

Prediction of Cell Permeability

Paul Selzer¹

Peter Ertl¹, Daniela Gabriel¹, Christian N. Parker¹, Meir Glick², Mariusz Milik²

¹ *Novartis Institutes for BioMedical Research, Basel, Switzerland*

² *Novartis Institutes for BioMedical Research Inc., Cambridge MA, USA*

High content imaging allows one to analyze biological processes at the sub-cellular level, providing valuable contributions to hit finding. However, this assay format can be limited in terms of throughput due to difficulties in assay protocol and reagent cost. Therefore, there is a need to screen focused sets of compounds instead of randomly screening the full set of available compounds. This highlights the need for reliable and robust methods to predict cell permeability that help to concentrate the screening resources on those compounds that have a higher probability of being cell permeable.

To support this goal an in-house expert system for the prediction of the cell permeability of small molecules has been developed. The system was developed by the application of several machine learning techniques to this problem leading to a combination of two complementary approaches:

- Bayesian Model
The model has been developed with data resulting from cell based screens providing a binary classification (cell permeable vs. non cell permeable) as output. The chemical structures were encoded as Pipeline Pilot fingerprints. The experimental data have been divided into training and test set (4:1). Applying the model to the test set yielded a 9.5 fold enrichment (compared to random) cell permeable compounds in a virtual screen.
- Random Forest Model
This model was built on Caco2 assay data. The output of the model is the permeation coefficient P which is then binned to provide a binary classification (highly permeable vs. medium and low permeable). The model was validated in a pseudo prospective study – meaning that it has been built with compounds that have been screened earlier in time and tested against compounds that have been screened later. This experiment has been repeated several times with different time-point thresholds yielding ROC curve AUCs of 0.8.

Both methods were combined to provide a consensus score to rank compounds according to their cell permeability potential. Currently the system is being validated to analyze it's applicability for the productive screening set selection process.