Performance Prediction of the Influence Relevance Voter

Chloé-Agathe Azencott, S. Joshua Swamidass, and Pierre Baldi

UCIrvine

Institute for Genomics and Bioinformatics
Bren School of Information and Computer Sciences

Virtual High-Throughput Screening

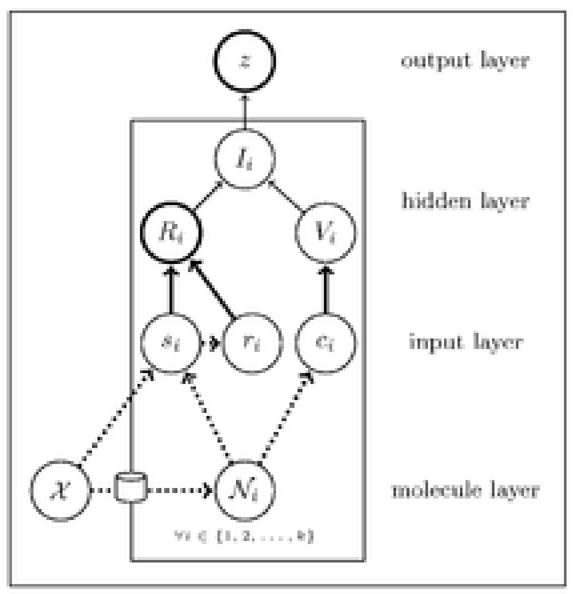
Virtual High-Throughput Screening (vHTS) is the cost-effective, in silico complement of experimental High-Throughput Screening (HTS). A vHTS algorithm uses data from HTS experiments to predict the activity of new sets of compounds in silico.

vHTS is most appropriately described as a ranking task, where the goal is to rank compounds such that active ones are close to the top of the prediction-sorted list as possible.

Influence Relevance Voter (IRV)

The k-Nearest Neighbors algorithm can be applied to chemical data, but does not perform optimally. The IRV uses a neural network architecture to learn how to best combine information from the nearest structural neighbors contained in the training set.

We compute nearest neighbors of chemicals using a standard MinMax similarity on structural fingerprints.



Benchmarked Performance

IJCNN07 Challenge HIV data: train on 4,229 compounds (149 actives), test on 38,449 compounds (1,354 actives).

McMaster 2005 DHFR data: train on 49,995 compounds (66 actives), test on 50,000 compounds (94 actives).

	BER	AUC
IJCNN07	0.283	0.771
SVM	0.269	0.764
IRV	0.271	0.762
MAXSIM	0.283	0.739

HIV data (IJCNN07 Challenge)

	EF1%	EF5%
McMaster	0.02	0.14
SVM	0.01	0.04
IRV	0.03	0.14
MAXSIM	0.00	0.03

DHFR data (McMaster Challenge)

Early Recognition

The BEDROC metric (Truchon and Bayly) quantifies the ability of a method to rank active compounds early at the top of the prediction-sorted test data.

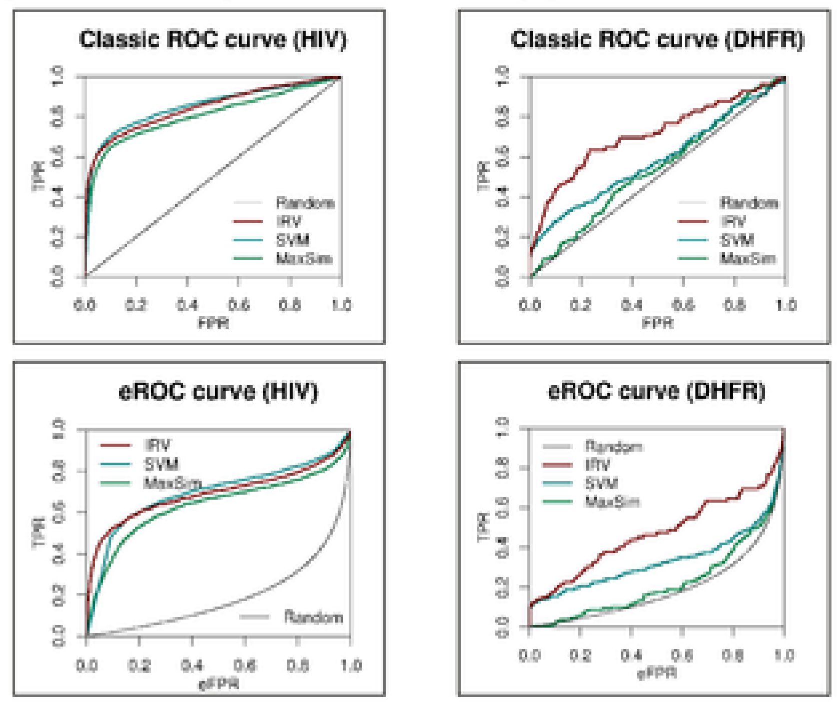
	Challenge	CV
SVM	0.469	0.573
IRV	0.500	0.630
MaxSim	0.439	0.526

BEDROC on the HIV data

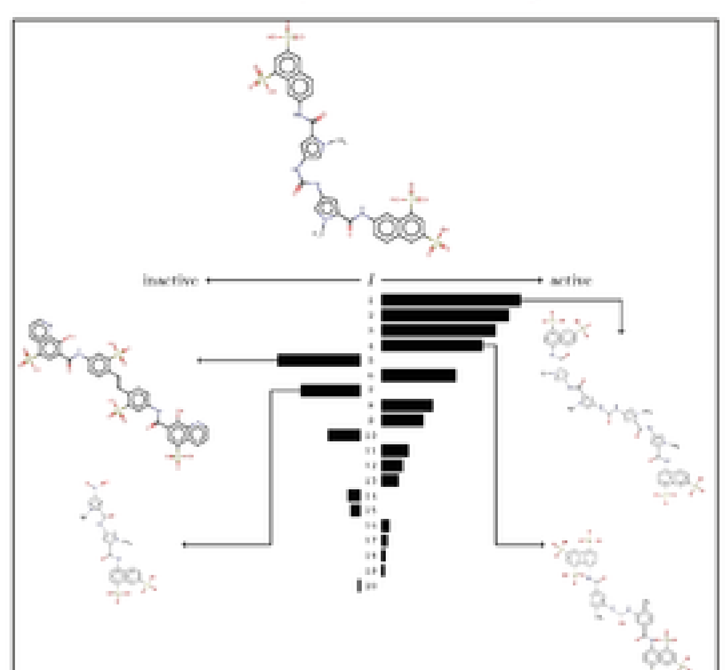
	Challenge	CV
SVM	0.084	0.200
IRV	0.100	0.251
MaxSim	0.045	0.062

BEDROC on the DHFR data

To better assess the results of vHTS experiments, we propose to replace traditional ROC curves with eROC curves, where an exponential transform has been applied to emphasize the importance of the early portion of the curve.



Interpretability



Performance Prediction

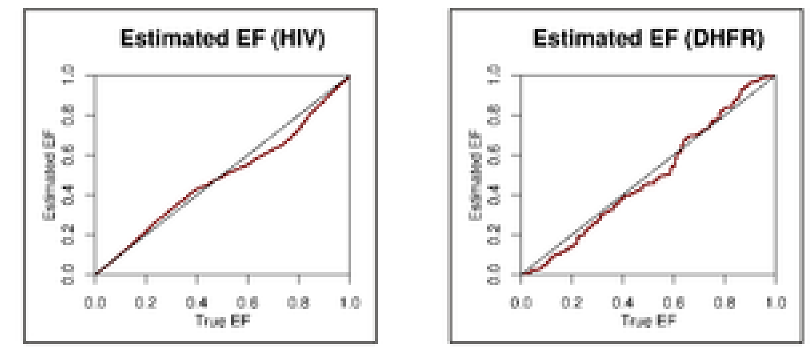
The output of the IRV is the probability that the corresponding instance is active. We can use the IRV to model the distribution of active compounds, and estimate the number of hits in a subset of the prediction-sorted list.

	Estimate	Actual
EF1%	0.25	0.23
EF5%	0.53	0.56
# Actives	1510	1503

HIV data (cross-validated)

	Estimate	Actual
EF1%	0.07	0.11
EF5%	0.20	0.23
# Actives	161	160

DHFR data (cross-validated)



Conclusion

We proposed a new vHTS algorithm, the IRV, with the following advantages: (1) the algorithm is suitable for early recognition and achieves state-of-the-art performance; (2) the underlying inferences are interpretable; (3) the output predictions have a probabilistic semantic; (4) the training time is very short; (5) the risk of overfitting is minimal, due to the small number of free parameters; (6) additional information can easily be incorporated into the architecture.

Moreover, we proposed a new visualization method, the eROC curve, to better assess the results of vHTS experiments.

Further Information

S. Joshua Swamidass, Chloé-Agathe Azencott, Ting-Wan Lin, Hugo Gramajo, Sheryl Tsai, and Pierre Baldi. The Influence Relevance Voter: an Accurate and Interpretable Virtual High Throughput Screening Method, J. Chem. Inf. Model., March 2009. DOI: 10.1021/ci8004379.

Contact: `cazencot@ics.uci.edu`

Acknowledgements

S. Joshua Swamidass and Prof. Pierre Baldi.

OpenBabel and OpenEye Scientific Software Academic Licences; SVMTorch implementation; ROCR package for R.

NIH/NLM Biomedical Informatics Training Grant; NSF Grants 0321290 and 0513376; Microsoft Research Award to PB.