

MolClass: a web portal to interrogate diverse small molecule screen datasets with different computational models

Jan Wildenhain^{1,*}, Nicholas FitzGerald² and Mike Tyers^{1,3,*}

¹Wellcome Trust Centre for Cell Biology and School of Biological Sciences, The University of Edinburgh, Edinburgh EH9 3JR, UK, ²Department of Computer Science and Engineering, University of Washington, Seattle, WA 98122, USA and ³Institute for Research in Immunology and Cancer, Université de Montréal, Québec, Canada H3T 1J4

Associate Editor: Jonathan Wren

ABSTRACT

Summary: The MolClass toolkit and data portal generate computational models from user-defined small molecule datasets based on structural features identified in hit and non-hit molecules in different screens. Each new model is applied to all datasets in the database to classify compound specificity. MolClass thus defines a likelihood value for each compound entry and creates an activity fingerprint across diverse sets of screens. MolClass uses a variety of machine-learning methods to find molecular patterns and can therefore also assign *a priori* predictions of bioactivities for previously untested molecules. The power of the MolClass resource will grow as a function of the number of screens deposited in the database.

Availability and implementation: The MolClass webportal, software package and source code are freely available for non-commercial use at <http://tyerslab.bio.ed.ac.uk/molclass>. A MolClass tutorial and a guide on how to build models from datasets can also be found on the web site. MolClass uses the chemistry development kit (CDK), WEKA and MySQL for its core functionality. A REST service is available at <http://tyerslab.bio.ed.ac.uk/molclass/api> based on the OpenTox API 1.2.

Contact: jan.wildenhain@ed.ac.uk or md.tyers@umontreal.ca

Received on February 13, 2012; revised on June 11, 2012; accepted on June 12, 2012

1 INTRODUCTION

Bioactive molecules can serve as powerful tools for interrogation of biological systems and/or as precursors in drug discovery. An objective in chemical systems biology is to model biological systems in order to understand the effects of small molecules on cellular processes, and thereby explain the basis for small molecule action (Hopkins, 2008). Realization of this ambitious goal will require extensive experimental datasets. The generation of chemical datasets from biological screening assays is usually limited by cost and throughput. Pharmaceutical companies and academic groups use high-throughput screens to test large libraries of small molecules that elicit a desired biological response, typically against a single target or at most a few related targets. However, chemical space is estimated to contain on the order of 10^{60} molecular entities, which greatly exceeds even the multi-million compound libraries at the disposal

of large pharmaceutical companies (Dobson, 2004). This vastness of chemical space requires that researchers devise rational approaches for identifying small bioactive molecules, particularly given the severe resource constraints on academic screening initiatives. The computational evaluation of potential bioactive molecules can drive down the high cost of screens and help extract potential drug-like compounds from pre-existing data in the public domain. To enable the extraction of information from existing chemical screen data, we have developed a suite of machine-based learning tools that statistically rank each compound for any given assay in a user-defined database. MolClass will thus facilitate the identification of specific bioactive molecules and allow the prediction of moieties that underpin biological activity.

2 WORKFLOW FEATURES

Existing resources for chemical screen data, notably PubChem, ChEMBL and ChemBank, are passive repositories that house an incomplete matrix of small molecule activity across submitted screens of various types, ranging from *in vitro* binding and enzyme assays to complex cellular and whole organism phenotypic assays (Fig. 1A). To interrogate such data, MolClass generates a complete matrix of compound activities across many screens and thereby enables functional predictions for all molecules, even if not tested in a specific screen. The user can upload input datasets of up to 20 000 molecules in SDF file format, in which tags distinguish hit from non-hit compounds in one or several screens. MolClass combines the datasets to generate a computational model for each screen submitted (Fig. 1B). These models are then applied to all molecules stored in MolClass to predict activity. MolClass currently provides either a composite of all molecular 2D chemical descriptors (2529 bit) or the user can independently choose 152 property descriptors, MACCS (166 bit), Substructure (306 bit), CDK extended (1024 bit) or PubChem (881 bit) fingerprints. As different machine-learning algorithms tend to generate slightly different likelihood values, a variety of algorithms are provided in MolClass including Random Forest, Naïve Bayes, SVM, KNN, Logistic Model Tree and J48. The user can apply one or several algorithms to any dataset of interest. Unbalanced datasets are boosted, to maximally double the size of the smaller part, using SMOTE (Nitesh *et al.* 2002) and further, if they exceed a ratio 1:5 of active versus inactive compounds, are adjusted using the WEKA under-sampling method. All models in MolClass are then applied to these molecules to generate activity fingerprints. For training and testing, MolClass uses 10-fold cross-validation.

*To whom correspondence should be addressed.

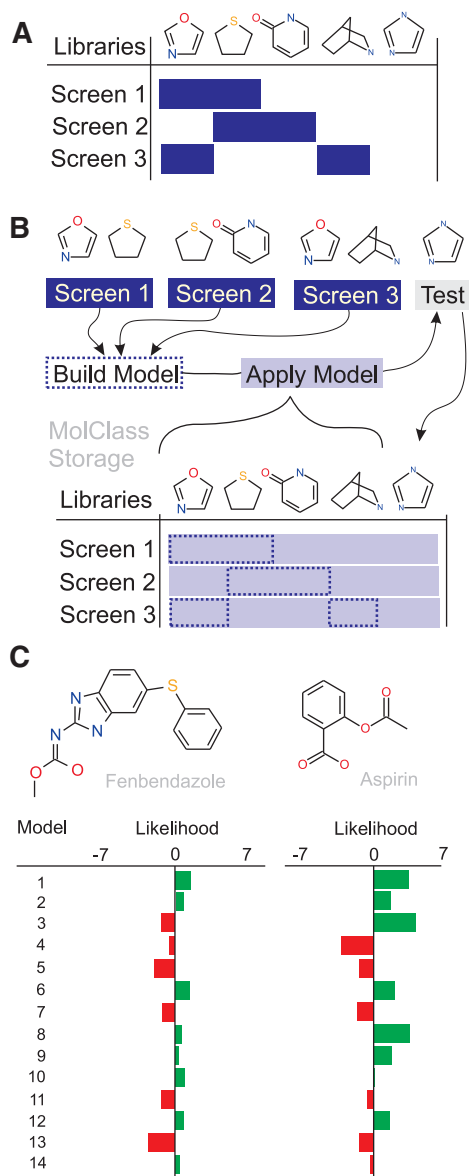


Fig. 1. MolClass features. (A) current state of data from public resources such as PubChem and ChemBank. (B) MolClass workflow from experimental data to activity likelihoods. (C) Likelihood scores for fenbendazole and aspirin in 14 different models: (1) neurosphere proliferation, +/-none (Diamandis *et al.* 2007); (2) Caco-2 permeation, +/- (Hou *et al.* 2004); (3) flucanazole synergizer, +/-none (Spitzer *et al.* 2011); (4) *Caenorhabditis elegans* drug bioaccumulation, none/+ (Burns *et al.* 2010); (5) Ames mutagenicity benchmark, none/+ (Hansen *et al.* 2009); (6) mutagenicity prediction, +/-none (Kazius *et al.* 2005); (7) blood-brain barrier penetration, +/- (Li *et al.* 2005); (8) PubChem AID 1828 +/-none; (9) PubChem AID 595 +/- (10) ChemBank 1000423 +/- (11) ChemBank 1001644 +/- (12) ChemBank 1000359 +/- (13) autofluorescence none/+ and (14) ChEMBL TargetID CHEMBL204 none/+. '+' activating, '-' inhibiting and 'none' no effect

The user can examine the model statistics, the likelihood scores for screens of interest and, as shown in Figure 1C, single molecule likelihood fingerprints for existing models. Finally, MolClass also

enables a substructure search using the JME Editor in the event a molecule of interest is not present in the database.

3 CONCLUSION

MolClass provides a comprehensive overview of compound activity in different screens. These profiles can reveal promiscuous activities across several screens, which may reflect undesirable off-target effects. For experimental datasets, the user can discover structure activity relationships because similar structures and activities will lead to specific likelihood patterns. As the data collection is expanded by users to different biological responses and assay formats, the classification power of the portal will increase, and thereby facilitate chemical systems biology.

4 IMPLEMENTATION

MolClass is implemented in Java and Perl using CDK (Steinbeck *et al.* 2003), WEKA (Hall *et al.* 2009) and moldb4 (Haider, 2010). The web interface and REST service are written in PHP5, Slim and PEAR and run on a Fedora Linux 8 server, as an Apache HTTP service. The data are stored in a MySQL 5.5 database running on a separate Fedora Linux 16 server.

ACKNOWLEDGEMENTS

We thank Lutz Fischer for support on improving MolClass Java code, Michaela Spitzer for comments on this article, Ryusuke Kimura for improvements of the web interface and anonymous reviewers for helpful suggestions.

Funding: European Research Council (233457-SCG to M.T.).

Conflict of Interest: none declared.

REFERENCES

- Burns,A.R. *et al.* (2010) A predictive model for drug bioaccumulation and bioactivity in *Caenorhabditis elegans*. *Nat. Chem. Biol.*, **6**, 549–557.
 Diamandis,P. *et al.* (2007) Chemical genetics reveals a complex functional ground state of neural stem cells. *Nat. Chem. Biol.*, **3**, 268–273.
 Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.
 Hall,M. *et al.* (2009) The WEKA data mining software: an update. *SIGKDD Explorations*, **11**, 10–18.
 Haider,N. (2010) Functionality pattern matching as an efficient complementary structure/reaction search tool: an open-source approach. *Molecules*, **15**, 5079–5092.
 Hansen,K. *et al.* (2009) Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.*, **49**, 9:2077–2081.
 Hopkins,L.A. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.
 Hou,T.J. *et al.* (2004) ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comput. Sci.*, **44**, 1585–1600.
 Kazius,J. *et al.* (2005) Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, **48**, 312–320.
 Li,H. *et al.* (2005) Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.*, **45**, 1376–1384.
 Nitesh,V.C. *et al.* (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.
 Spitzer,M. *et al.* (2011) Cross-species discovery of synergistic drug combinations that potentiate the antifungal flucanazole. *Mol. Syst. Biol.*, **7**, 499.
 Steinbeck,C. *et al.* (2003) The Chemistry Development Kit (CDK): an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.