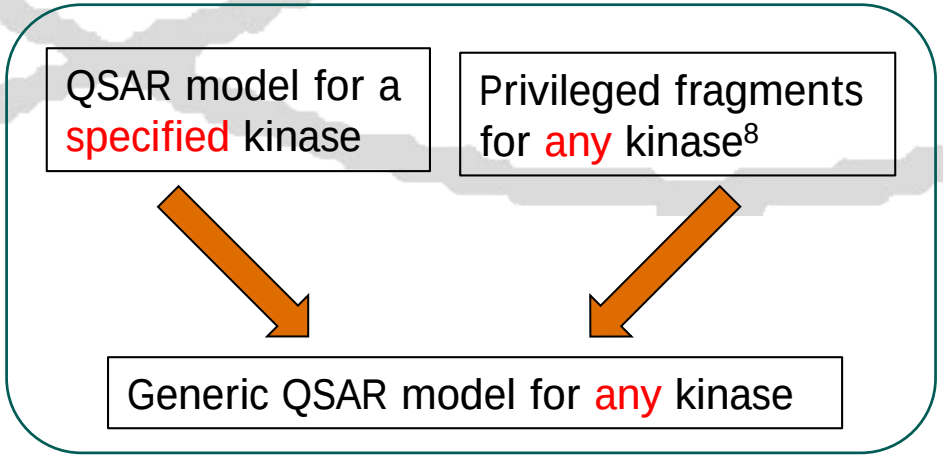
Galápagos

ICCS
International Conference on Chemical Structures
27 – 31 May 2018
#250, P-78

A new, improved model to predict kinase inhibition

Cornel Catana and Pieter FW Stouten
Galapagos NV, Generaal de Wittelaan L11 A3, 2800 Mechelen, Belgium



Abstract

Kinases constitute an important family of targets for Galapagos, as exemplified by filgotinib, which is currently in phase III clinical trials for RA and IBD. As part of its kinase HTS campaigns, Galapagos routinely and successfully screens a set of around 80,000 kinase-focused compounds. The set of compounds selected based on our home-grown kinase inhibition propensity ("kinase likeness") models exhibits a high hit rate. In order to further enhance these models (dating back to 2008 and 2011), we have recently developed new models since:

1) larger and new data sets, new descriptors, and improved statistical techniques have become available; and
2) while we previously exclusively selected models on the basis of their performance on IC_{50} data, our current goal is to develop a model that performs well both on HTS PIN (%inhibition) and IC_{50} data.

The training set contained ~88 k kinase-active compounds and ~84 k kinase-inactive compounds. A random forest (RF 2018 all) classification model was developed using Pipeline Pilot. In order to have an unbiased assessment of model performance against in-house data, a model was also developed without the ~22 k Galapagos compounds (RF 20018 NoG).

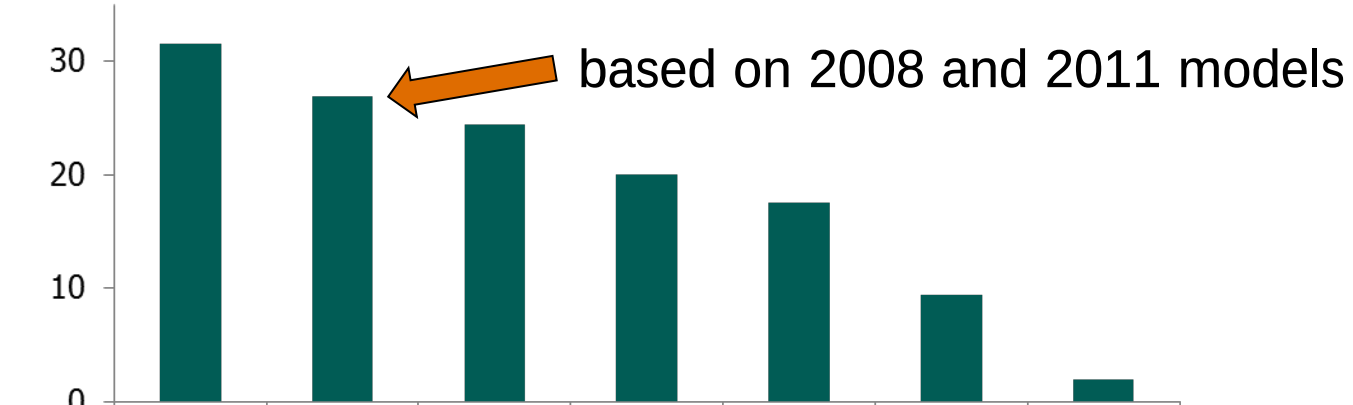
The models were tested on in-house HTS (PIN) data against 20 kinases. A compound was considered kinase-active if it was at least 2x a hit (PIN>75%) irrespective of the number of assays. It was considered kinase-inactive if it was assayed at least 5x and never was a hit. Other compounds were ignored. This stringency was used to account for the variability in single dose experiments. This test set comprised a total of 45,569 unique compounds, of which 3,339 were active.

Although the 2 new models performed less well on the test sets than on the training sets, they performed better than the 2 previous models, which had already been very useful in identifying kinase inhibitors. The "RF 2018 all" model is already being used in the process of selecting and acquiring kinase-focused compounds to expand our kinase-focused collection.

Introduction

- Galapagos routinely and successfully screens a set of around 80,000 kinase-focused compounds against kinase targets
- Compounds selected based on home-grown 2D QSAR models for generic kinase inhibition (“Kinase likeness models”) exhibit a high hit rate (Fig. 1)

Sources of hits (historical kinase HTS)



	In-house kinase projects	Kinase likeness models	BioFocus kinase library	Kinase virtual screens	Kinase others	Others	BioFocus non-kinase library
Hit rate	31.6	26.9	24.4	20.1	17.6	9.4	2.0

Fig. 1: Sources of hits (PIN>50%) for a representative kinase HTS

Objectives

1) Improve on previous models, taking advantage of larger data sets (more compounds, more kinases), new descriptors, and improved statistical techniques

2) Models to perform well when applied to IC_{50} test sets as well as to HTS PIN (%inhibition) test sets (where in the past only IC_{50} test sets were used)

Methods

- Training set
  - Kinase-active (model RF 2018 all): ~88 k compounds. From ref 1 and the Galapagos collection (pIC50 > 6)
  - Kinase-active (model RF 2018 NoG): ~66 k compounds. From ref 1 alone (no Galapagos compounds) (pIC50 > 6)
  - Kinase-inactive (both models): ~84 k compounds. From ref 2 (pIC50 < 5) and the decoys from ref 3
- Technique: Random Forest classification model developed with Pipeline Pilot[4] (customized R script), using “C” descriptors[7]
- Out-of-bag training set statistics
  - RF 2018 all: kappa = 0.940, accuracy = 97.0%
  - RF 2018 NoG: kappa = 0.935, accuracy = 96.8%
- Test sets
  - In-house HTS (PIN) data against 20 kinases (Table 1). Categories: 3,339 kinase-active (at least 2x a hit (PIN>75%)), 42,230 kinase-inactive (tested at least 5x and never a hit), 42,358 (all other compounds) are ignored. Stringency supposedly accounts for the variability in single dose experiments
  - Christmann-Franck[5] (Table 2): 2,101 compounds with IC_{50} data, 1,681 active (max pIC50 >= 6), 420 inactive (max pIC50 < 6)
  - Martin[6] (Table 2): 3,814 compounds with IC_{50} data, 856 active (max pIC50 >= 6), 2,958 inactive (max pIC50 >= 6)

Statistics on the test sets

Table 1: Model statistics for in-house test set (HTS PIN values)

Model	BEDROC α = 5	Number of hits retrieved and enrichment factor			kappa	AU ROC
		Top 1% (450)	Top 2% (900)	Top 5% (2,250)		
Bayes 2008	0.265	184 / 6	259 / 4	444 / 3	0.051	0.588
RF 2011	0.310	216 / 7	324 / 5	562 / 3	0.131	0.573
RF 2018 NoG	0.354	162 / 5	297 / 5	612 / 4	0.106	0.699
RF 2018 all	0.358	191 / 6	428 / 6	920 / 6	0.106	0.711

Table 2: Model statistics for two literature test sets (IC_{50} values)

Model	Christmann-Franck[5]		Martin[6]	
	kappa	AU ROC	kappa	AU ROC
RF 2011	0.230	0.687	0.187	0.570
RF 2018 NoG	0.475	0.707	0.434	0.739
RF 2018 all	0.450	0.694	0.413	0.723

Difference between training and test sets statistics

- Overtraining? No: subset of descriptors, subset of compounds (1/3 out-of-bag) used for each tree; convergence reached after 1,000 ~ 1,500 trees
- Activity cut-offs not the same, especially between training and PIN test set (pIC50 = 6 ~ PIN@1 µM = 50%;
  pIC50 = 5 ~ PIN@1 µM = 9%)
- Applicability domain: Galapagos kinase compounds differ from public compounds; assay conditions may differ as well
- Compounds regarded as inactive against an initial set of kinases may later need to be regarded as active when activity against other kinases is detected. In fact, of the 42,230 supposed inactives in the in-house test set, in the mean time 1,754 have been tested and 608 of those (35%) have been found to actually be active

Conclusions

- Our previous models (2008 and 2011) have proven very useful in identifying kinase inhibitors (see Fig. 1)
- We have developed two new and improved models.
  The ”RF 2018 all” model is already being used in the process of selecting and acquiring kinase-focused compounds to expand the Galapagos kinase-focused collection

References

1) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 2014, 42, 1083-1090.
2) Bora, A.; Avram, S.; Ciucanu, I.; Raica, M.; Avram, S. Predictive Models for Fast and Effective Profiling of Kinase Inhibitors. *J. Chem. Inf. Model.* 2016, 56, 895-905.
3) Rohner, S. G.; Baumann, K. Maximum unbiased validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* 2009, 49, 169-184.
4) Biovia Pipeline Pilot, 17.2.0, San Diego: Dassault Systèmes, 2017.
5) Christmann-Franck, S.; Van Westen, G. J. P.; Papadatos, G.; Beltran Escudie, F.; Roberts, A.; Overington, J. P.; Domine, D. Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound-Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* 2016, 56, 1654-1675.
6) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50 for Realistically Novel Compounds. *J. Chem. Inf. Model.* 2017, 57, 2077-2088.
7) Catana, C. Simple idea to generate fragment and pharmacophore descriptors and their implications in chemical informatics. *J. Chem. Inf. Model.* 2009, 49, 543-548.
8) Aronov, A. M.; McClain, B.; Moody, C. S.; Murcko, M. A. Kinase-likeness and kinase-privileged fragments: toward virtual polypharmacology. *J. Med. Chem.* 2008, 51, 1214-1222.