A Case Study in Predicting hERG Activity

Investigating Multiple (Q)SARs and Data Sources

Jeffrey Plante

Senior Research Cheminformatician

jeffrey.plante@lhasalimited.org

Background

- Collaboration between Merck KGaA in Darmstadt and Lhasa Limited in Leeds.
- Investigate the ability of using the SOHN methodology to predict pharmacophoric endpoints.
- hERG was chosen as the proof of concept as there is a large amount of public and private data obtained using established assays.

Experiment Design

- We investigated the interplay between using different (Q)SAR models and Data Sources.
- Data Sources
  - Public – ChEMBL
  - Private – Merck
- Models
  - Derek Nexus – Expert System (Lhasa)
  - SOHN – Statistical System (Lhasa)
  - Random Forest – Statistical System (Merck)

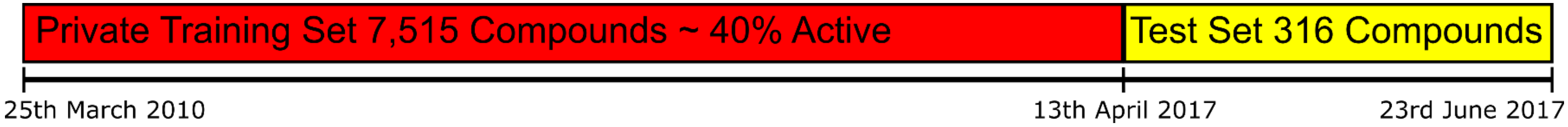
Dataset Design – Public Data

- ChEMBL 22 was mined by obtaining all compounds that had been assayed against hERG.
- These were grouped by compound and binarised using a 10 micromolar threshold.
- Where multiple results were present for an individual compound a single active call was sufficient to make the compound active.
- 7,861 Compounds 46.7% Active.

Dataset Design – Private Data

- Merck provided development compounds that had been assayed against their in house Patch-Clamp assay.
- These were grouped by compound and binarised using a 10 micromolar threshold.
- Where multiple results were present for an individual compound a single active call was sufficient to make the compound active.
- 7,515 compounds 40% Active.

Dataset Design – Test Set



Private Training Set 7,515 Compounds ~ 40% Active

Test Set 316 Compounds

25th March 2010

13th April 2017

23rd June 2017

- Temporally split test set.
- RF was used in the selection of compounds to synthesise.
- Test set results were held locked until the final predictions were made.

Model Design – Derek

- An expert system developed by Lhasa Limited.
- It consists of structural alerts that contain patterns that define a structure-activity relationship.
- The hERG endpoint has 5 alerts that contain a total of 38 patterns defined by a human expert after examining activity data from public and private sources.

Model Design – RF

- The RF was developed and optimised at Merck.
- It was implemented in Scikit-learn 0.17 in python 2.7.11 and trained using a number of physiochemical descriptors from RDKit as well as ECFP4 fingerprints.
- Each prediction was given a confidence score, which was the average of the maximum similarity of the query to the training data along with the overall model probability.

Model Design – SOHN

- The SOHN methodology was developed at Lhasa Limited.
- The descriptors used to generate a hypothesis are topological atom-pairs.
- These hypotheses are assembled into a network, which enables the system to use the most specific local kNN model for each appropriate hypothesis.
- Each prediction also reports a confidence score, which will range from 0.5 to 1 with a higher number indicating a higher confidence.

Results

Expert model	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🔵 Derek Nexus	0.75	0.63	0.43	0.84	0.44	0.84	0.27	0.27

🔵 Derek Nexus 🟣 SOHN 🔴 Private Data
🟠 Random Forest 🟢 Public Data

Lhasa Limited

Results

Expert model	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🔵 Derek Nexus	0.75	0.63	0.43	0.84	0.44	0.84	0.27	0.27

Statistical models (public)	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🟢 RF$_{\text{ChEMBL}}$	0.74	0.57	0.26	0.88	0.37	0.81	0.16	0.15
🟣🟢 SOHN$_{\text{ChEMBL}}$	0.73	0.66	0.54	0.78	0.42	0.86	0.30	0.29

🔵 Derek Nexus 🟣 SOHN 🔴 Private Data
🟠 Random Forest 🟢 Public Data

Lhasa Limited

Results

Expert model	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🔵 Derek Nexus	0.75	0.63	0.43	0.84	0.44	0.84	0.27	0.27

Statistical models (public)	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🟢 $\text{RF}_{\text{ChEMBL}}$	0.74	0.57	0.26	0.88	0.37	0.81	0.16	0.15
🟣🟢 $\text{SOHN}_{\text{ChEMBL}}$	0.73	0.66	0.54	0.78	0.42	0.86	0.30	0.29

Statistical models (private)	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴 RF_{Merck}	0.82	0.73	0.57	0.89	0.61	0.88	0.48	0.47
🟣🔴 $\text{SOHN}_{\text{Merck}}$	0.82	0.75	0.63	0.87	0.59	0.89	0.49	0.48

🔵 Derek Nexus 🟣 SOHN 🔴 Private Data
🟠 Random Forest 🟢 Public Data

Lhasa Limited

Results

Models	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴🟢 $\text{RF}_{\text{Merck+ChEMBL}}$	0.84	0.75	0.59	0.91	0.65	0.89	0.52	0.51
🟣🔴🟢 $\text{SOHN}_{\text{Merck+ChEMBL}}$	0.83	0.76	0.63	0.89	0.62	0.89	0.52	0.52

🔵 Derek Nexus 🟣 SOHN 🔴 Private Data
🟠 Random Forest 🟢 Public Data

Lhasa Limited

Results

Models	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴🟢 $\text{RF}_{\text{Merck+ChEMBL}}$	0.84	0.75	0.59	0.91	0.65	0.89	0.52	0.51
🟣🔴🟢 $\text{SOHN}_{\text{Merck+ChEMBL}}$	0.83	0.76	0.63	0.89	0.62	0.89	0.52	0.52

Statistical + Expert	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴🔵 RF + Derek	0.83	0.73	0.54	0.92	0.64	0.88	0.49	0.49
🟣🔴🔵 SOHN + Derek	0.84	0.75	0.59	0.91	0.64	0.89	0.51	0.51

🔵 Derek Nexus 🟣 SOHN 🔴 Private Data
🟠 Random Forest 🟢 Public Data

Lhasa
Limited

Results

Models	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴🟢 RF$_{\text{Merck+ChEMBL}}$	0.84	0.75	0.59	0.91	0.65	0.89	0.52	0.51
🟣🔴🟢 SOHN$_{\text{Merck+ChEMBL}}$	0.83	0.76	0.63	0.89	0.62	0.89	0.52	0.52

Statistical + Expert	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴🔵 RF + Derek	0.83	0.73	0.54	0.92	0.64	0.88	0.49	0.49
🟣🔴🔵 SOHN + Derek	0.84	0.75	0.59	0.91	0.64	0.89	0.51	0.51

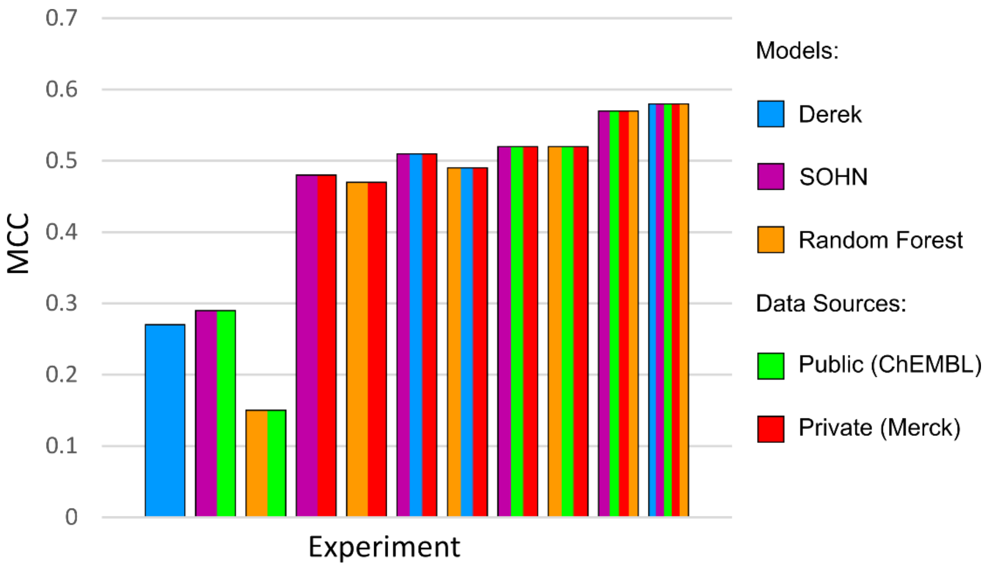
Models	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
🟠🔴🟢🟣 RF-SOHN	0.85	0.78	0.66	0.91	0.67	0.90	0.57	0.57

🔵 Derek Nexus 🟣 SOHN 🔴 Private Data
🟠 Random Forest 🟢 Public Data

Lhasa
Limited

Results Overall

Models	ACC	BA	SENS	SPEC	PPV	NPV	MCC	KAPPA
Pure Expert (Derek)	0.75	0.64	0.43	0.80	0.40	0.84	0.27	0.27
RF-SOHN	0.85	0.78	0.66	0.91	0.67	0.90	0.57	0.57
RF-SOHN-Derek	0.86	0.77	0.61	0.93	0.72	0.89	0.58	0.57



🔵 Derek Nexus 🟣 SOHN 🔴 Private Data

🟠 Random Forest 🟢 Public Data

Conclusions

- The SOHN methodology is able to successfully model hERG activity and should be applicable to other pharmacophoric models.
- Combining statistical systems together with expert systems, as well as using multiple sources of data results in increased performance.

Published as Hanser et al. Journal of Cheminformatics 2019 11:9