Toxico-Cheminformatics and QSAR Modeling of the Carcinogenic Potency Database

Kun Wang[1], Ann Richard[2], Ivan Rusyn[3], and Alexander Tropsha[1]

[1]Laboratory for Molecular Modeling, School of Pharmacy, UNC-Chapel Hill; [2]National Center for Computational Toxicology, Office of Research & Development, US EPA; [3]Department of Environmental Sciences and Engineering, School of Public Health, UNC-Chapel Hill

INTRODUCTION

Fragment-based structure-activity relationship approaches to carcinogenicity and mutagenicity prediction, involving identification of toxicophores or structure-alerting features associated with activity classes (e.g., MultiCase, Derek, etc), are commonly employed methods for toxicity and virtual library screening of pharmaceuticals and industrial chemicals. The popularity of these approaches is due in large part to their simplicity, efficiency, and most importantly, the intuitive chemical/biological interpretability of the results. Whereas such approaches do well at identifying gross chemical features associated with activity, they do less well at predicting modulators of activity within structural classes due to lack of sufficient statistical representation of modulating fragment features within the dataset. In addition, mutagenicity evaluation, which is experimentally feasible in a medium-throughput screening mode and can be more reliably predicted than carcinogenicity, does not reliably predict “non-genotoxic” carcinogens. Both fragment-based approaches to prediction and mutagenicity as a predictor of carcinogenicity typically have high false-positive rates, which screen out many potentially useful drugs and chemicals unnecessarily.

A kNN (k Nearest Neighbors) Quantitative Structure-Activity Relationship (QSAR) approach is employed in this study that is built on the MolConnZ algorithm for chemical descriptor generation and a consensus model approach. MolConnZ descriptors span multiple facets of chemical structure, including structural functional groups (pre-defined fragments), topological, and electronic descriptors. Fragment and fragment descriptions do not delineate distinct activity classes in this approach, but rather are weighted and combined to provide optimal discriminatory power in the classification problem (active vs. inactive). In addition to the ability to identify nearest neighbors (or similarity neighborhoods in activity space), the presence of weighted contributions of fragment groups in the final kNN discrimination models can offer added interpretability. The generation of multiple kNN models, involving shuffling and optimization of training and test sets, and the use of performance thresholds to extract consensus models for the final prediction “model”, furthermore, have been shown to increase the stability and reliability of prediction models on external validation sets.

A number of kNN QSAR Consensus Prediction models have been generated for this study with the objective of using mutagenicity as a strong, but insufficient biological classifier for carcinogenicity, in conjunction with chemical structure determinants. To this end, different consensus prediction models have been generated for distinguishing:

- mutagens vs. non-mutagens **Model 1**
- carcinogens vs. non-carcinogens **Model 2**
- genotoxic carcinogens vs. non-genotoxic carcinogens **Model 3**
- genotoxic carcinogens vs. genotoxic non-carcinogens **Model 4**

Because these models capture different information in the biological activity and structure domain relevant to prediction, it is proposed that the use of these models in a tiered, confirmatory fashion can reduce the incidence of false positives and strengthen the overall prediction performance of the models. This concept can be generalized and extended to better integrate other types of carcinogenicity characteristics in aiding classification, e.g., tumor sites, TD50 range, multisite, multisex, multispecies tumor incidences, etc.

DATA

- All chemical structures and summary carcinogenicity and mutagenicity activity calls used in this study were extracted from the EPA DSSTox website (http://www.epa.gov/ncct/dsstox) Carcinogenic Potency Database – All Species SD file: **CPDBAS_v3b_1481_10Apr2007.sdf** (Source collaborator, L.S. Gold; Source website http://potency.berkeley.edu/). Summary mutagenicity and carcinogenicity activity data were obtained from the DSSTox CPDBAS fields: Mutagenicity_SAL_CPDB and ActivityCategory_SingleCellCall.
- CPDBAS contains 1481 chemical records. For kNN QSAR modeling purposes, the following conditions for chemical record inclusion applied: a mutagenicity call was available, a structure was available, not a mixture, no inorganic elements, no chirality, and no duplicated entry allowed. This left 693 unique chemical records for which a structure and both mutagenicity and carcinogenicity activity calls were available. Ability of mutagenicity to predict carcinogenicity in this set is 61% (30% false positives, 25% false negatives).

Number of Compounds	Mutagenic	Non-mutagenic	Total
Carcinogenic	252	172	424
Non-carcinogenic	85	184	269
Total	337	356	693

Model 3 · Model 4 · Model 2 · Model 1

METHODS

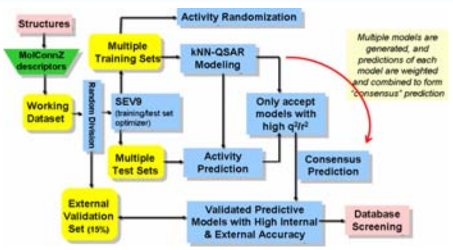

Structures
MolConnZ descriptors
Working Dataset
Random Division
External Validation Set (15%)
Multiple Training Sets
SEV9 (training/test set optimizer)
Multiple Test Sets
Activity Randomization
kNN-QSAR Modeling
Activity Prediction
Only accept models with high q²/r²
Consensus Prediction
Validated Predictive Models with High Internal & External Accuracy
Database Screening
Multiple models are generated, and predictions of each model are weighted and combined to form "consensus" prediction

Figure 1. kNN QSAR Consensus Prediction Approach

RESULTS & DISCUSSION

Model 1 **Mutagenicity models** (mutagenic vs. non-mutagenic)
55 models common to both Training and Test Sets, with prediction accuracy higher than 0.85, were used to formulate consensus prediction of validation set.

Dataset (693)	Maximum Prediction Accuracy	# Models in Prediction Accuracy Range		
		0.70-0.75	0.75-0.80	**0.85-0.90**
Training	0.92	23	1351	5067
Test	0.85	3183	772	69
Validation (105)	**Consensus Predic. Accuracy (55 models)**			**0.89**
	Sensitivity			0.84
	Precision			0.84

Model 2 **Carcinogenicity models** (carcinogenic vs. non-carcinogenic)
29 models, with prediction accuracy higher than 0.7 in Training Set and 0.65 in Test Set, were used to formulate consensus prediction of validation set.

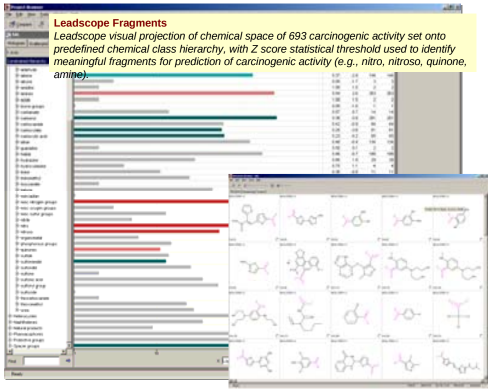
Dataset (693)	Maximum Prediction Accuracy	# Models in Prediction Accuracy Range		
		0.65-0.70	**0.70-0.75**	0.75-0.80
Training	0.82	1579	4487	812
Test	0.70	1085	0	0
Validation (105)	**Consensus Predic. Accuracy (29 models)**			**0.65**
	Sensitivity			0.85
	Precision			0.67

Model 3 **Carcinogenicity models** (genotoxic vs. non-genotoxic carcinogens)
59 models common to both Training and Test Sets, with prediction accuracy higher than 0.80, were used to formulate consensus prediction of validation set.

Dataset (424)	Maximum Prediction Accuracy	# Models in Prediction Accuracy Range		
		0.70-0.75	0.75-0.80	**0.80-0.90**
Training	0.94	40	729	5974
Test	0.89	1921	481	66
Validation (63)	**Consensus Predic. Accuracy (59 models)**			**0.80**
	Sensitivity			0.84
	Precision			0.87

Model 4 **Carcinogenicity models** (genotoxic carcinogens vs. genotoxic non-carcinogens)
20 models common to both Training and Test Sets, with prediction accuracy higher than 0.80, were used to formulate consensus prediction of validation set.

Dataset (337)	Maximum Prediction Accuracy	# Models in Prediction Accuracy Range		
		0.70-0.75	0.75-0.80	**0.80-0.90**
Training	0.92	1563	2470	1584
Test	0.88	1433	627	94
Validation (50)	**Consensus Predic. Accuracy (56 models)**			**0.80**
	Sensitivity			0.97
	Precision			0.79



Leadscope Fragments
Leadscope visual projection of chemical space of 693 carcinogenic activity set onto predefined chemical class hierarchy, with Z score statistical threshold used to identify meaningful fragments for prediction of carcinogenic activity (e.g., nitro, nitroso, quinone, amine).

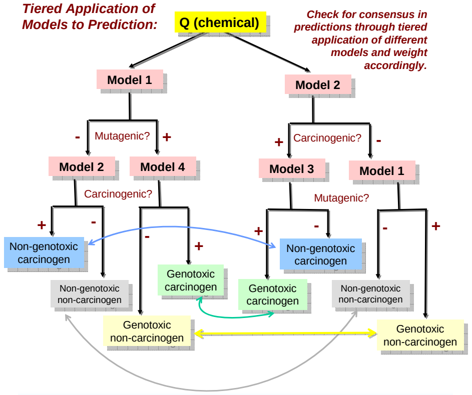
kNN QSAR Consensus Prediction Approach – MolConnZ descriptors
Frequent descriptor analysis of MolConnZ descriptors contributing to Model 2 (carcinogens vs. non-carcinogens) was performed following kNN QSAR Consensus Prediction Approach, identified groups common to Leadscope as well as additional groups, including carbonyl, aldehyde, peroxide.

Model 2	Descriptors		Descriptors	
	Count	Freq.	E-state	Freq.
Nitroso	nnitroso	117	Snitroso	124
Misc N	nssssNp	92	naaN	106
Nitro			Snitro	24
Hydrazine	Hhydrazine	83	Shydrazine	61
Carbonyl	ncarbonyl	42	Scarbonyl	38
Peroxide			Speroxide	41
Aldehyde	naldehyde	94		

Model 3	Descriptors		Descriptors	
	Count	Freq.	E-state	Freq.
Nitroso	nnitroso	42		
Misc N	nssssNp	36		
Nitro			Snitro	55
Amide	nsNH2	69	Hamide	124
Carbonyl			Sthiocarbonyl	36
Phosphate	nphosphate	41		
Sulfonate	nsulfonate	40		

- kNN QSAR Consensus Prediction Models 1-4 were built from the CPDB carcinogenicity/ mutagenicity dataset. Consensus model prediction accuracy ranged form 0.89 for Model 1 (Mutagenicity) to 0.65 for Model 2 (Carcinogenicity), with sensitivity (positive predictivity) 0.84 or higher for all 4 models.
- Mutagenicity alone predicts carcinogenicity in this dataset with 61% accuracy. In comparison, Model 2 Consensus Prediction Accuracy based on the MolConnZ descriptors alone (without mutagenicity) predicts carcinogenicity with 65% accuracy, slightly higher.
- MolConnZ group contribution descriptors overlap significantly with Leadscope-identified fragments and are well known structural alerts to carcinogenicity
- MolConnZ descriptors contributing to Models 1-4 provide coverage of different regions of chemical and activity space, with differences reflected in MolConnZ group contribution descriptors.

Tiered Application of Models to Prediction:



Check for consensus in predictions through tiered application of different models and weight accordingly.

CONCLUSIONS

- kNN QSAR Consensus Prediction Models 1-4 sample different areas of chemical and activity (carcinogenic and mutagenic) space, also as reflected in the different MolConnZ descriptors that contribute to the models
- Consensus model building optimally incorporates Training and Test set information, and creates stable models and improved validation statistics over single models in all cases
- Models 1,3 and 4 all have Consensus Prediction Accuracies of 0.80 or greater, whereas the Model 2 (carcinogenicity) is the least predictive at 0.65.
- Non-genotoxic carcinogens were well discriminated from genotoxic carcinogens by Model 3.
- We propose using a tiered approach in which Carcinogenicity prediction confidence is increased by incorporating a biological layer (genotoxic vs. non-genotoxic) and alternative routes to a “biological” consensus prediction.
- The CPDB as represented in the DSSTox CPDBAS data file relays a rich spectrum of activity information for each chemical substance. Future work will examine model dependence on alternative activity representations within the CPDB (e.g., tumor site, TD50 range, sex, species, multisite) and attempt to incorporate richer activity information into prediction schemes.

REFERENCES

- Richard AM, *Chem Res Toxicol 2006 Oct;19(10):1257-62.*
- Carcinogenic Potency Database (L.S. Gold): http://potency.berkeley.edu/
- EPA DSSTox Website: http://www.epa.gov/ncct/dsstox/
- Golbraikh A, Tropsha A, *J Comput Aided Mol Des. 2002, 16(5-6):357-69.*
- Roberts G, Myatt GJ, Johnson WP, Cross KP, Blower PE Jr., *J Chem Inf Comput Sci 2000;40(6):1302-14.*
- Benigni R., Giuliani A. *Med Res Rev. 1996 May;16(3):267-84.*
- Lauginin AA et al *Mutat Res 2005 Oct 3;586(2):138-46.*
- Helma C*, Mol Divers 2006 May;10(2):147-58.*

This poster does not necessarily reflect EPA policy. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.