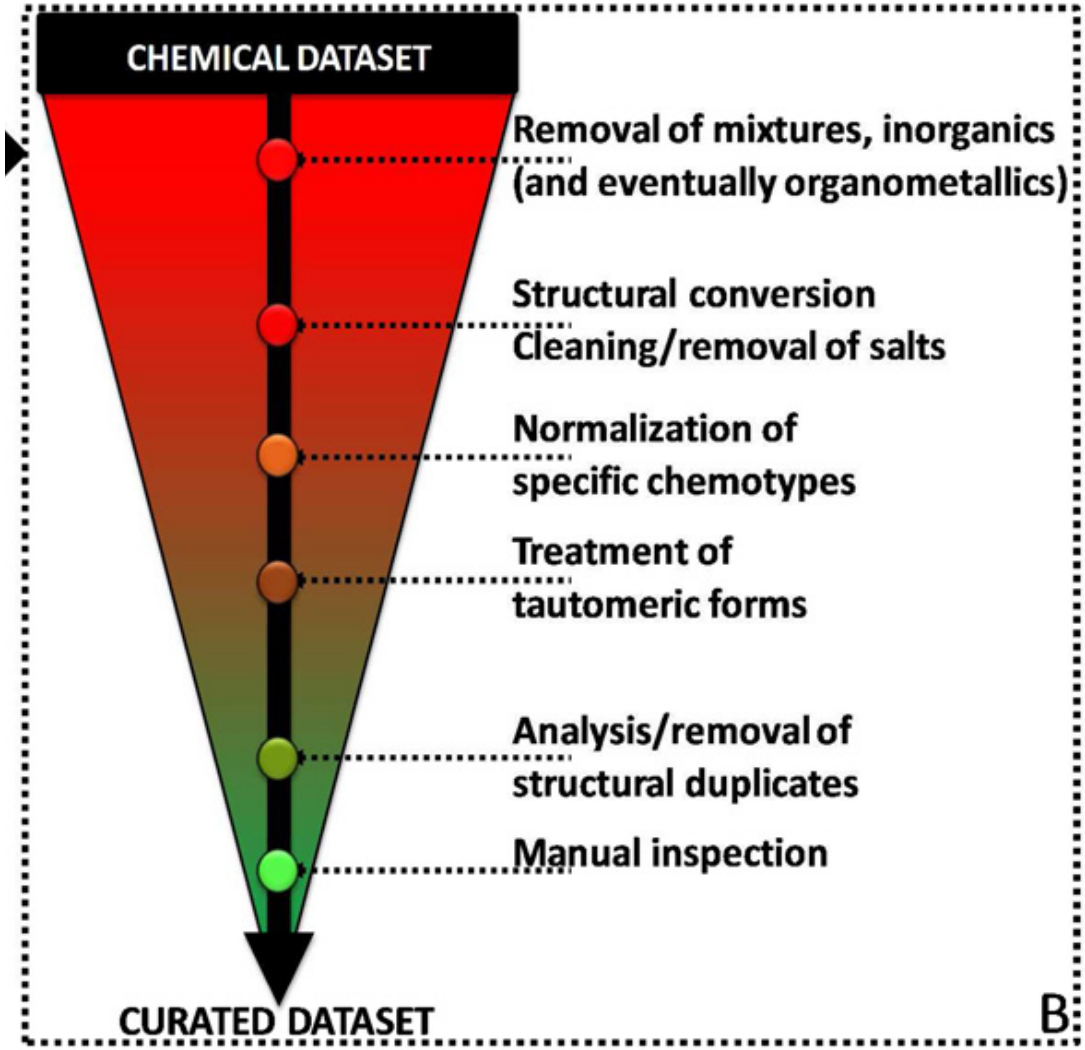
Statistical evaluation

R
R^2
Q^2
MSE
RMSE

Predicted activity



QSAR: Data preprocessing



QSAR:Descriptors

Lead Discovery Center LDC

Representation of a molecule – molecular descriptors: numerical values describing the properties of a molecule

Descriptors representing properties of complete molecules:

- log P, dipole moment, polarizability

Descriptors calculated from 2D graphs:

- topological indices, 2D fingerprints

Descriptors requiring 3D representations:

- Pharmacophore descriptors

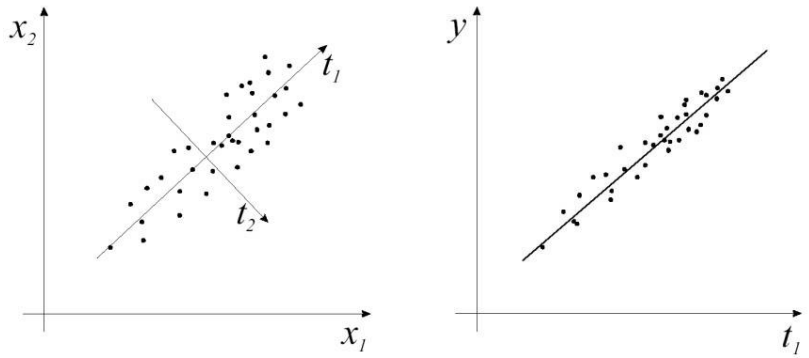
Data processing: Regression methods

Multiple linear regression: *Y = a0 + a1 * X1 + a2 * X2 + a3 * X3 + …+ an * Xn*

Stepwise multiple linear regression: A multiple-term linear equation is produced, but not all independent variables are used. Each variable is added to the equation in turn. A new regression is performed. The new term is retained only if the equation passes a test for significance.

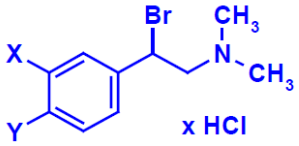
Principal components regression (PCR): A multiple-term linear equation is created based on a principal-components analysis transformation of the independent variables. Some of the last components are discarded to reduce the size of the model and avoid over-fitting.

Partial least squares: Useful when number of descriptors large. Latent variables are chosen as orthogonal linear combinations of original descriptors to provide maximun correlation with dependent variable.



QSAR

Example: Antiadrenergic Activities of N,N-Dimethyl-a-bromophenetylamines



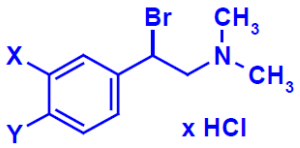
SAR Table with activities (-log IC50 = log 1/c)

meta	para	log 1/C	meta	para
H	H	7.46	Cl	F
H	F	8.16	Br	F
H	Cl	8.68	Me	F
H	Br	8.89	Cl	Cl
H	I	9.25	Br	Cl
H	Me	9.30	Me	Cl
F	H	7.52	Cl	Br
Cl	H	8.16	Br	Br
Br	H	8.30	Me	Br
I	H	8.40	Me	Me
Me	H	8.46	Br	Me

QSAR

Lead Discovery Center LDC

Example: Matrix for Hansch Analysis



meta (X)	*para* (Y)	log 1/C obsd.	π	σ+	E_s^{meta}
H	H	7.46	0.00	0.00	1.24
H	F	8.16	0.15	-0.07	1.24
H	Cl	8.68	0.70	0.11	1.24
H	Br	8.89	1.02	0.15	1.24
H	I	9.25	1.26	0.14	1.24
H	Me	9.30	0.52	-0.31	1.24
F	H	7.52	0.13	0.35	0.78
Cl	H	8.16	0.76	0.40	0.27
Br	H	8.30	0.94	0.41	0.08
I	H	8.40	1.15	0.36	-0.16
Me	H	8.46	0.51	-0.07	0.00
Cl	F	8.19	0.91	0.33	0.27
Br	F	8.57	1.09	0.34	0.08
Me	F	8.82	0.66	-0.14	0.00
Cl	Cl	8.89	1.46	0.51	0.27
Br	Cl	8.92	1.64	0.52	0.08
Me	Cl	8.96	1.21	0.04	0.00
Cl	Br	9.00	1.78	0.55	0.27
Br	Br	9.35	1.96	0.56	0.08
Me	Br	9.22	1.53	0.08	0.00
Me	Me	9.30	1.03	-0.38	0.00
Br	Me	9.52	1.46	0.10	0.08

π: hydrophobic additive parameter for a specific substituent (amount of change caused by this substituent)

π−values of aromatic substituents

	π		π
-CH3	0.52	-NO2	-0.28
-C(CH3)3	1.98	-OH	-0.67
-C6H5	1.96	-CO2H	-0.32
-C6H11	2.51	-NH2	-1.23
-CF3	0.88		

σ: Hammett sigma constant for the influence of meta and para-substituents on acidity of benzoic acid (σ+ for meta- para conjugated systems)

Esmeta: Taft equation – measures steric effect of substituents on ester hydrolysis (Me = 0)

Hansch Equation

Hansch equation

Hansch equation: log 1/C = 1.151 π -1.464 σ+ +7.817

(n = 22; r = 0.945; s = 0.196; F = 78.6; Q2 =0.841; Spress = 0.238)

N: nr. of compounds

R: internal validation correlation coefficient

S: standard deviation; measure of absolute quality of model (should be < 0.3)

Q2: cross validated R2. Repetition of regression on subsets of data many times (LMO: Leave many out), compute R on predicted values for molecules in test set. Overfitting if R2 >> Q2. R2 – Q2 should not exceed 0.3.

QSAR:COMFA

COMFA (Comparative molecular field analysis) 3D QSAR

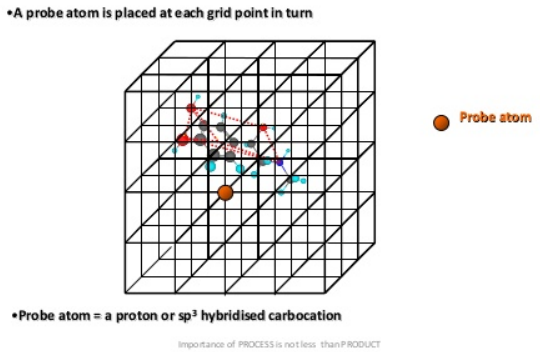
Derive a correlation between the biological activity of a set of molecules and their 3D shape, electrostatic and hydrogen bonding characteristics.

Superposition of 3d structures of ligands based upon a pharmacophore hypothesis

A grid box is placed around the superimposed molecules

Grid probes (C-atom, positive and negative charges) to calculate grid field values

Predict activitiy from energy values at grid points suing PLS (partial least squares)



Machine learning methods

Most often used as classification methods distinguishing actives from non-actives.

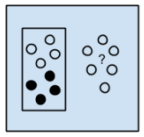
Training set: Active and inactive compounds and their descriptors.

Supervised learning methods:

Regression and classification problems

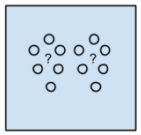
Training set with a known label (active/inactive)

Neural network, SVM, RF, regression methods



Unsupervised learning methods:

No label in training data, eg clustering, PCA



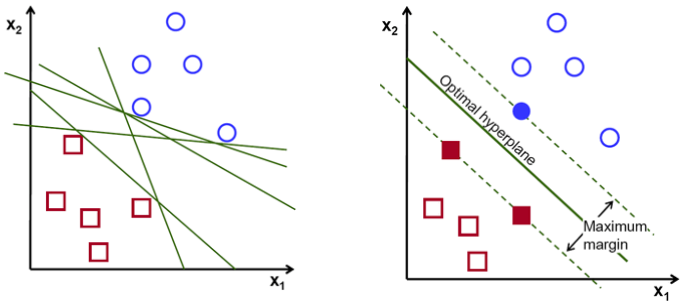
Machine learning methods

Two popular methods

SVM: Support vector machine defined by a hyperplane separating to classes of data.

Find the line passing as fas as possble from all points.

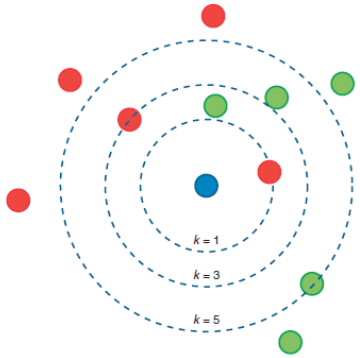
The optimal separating hyperplane maximizes the margin of the training data set.



kNN: Classifies an instance by majority vote of its k neighbours. Neighbours are usually identified on the basis of distance in feature space (eg Euclidian). Once labeled instances are available no explicit training is required → lazy learning algorithm.

For k =1 and 3 the blue instance will be classified as a member of the red class.

For k = 5 as a member of the green class (3 green vs 2 red)



Machine learning methods

Machine learning in virtual screening

Comparison of six different machine learning methods: SVM, ANN, RF (random forest), NB (naive Bayesian), kNN, DT (decision tree), TV (trend vector)

Method:

Five different targets: HIV-RT, Cox2, dihydrofolate reductase, estrogen receptor, thrombin.

Non-actives were taken at random from commercial databases.

Training set contained one third of actives published first. Test set with actives patented afterwards → IP novel (time-split cross validation)

Table 2. Prediction of Recently Published Compounds on the Basis of Earlier Published Compounds[a]

protein target	COX2	DH	TH	RT	AE
		Set I			
train (positive/negative)	36/2106	12/2151	34/2032	34/2353	13/2353
test (positive/negative)	77/8346	16/8390	78/8423	81 /8323	22/9053
		Set II			
train (positive/negative)	77/2106	16/2151	78/2032	81/2353	22/2353
test (positive /negative)	36/8346	12/8390	34/8423	34/8323	13/9053

[a] All compounds were taken from the VA set. Set I: one-third of the compounds published first used to generate a model to predict the later two-thirds of the compounds. Set II: two-thirds of the compounds published first used to generate a model to predict the later one-third of the compounds.

D. Plewczynski et al. ***JCIM*** 2006, 46, 1098

Machine learning methods

Machine learning in virtual screening

Model validation:

Classfication error E provides an overall error measure

$$E = \frac{\mathrm{fp} + \mathrm{fn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}} \times 100\%$$

Recall R measures % age of actives retrieved

$$R = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \times 100\%$$

Precision P: % age of positives (tp) correctly predicted

$$P = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \times 100\%$$

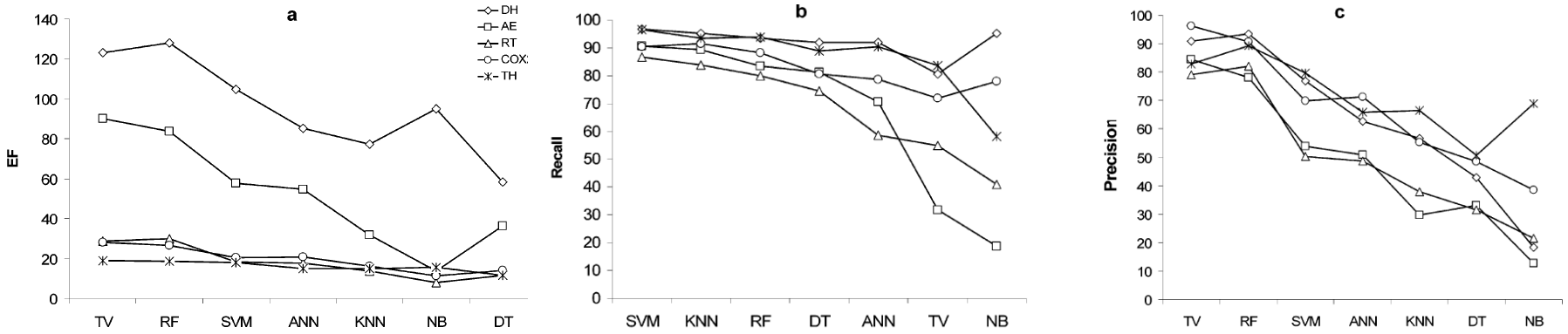
Enrichment factor EF

$$\mathrm{EF} = \frac{\mathrm{tp}/(\mathrm{tp} + \mathrm{fp})}{(\mathrm{tp} + \mathrm{fn})/(\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn})} \times 100\%$$

tp: true positives, fp: false positives, tn: true negatives, fn: false negatives

Machine learning methods

Test: Extracting small number of actives (<<1%) from a large compound collection



Enrichment:

Depends on Target (DH >> TH)

Ligands with characteristic features

High enrichment can be achieved with

Large number of fn → Recall

Recall (%age of actives retrieved):

SVM/kNN retrieve > 90% of actives

TV achieves high enrichment with poor Recall

Precision (%of tp of all positives):

TV and RF predict small mumber of false positives

ML methods can identify recently patented compounds based upon earlier piblications (scaffold hoppi

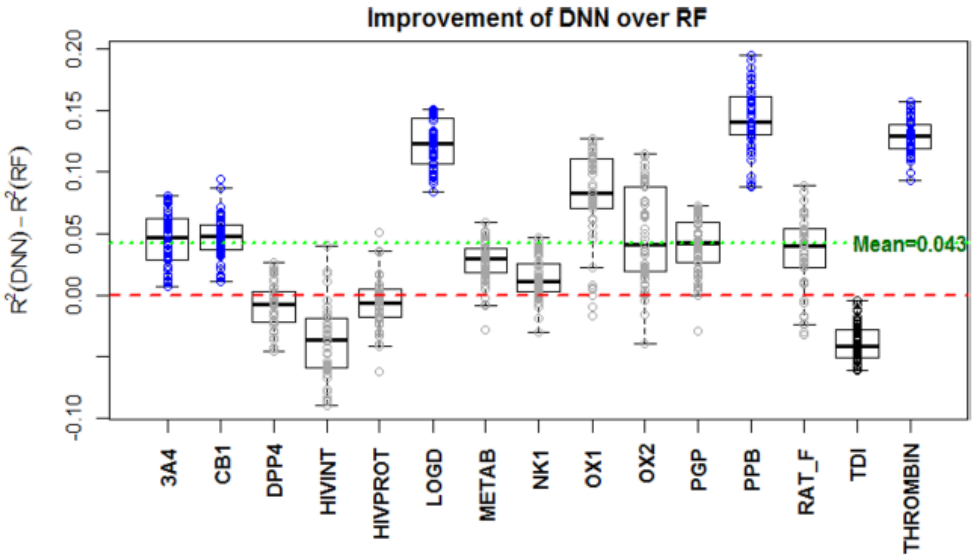
If one wants to retrieve as many actives as possible → SVM followed by kNN and RF

If one wants to avoid false positives → TV and RF followed by SVM

Machine learning methods

Recent Development: Deep Neural Nets (DNN)

In a recent Kaggle competition (www.kaggle.com) DNN performed better than RF and SVM



Compared to classical ANNs DNNs have more than one hidden layer and more neurons in each layer

Classical ANN could handle only a limited number of descriptors requiring descriptor selection.

J. Ma et al. ***J. Chem. Inf. and Mod***. 2015, 55, 263-272.

QSAR: Applicability domain

Even a robust and validated QSAR model can not predict activity of the entire universe of chemicals. Prediction is valid only if the compound is within the applicability domain of the model.

Four methods for definition of the applicability domain:

Range based

eg. bounding box: based on minimum and maximum values of each descriptor.

Geometric methods

Distance based

eg. K neares neighbors: similarity threshhold to nearest neighbour in the training set

Probability density distribution based

calculates probability density function for data set. Can identify internal empty regions.

R.Todeschini et al. ***Molecules*** 2012, 17, 4791-4810