COMBINING *IN SILICO* MODELS TO DEVELOP MORE ACCURATE CLASSIFICATION SCHEMES IN A SINGLE WORKFLOW TO PREDICT CYP3A4 INHIBITION AND HERG K+ CHANNEL AFFINITIES

Elena Fioravanzo (a), Nicola Cazzolla (b), Massimo Mabilia (a), Rosella Ombrato (b), Marco Daniele Parenti (a), Matteo Stocchero (a).

(a) S-IN Soluzioni Informatiche Via Salvemini 9, 36100, Vicenza, Italy (www.s-in.it)
(b) Angelini Research Center, ACRAF, P.le Stazione I-00040, S. Palomba, Pomezia (Rome), Italy

INTRODUCTION

In the late 1990s poor pharmacokinetics and toxicity were important causes of costly late-stage failures in drug development. Therefore, early consideration of ADME/Tox - Absorption, Distribution, Metabolism, Excretion and Toxicity - properties is increasingly seen as essential for efficient discovery and development of new drugs and drug candidates. Moreover, recently developed high-throughput ADME/Tox assays allowed accumulation of experimental data that are expected to provide us with statistically significant *in silico* models to be employed at the design stage of new compounds and compound libraries.

At present, blockade of hERG K+ channel is an unwanted side effect that must be detected as early as possible during drug development [1]. As the experimental data available suggest the binding affinity of a drug may vary as a function of the channel state (activated / inactivated) and that different drugs may bind to different known binding sites with different binding modes, ligand-based and structure-based approaches may be of limited use in screening databases. Also the inhibition of CYPs is detrimental for a potential drug candidate. The state of the art of computational methods used in CYP modeling varies considerably. This arises from limitations of the methods applied as well as from difficulties related to the complexity of the CYP enzymatic action itself. QSAR approaches appear challenging being a rapid, and rather successful approach. A single *in silico* ADME-Tox prediction model may provide acceptable results. As by definition all models are simulation of reality, and therefore they will never be completely accurate, sometimes a single model will not work. When multiple models and multiple approaches are combined in a single consensus score, however, more accurate predictions can be achieved. This idea prompted us to develop different QSAR models and to employ different prediction approaches in order to be able to get a consensus score more accurate then the single method to be used as a filter in the discovery process.

MATERIALS AND METHODS

Herg Dataset
62 compound from literature[3] with experimental IC_{50} values

CYP3A4 Dataset
64 compound from literature[2] with experimental IC_{50} and Ki values

Ligand Preparation with LIGPREP (Schrödinger)
- 2D-3D Conversion
- Check on Ionization state and tautomers
- Minimization

Molecular Descriptors Generation
- **Dragon** Descriptors (0D, 1D, 2D, 3D)
- **QikProp** Descriptors
- **ACD/Labs** Physchem Properties
- Molecular Fingerprints (**CANVAS**)

Variable Selection
- Costant or near costant variable filter (variance <0.1)
- Internal Correlation among X variables: maximum threshold 0.8/0.9
- Correlation between X and responses: minimum threshold 0.01 to 0.1

Applicability Domain Check
- Outlier detection with PCA analysis
- Hotelling's T2 and DmodX graphs

Training and Test set selection
- Onion D-Optimal Design on PCA scores

Regression Models (only for Herg prediction)
PLS and OPLS Modelling
QP Herg prediction

Classification Models
PLS-DA and OPLS-DA (Herg)
Naive Bayesian (CYP)
K Nearest Neighbour (CYP)
SVM (CYP)

Dragon is a software package for the calculation of molecular descriptors developed by Milano Chemometrics and QSAR Research Group. It allows calculation of more than 1600 molecular descriptors for thousands of molecules [6].

QikProp [7] (Schrödinger) has been developed by Prof. Bill Jorgensen at Yale University to rapidly predict ADMET properties of drug candidates. QikProp results have been fitted to datasets of drug-like molecules, based on 2-D and 3-D descriptors.

ACD/Labs LogD Sol Suite [8] is a complete array of tools for the prediction of molecular physical properties including pKa, logP, logD, and pH-dependent solubility. In addition, these packages predict an array of molecular properties including polar surface area, number of freely rotatable bonds, pH profiles of bioconcentration factor (BCF), adsorption coefficient (Koc), and "Rule-of-5" properties.

Canvas [9] is the chemo-informatics module of Schrödinger Suite; include, among other tools, the state-of–the-art tools for fingerprint and molecular descriptors calculation, statistical and cluster analysis, similarity/diversity searches.

Variables selection was carried out using a proprietary algorithm developed by S-IN.

Simca-P+ [4] was employed for a preliminary analysis of data matrices by means of PCA-X to detect both outliers and data structure.

Modde [4] was used for training and test set selection by means of a Onion/D-Optimal Design based on PCA scores.

Simca-P+ [4] was used for statistical analysis using different approaches: PLS and OPLS modeling has been used to investigate likely correlations between molecular descriptors and experimental activity values, PLS-DA and OPLS-DA has been employed to generate classification models. The optimal number of components in each PLS model was determined by the default cross-validation procedure.

KNIME [5] **and KNIME Extensions** [10] were employed both for statistical analysis (Naive Bayesian, K Nearest Neighbour and SVM) and for developing a general automation workflow. KNIME is a popular modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models. The Schrödinger KNIME Extensions are built upon the existing KNIME infrastructure and provide access to a wealth of ligand- and structure-based tools from the Schrödinger Suite. Glide, Prime, Phase, MacroModel, Jaguar, and other programs and utilities have Schrödinger nodes that enable core functionality.

Experimental classification was made considering 5.0 as threshold value of pIC_{50} and pKi, obtaining the following results:

Model	Active Molecules	% of Active Molecules	Inactive Molecules	% of Inactive Molecules
Herg	41	66%	21	34%
CYP3A4	21	33%	43	67%

RESULTS

hERG Models

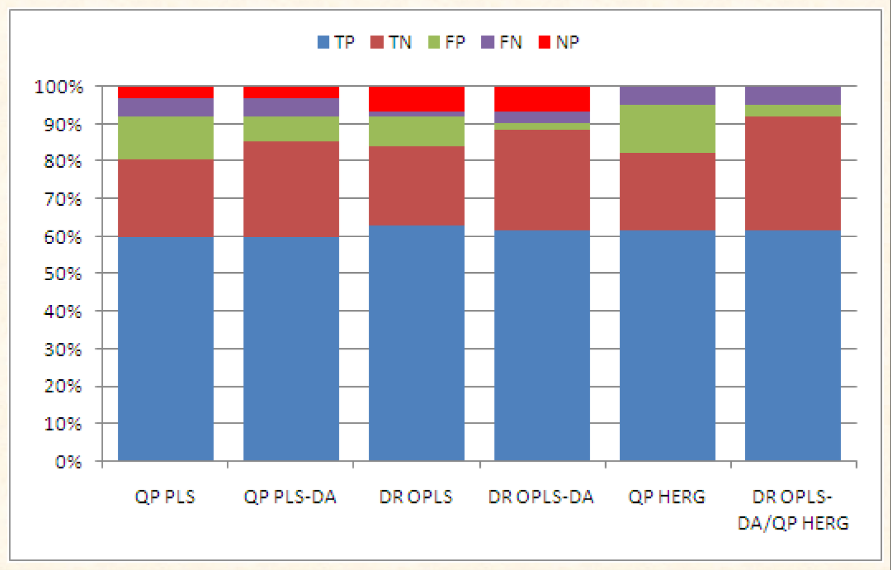
Table 1. Regression Models Results

Model	X	PCs	R2	Q2	SDEC	SDEP
QP PLS	4	2	0.64	0.53	0.83	0.92
QPHERG						0.97
QP PLS-DA	4	2	0.5	0.34		
DR OPLS	109	1+2	0.9	0.7	0.44	0.76
DR OPLS-DA	283	1+2	0.84	0.4		

Table 2. Classification Models Results

Model	TP	TN	FP	FN	NP	Accuracy	Selectivity	MCC
QP PLS	37	13	7	3	2	84	93	0.61
QP PLS-DA	37	16	4	3	2	90	93	0.73
DR OPLS	39	13	5	1	4	89	98	0.75
DR OPLS-DA	38	17	1	2	4	97	95	0.88
QP HERG	38	13	8	3	0	83	93	0.59
DR OPLS-DA/QP HERG	38	19	2	3	0	95	93	0.82

Figure 1. Training + test set Results



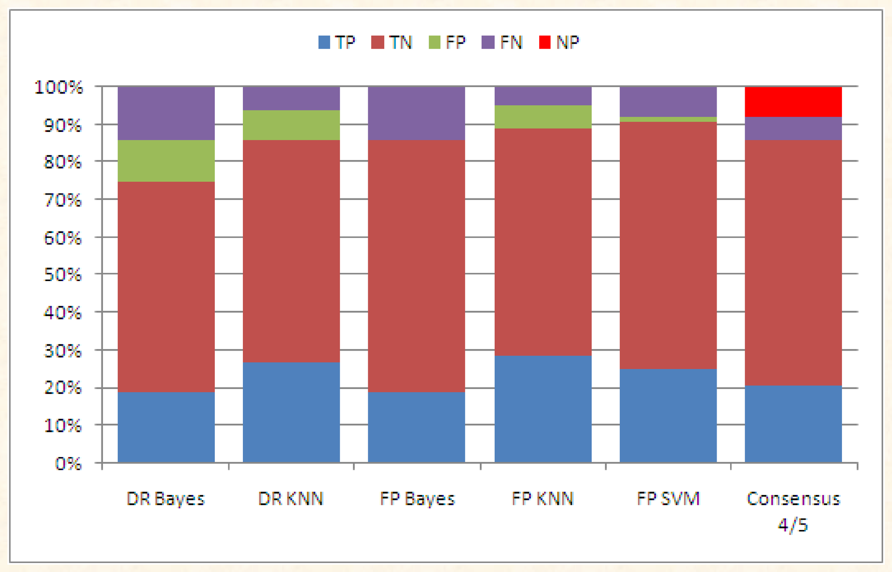
All hERG K+ channel affinities QSAR models showed a SDEP value of slightly less than 1 log unit [Table 1]: this is an interesting result if we consider the discrepancy observed between experimental IC50 values for hERG inhibition, determined for the same molecules in different laboratories, which may be greater than 1 log unit. Classification yielded a MCC between **0.59** and **0.88** [Table 2 and Figure 1]. Best classification scheme related to test set (results not shown) resulted Dragon OPLS, with an external test set of 20 structures a single FP was obtained. QPHERG may be extremely useful for structures outside other models applicability space.

CYP3A4 Models

Table 3. ClassificationModels Results

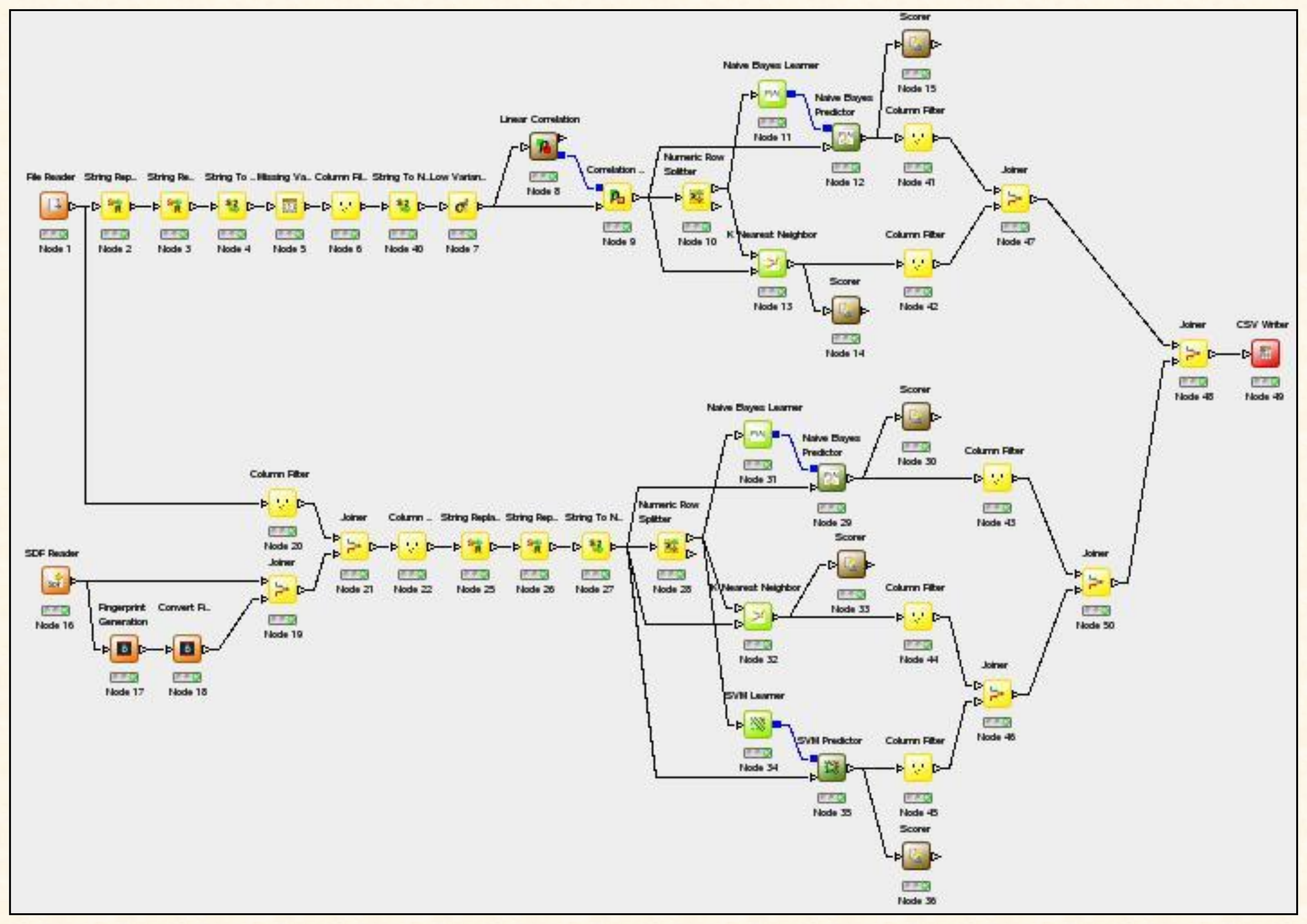
Model	TP	TN	FP	FN	NP	accuracy	selectivity	MCC
DR Bayes	12	36	7	9	0	63	57	0.42
DR KNN	17	38	5	4	0	77	81	0.69
FP Bayes	12	43	0	9	0	100	57	0.69
FP KNN	18	39	4	3	0	82	86	0.76
FP SVM	16	42	1	5	0	94	76	0.79
Consensus 4/5	13	42	0	4	5	100	76	0.84

Figure 2. Prediction Results



This dataset is not suitable for a quantitative regression model development as activity data for inhibition were determined with different experiment. Just qualitatively classification models were generated. Good results were obtained using Torsional Fingerprints as descriptor, and K Nearest Neighbour (KNN) or Support Vector Machine (SVM) as classification method yielding a MCC of **0.76** and **0.79** respectively [Table 3]. The Naïve Bayesian classifier showed a perfect accuracy (no FP), and the worst performance in terms of selectivity. Consensus criterion assigning the activity class according to at least 4 out of the 5 prediction available, otherwise the molecule is defined as not predictable (NP) led to the best overall results, with only 5 molecules classified as NP (8% of the total) [Figure 2].

KNIME Automation



KNIME Workflow for CYP3A4 Prediction

All developed models were implemented into a workflow built in KNIME employing both native processing nodes and Schrödinger KNIME Extensions. The developed workflow allows the user to easily predict a new compound according to different classification schemes leading him from structure preparation, across prediction and prediction validation towards final classification.

CONCLUSIONS

HERG: different PLS models were developed obtaining a SDEP values of about 1 log unit, very close to mean experimental errors. Classification models were developed obtaining a MCC of 0.88 (DR OPLS).
CYP3A4: a consensus classification scheme was developed obtaining a MCC of 0.84 and no FP.

The availability of different and independent methods and models able to predict hERG activity and CYP3A4 inhibition, allow these models to be used as a powerful *in silico* screening tool for drug discovery process.
The developed workflow allows the user to easily predict employing different models both a single molecule and a whole database of molecules starting from a simple SDF 2D file.

REFERENCE

[1] (a) Mitcheson et al. PNAS, 2000, 97, 12329. (b) Doyle et al. Science, 1998, 280, 69.
[2] (a) Ekins et al. J. Pharmacol. Exp. Ther. 1999, 290(1):429-38. (b) Pichard et al. Biochem. Pharmacol. 1996, 51:591–598. (c) Zhao et al. Br. J. Clin. Pharmacol. 1997, 44:505–511. (d) Ahlstrom et al. J. Med. Chem. 2008, 51, 1755–1763.
[3] G.M. Keserü, Bioorg. Med. Chem. Lett. 2003, 13, 2773.
[4] Umetrics AB, www.umetrics.com.
[5] Prof. Dr. Michael Berthold, University of Konstanz, Germany, http://www.knime.org/index.html
[6] R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry, WILEY-VCH, Weinheim, Germany, 2000.
[7] (a) E.M. Duffy, W.L. Jorgensen. J. Am. Chem. Soc. 2000, 122, 2878-2888. (b) QikProp, version 3.02, Schrödinger, LLC, New York, NY, 2007, http://www.schrodinger.com
[8] ACD/Labs LogD Sol Suite, version 11.0, Advanced Chemistry Development Inc. www.acdlabs.com
[9] Canvas, version 1.1110 (beta), Schrödinger, LLC, New York, NY, 2008, http://www.schrodinger.com
[10] KNIME Extensions, Schrödinger, LLC, New York, NY, 2007 http://www.schrodinger.com

ANGELINI

S-IN Soluzioni Informatiche
for Chemistry and Pharmaceutical Chemistry