Bioinformatics Advance Access published June 17, 2012

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

Jan Wildenhain[1,*], Nicholas FitzGerald[2] and Mike Tyers[1,3]

[1]Wellcome Trust Centre for Cell Biology and School of Biological Sciences, The University of Edinburgh, Edinburgh, UK, EH9 3JR.

[2]Department of Computer Science and Engineering, University of Washington, Seattle, US, 98122

[3]Institute for Research in Immunology and Cancer, Université de Montréal, Québec H3T 1J4, Canada

Associate Editor: Dr. Jonathan Wren

ABSTRACT

Summary: The MolClass toolkit and data portal generates 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 is 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 website. 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, md.tyers@umontreal.ca

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 *et al*. 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 *et al*. 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 (Figure 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 data sets 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 (Figure 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 (2529bit) or the user can independently choose 152 property descriptors, MACCS (166bit), Substructure (306bit), CDK extended (1024bit) or PubChem (881bit) 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

[*]To whom correspondence should be addressed.

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. 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 *et al.* 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 is 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 the manuscript, Ryusuke Kimura for improvements of the web interface and anonymous reviewers for helpful suggestions. This work was supported by the European Research Council grant 2007-223411 to MT.

REFERENCES

Burns, A.R., *et al.* (2010) A predictive Model for Drug Bioaccumulation and Bioactivity in Caenorhabditis elegans, *Nat Chem Biol.*, **6(7)**, 549-57.

Diamandis, P., *et al.* (2007) Chemical genetics reveals a complex functional ground state of neural stem cells, *Nat Chem Biol.*, **3(4)**, 268-73.

Dobson, C.M., *et al.* (2004) Chemical space and biology, *Nature*, **432**:824-828.

Hall, M., *et al.* (2009) The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, **11**, Issue 1

Haider, N., *et al.* (2010) Functionality Pattern Matchin as an Efficient Complemntary Structure/Reaction Search Tool: an Open-Source Approach, *Molecules*, 1**5**, 5079-5092

Hansen, K., *et al.* (2009) Benchmark data set for in silico Prediction of Ames Mutagenicity, *J Chem Inf Model*, **49**, 9:2077-81.

Hopkins, L.A., *et al.* (2008) Network pharmacology: the next paradigm in drug discovery, *Nat Chem Biol.*, **4**(11):682-90.

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**(5):1585-600.

Kazius, J., *et al.* (2005) Derivation and validation of toxicophores for mutagenicity prediction., *J Med Chem*, **48**(1):312-20

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**(5):1376-84.

Nitesh V.C., *et al.* (2002) SMOTE: Synthetic Minority Over-sampling TEchnique, *Journal of Artificial Intelligence Research*, **16**:321-357.

Spitzer, M., *et al.* (2011) Cross-Species Discovery of Syncretic Drug Combinations that Potentiate the Antifungal Flucanozole, *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**(**2**), 493-500

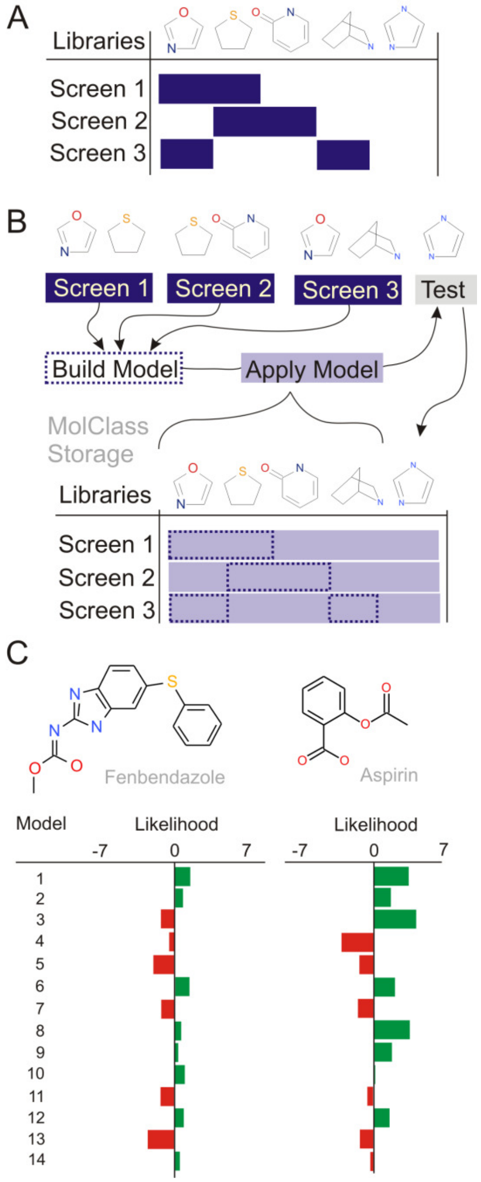

Figure 1. MolClass features **(A)** Current state of data from public resources such as Pubchem and Chembank. **(B)** MolClass Workflow from experimental data to activity likelihoods. **(C)** Likelihood scores for fenbendazole and aspirin in 14 different models: : 1 Neurosphere proliferation, +/none, Diamandis *et al.* 2 Caco-2 permeation, +/-, Hou *et al.* 3 Flucanozole Synergizer, +/none, Spitzer *et al.* 4 *C. elegans* Drug Bioaccumulation, none/+, Burns *et al.* 5 Ames mutagenicity Benchmark, none/+, Hansen *et al.* 6 Mutagenicity prediction, +/none, Kazius *et al.* 7 Blood-Brain Barrier penetration, +/-, Li *et al.* 8 PubChem AID 1828 +/none 9 PubChem AID 595 +/- 10 Chembank 1000423 +/- 11 Chembank 1001644 +/- 12 Chembank 1000359 +/- 13 Autofluoresence none/+ 14 ChEMBL TargetID CHEMBL204 none/+. '+' activating, '-' inhibiting, 'none' no effect.