Dataset: Recognizing mechanism of activity (MOA) of antitumor compounds from the NCI-60 cancer cell line biological activity screen

The NCI-DTP (Developmental Therapeutics Program) has tested thousands of small molecules for *in vitro* antiproliferative activity against 60 human cancer cell lines – the NCI-60 screen [1]. The mode of action for 100s of these compounds is characterized well enough so they can be assigned to a mechanistic class. This data can be used to develop models that recognize the MOA class of a compound from its pattern of differential activity across the 60 cell lines [2, 3]. Then, such models could be applied to a novel compound to find its MOA after having tested the compound for growth inhibition on the 60 cell lines.

In particular, the dataset contains 11,999/13,404/12,998 molecules (depending on the preprocessing, see below), 475/552/615 of which are assigned to exactly one of 12 possible MOA classes. For each molecule, its cytostatic activity against 60 cell lines is given, expressed as $-\log_{10}$ of the GI_{50} [also called IC_{50}] concentrations. For example, activity of 6.0 means the compound is active (inhibits growth by 50%) in the 10^{-6} M concentration (micromolar range). There are some missing values: overall 2.2%/6.3%/2.1%, and varying substantially across cell lines. We have excluded compounds that are either completely inactive or active against very few (<10) cell lines, and those with many missing values or experimental results removed in preprocessing (see Methods); the unfiltered NCI-60 dataset would have >47,000 compounds.

The possible MOA classes are: alkylating agent, DNA antimetabolite, nucleoside/nucleobase analog, cytoskeleton/mitotic agent, antineoplastic antibiotic, kinase targeting agent, membrane active agent, DNA intercalator, steroid compound, ion channel agent, topoisomerase I poison, or topoisomerase II poison. The MOA labels were filtered to retain only compounds where the GI_{50} patterns across cell lines were found to be consistent with other compounds in the same MOA class, using a Random Forest analysis (see Methods).

This specific dataset has been used to find the MOA of novel compounds by experimentally measuring cytostatic activity against a subset of NCI-60 cell lines, and applying a Random Forest classifier model on such measurements. This is described in: (i) Ester, Supek et al. 2012 *Inv New Drugs*. "Putative mechanisms of antitumor activity of cyano-substituted heteroaryles in HeLa cells", and (ii) Supek, Kralj et al. 2010 *Inv New Drugs*. "Atypical cytostatic mechanism of N-1-sulfonylcytosine derivatives determined by in vitro screening and computational analysis".

If you found this data useful, please cite Ester et al. (2012) doi:10.1007/s10637-010-9571-7 and an original NCI publication with MOA analyses (for instance, [2]).

Methods:

Preprocessing – handling defaults. This data set is replete with *default* GI_{50} values, meaning that the actual GI_{50} concentration falls outside of the experimentally tested range. Given that compounds are often tested in the concentration range of 10^{-8} M to 10^{-4} M, inactive compounds will have the default of $GI_{50} >= 10^{-4}$ M, and (less frequently) very active compounds will have the default $<= 10^{-8}$ M. As a complicating circumstance, compounds are sometimes not tested in this typical range 10^{-8} - 10^{-4} M, and moreover the same compound may be tested in multiple experiments which may have different ranges (and thus different default values). Data for some compounds was not provided in some cell lines.

The three ways of preprocessing depend on how these *defaults* are handled:

(1) ***Force typical range***: all experiments which do not range from 10^{-8} to 10^{-4} M are fully discarded; some compounds will thus be left out of the final data. The high activity ($GI_{50} \leq 10^{-8}$ M) and the low activity ($GI_{50} \geq 10^{-4}$) defaults are

kept as observed measurements, although the actual activity may be a lot higher/lower than the given value. Multiple experiments per compound are averaged over. Max. 3/60 missing values per compound are tolerated. At least 10/60 cell lines must have a non-default value or the compound is discarded.

(2) ***Force non-defaults***: all experiments are kept, regardless of tested range. All default values are discarded, and represented by missing values. This results in many missing values for the inactive (or rarely, very active) compounds. Multiple experiments for the same compound are averaged over. Max. 10/60 missing values per compound are tolerated (here this includes both the actual missing values, and the defaults) or the compound is discarded.

(3) ***Smart filter*** (RECOMMENDED): all experiments are kept regardless of range; the defaults are also kept and recorded as observations. If the same compound has multiple experiments and gives >1 default value, the most extreme default is kept. If some experiment records a default and others record non-defaults for the same compound, the default is completely discarded, and the non-defaults are averaged over. Max. 3/60 missing values per compound are tolerated. At least 10/60 cell lines must have a non-default value or the compound is discarded.

Methods (1) and (3) have little missing data but may be noisier as they include defaults; method (3) has higher coverage. Option (2) has the most reliable data but also more missing values.

This data is not normalized and the $-\log_{10} GI_{50}$ values are given as-is. However in the Ester *et al.* and Supek *et al.* papers above, we used the data after standardizing (scaling) each compound (row in table) to a mean value of 0 and standard deviation of 1 across the cell lines.

Data sources for GI_{50} values and MOA class labels. The growth inhibition data is the Dec-2010 version downloaded from the NCI-DTP site [4] and preprocessed as above. The putative mechanism-of-action (MOA) class labels were collected from the DTP website [5], a broader set derived from GI_{50} profile clustering analyses in [2,3] was kindly provided by the DTP (pers. comm.), and was further edited by removing and/or merging smaller classes, by manually curating further compounds from the literature, and by assigning each compound to at most one class. Importantly, these putative MOA labels were further filtered to retain only the compounds where the pattern of GI_{50} scores across cell lines was consistent within one MOA class. In particular, the Random Forest was used to classify the MOA classes based on their log GI_{50} values (dataset: no GI_{50} defaults allowed, max. 30/60 missing values tolerated per compound; GI_{50} scores of each compound were standardized to a mean=0 and sd=1.0 across the cell lines; 938 compounds of known MOA meet these criteria). The Random Forest crossvalidation predictions (by out-of-bag method) were examined and all the compounds marked as classification errors were discarded. In other words, if a compounds' differential GI_{50} pattern did not agree with other compounds sharing its putative MOA class, this MOA assignment was removed, leaving 714 compounds with known MOA at this step. The supplied SMILES molecular structures for the subset of compounds with known MOA are also from the DTP website, where the SMILES for the complete set of compounds in the NCI-60 screen can be found.

References.

[1] Shoemaker, R. H. The NCI60 Human Tumour Cell line Anticancer Drug Screen. Nat Rev Cancer, 6: 813-823, 2006.

[2] Rabow AA, Shoemaker RH, Sausville EA, Covell DG (2002) Mining the National Cancer Institute's tumor-screening database: identification of compounds with similar cellular activities. J Med Chem 45:818‑840

[3] Huang R, Wallqvist A, Covell DG (2006). Assessment of *in vitro* and *in vivo* activities in the National Cancer Institute's anticancer screen with respect to chemical structure, target specificity, and mechanism of action. J Med Chem. 2006 Mar 23;49(6):1964-79.

[4] http://dtp.nci.nih.gov/webdata.html

[5] http://www.dtp.nci.nih.gov/docs/cancer/searches/standard_mechanism.html