Improving the Prediction of Toxicological Endpoints by Integrating Chemical, Biological and Phenotypic Data

Chad H. G. Allen,[1] Alexios Koutsoukas,[1] Isidro Cortes-Ciriano,[2] Daniel S. Murrell,[1]
Thérèse E. Malliavin,[2] Robert C. Glen[1] and Andreas Bender[1] (ab454@cam.ac.uk)

1. Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom
2. Unite de Bioinformatique Structurale, Institut Pasteur and CNRS URA 2185, Structural Biology and Chemistry Department, Paris, France

Introduction

The *in silico* prediction of *in vivo* toxicity of a compound is nontrivial, due in part to the lack of any direct, linear correlation between structural features and toxicity. However, chemical, protein target and phenotypic data provide complementary bioactivity information, and hence the hypothesis of this work was that more accurate toxicity predictions may be afforded by integration of such heterogeneous data. Recent studies[1,2] have demonstrated that integrating phenotypic descriptors derived from *in vitro* studies into QSAR models improve the accuracy of these models' predictions. We go further, additionally integrating so-called "protein affinity descriptors" into a toxicity prediction model.[3]

Protein target predictions as descriptors

To generate protein affinity descriptors, we have employed an *in silico* algorithm for predicting likely protein targets of compound based on their structure, trained on 190,000 ligand-protein interactions across 140,000 unique compounds extracted from ChEMBL v14. Using this approach, we can generate protein affinity descriptors in the form of "scores" corresponding to the likelihood of activity against a panel of 477 human proteins.[4] These scores may then be used as biological descriptors in toxicity prediction models.

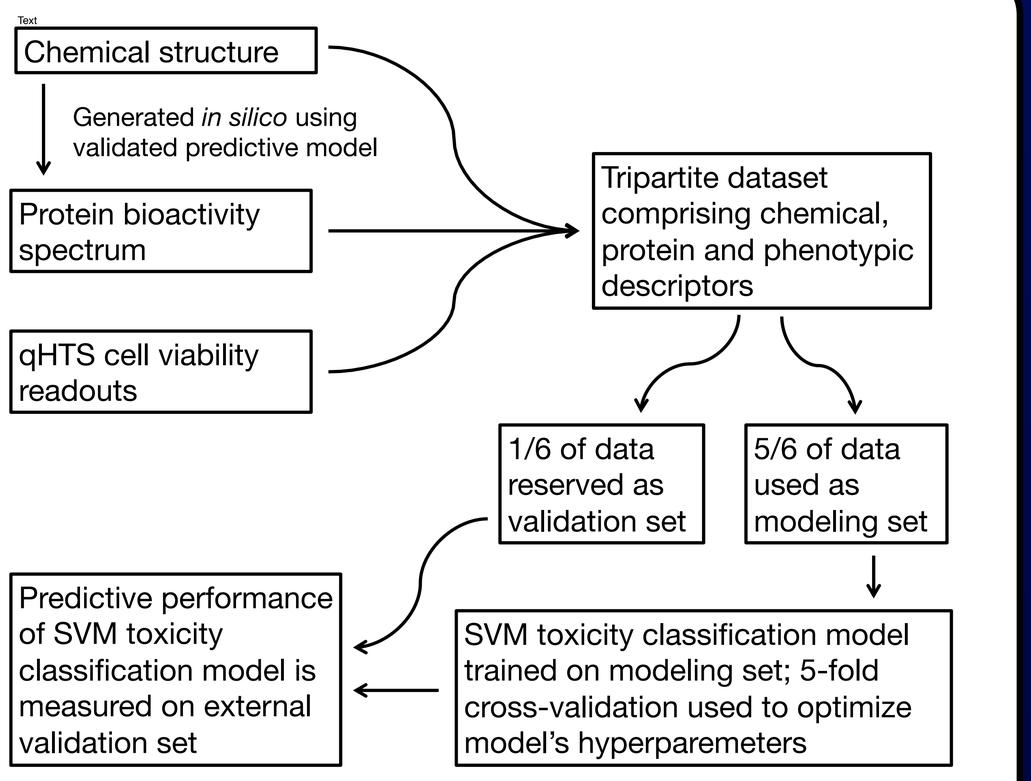

Figure 1: Modeling workflow diagram. The workflow was repeated using all three descriptors, chemical descriptors only and chemical descriptors along with either phenotypic or protein affinity descriptors

Data domain combination	CCR	Sensitivity	Selectivity
Chemical	0.74	0.63	0.85
Chemical & phenotypic	0.80	0.69	0.91
Chemical & biological	0.83	0.75	0.91
Chemical, phenotypic & biological	0.92	0.88	0.96

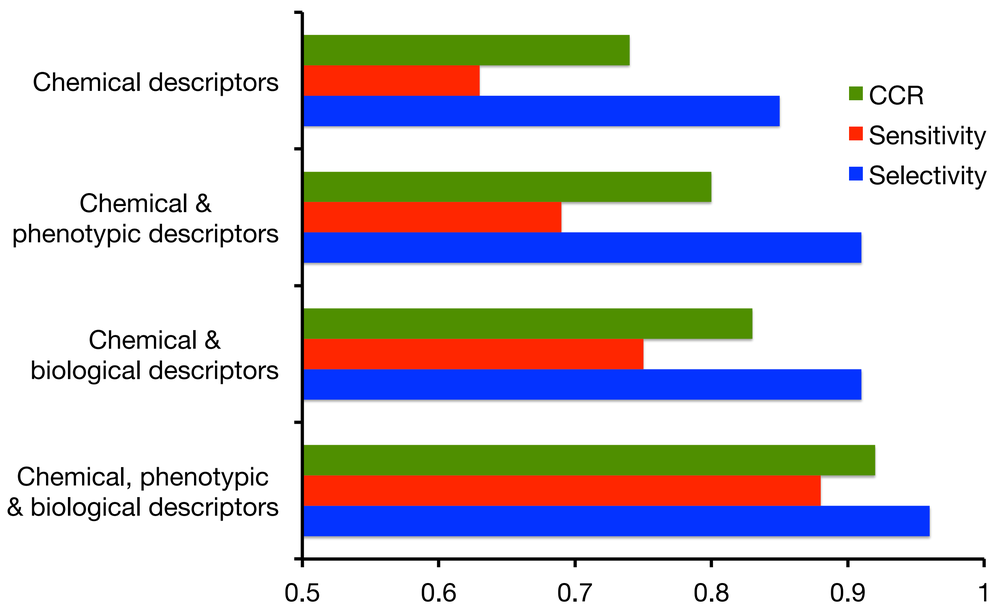

Figure 2: *top*, comparison of the predictive performance (measured by correct classification rate (CCR), sensitivity and selectivity) of various combinations of data domains; and *bottom*, graphical representation of these data. The performance of the model built from chemical, phenotypic and biological descriptors outperforms the other models in every performance metric.

Materials and methods

A tripartite dataset was collated, comprising 367 substances classified as toxic or nontoxic using rat 50% lethal dose values[5] along with three sets of descriptors for every compound in the set, providing data from chemical, protein affinity and phenotypic domains respectively. The three descriptors sets were 1D & 2D molecular descriptors calculated using PaDEL-Descriptor,[6] *in silico* protein affinity descriptors, and dose-response cell viability readouts from quantitative high-throughput-screening assays on 13 cell lines acquired from PubChem and filtered for noise as described by Sedykh *et al.*.[1]

As illustrated in Figure 1, this dataset was used to build four toxicity prediction models using all three data domains, chemical descriptors only, and chemical descriptors along with either phenotypic or protein affinity descriptors. One sixth of the data was selected at random to be kept apart as an external validation set, and a radial kernel support vector machine classification model was trained on the remainder of the data, employing five fold cross-validation to optimise the model's hyperparamaters.

Results, conclusions and further work

As shown in Figure 2, the correct classification rate (CCR) of the model generated using all three sets of descriptors was 0.92, compared with 0.74 for the analogous model built using only chemical descriptors, and 0.80 and 0.83 for models built using chemical descriptors along with either phenotypic or biological descriptors, respectively. In addition, the difference between sensitivity and selectivity improved from 0.22 (0.63 sensitivity, 0.85 selectivity) for the model built only with chemical descriptors to 0.08 (0.88 sensitivity, 0.96 selectivity) for the triply integrative model.

The increasing predictive accuracy of the models built using successively more data domains illustrates the utility of the heterogeneous data integration strategy. In future work we will apply the methodology developed to novel datasets such as those derived from ToxCast, as well as exploring further types of biological descriptors.

Acknowledgments

The work of Alexander Sedykh, Ivan Rusyn, Alexander Tropsha and colleagues in preparation of the data used in this study is gratefully acknowledged. The CEFIC-LRI is thanked for providing the funding for this work. ICC thanks Institut Pasteur and the Pasteur-Pasteur International PhD Programme for funding. TM thanks the Institut Pasteur and CNRS for funding.

References

1. A. Sedykh, H. Zhu, H. Tang, L. Zhang, A. Richard, I. Rusyn, and A. Tropsha, Environ. *Health Perspect.*, 2011, **119**, 364–370.
2. I. Rusyn, A. Sedykh, Y. Low, K. Z. Guyton, and A. Tropsha, *Toxicol. Sci.*, 2012, **127**, 1–9.
3. A. Bender, J. L. Jenkins, M. Glick, Z. Deng, J. H. Nettles, and J. W. Davies, *J. Chem. Inf. Model.*, 2006, **46**, 2445–2456.
4. A. Koutsoukas, R. Lowe, Y. Kalantar Motamedi, H. Mussa, W. Klaffke, J. Mitchell, R. C. Glen, and A. Bender, *J. Chem. Inf. Model.*, 2013, **53**, 1957-1966.
5. H. Zhu, T. M. Martin, L. Ye, A. Sedykh, D. M. Young, and A. Tropsha, *Chem. Res. Toxicol.*, 2009, **22**, 1913–1921.
6. C. W . Y ap, *J. Comput. Chem.*, 2010, **32**, 1466–1474.