Gene expression

**Cell-specific predictionof drug-induced gene expression profiles**

Rachel Hodos1,2, Hao Chih Lee1, Ping Zhang3, Qiaonan Duan1, Zichen Wang1, Neil R. Clark1, Avi Ma’ayan1, Fei Wang3,4, Brian Kidd1, David Sontag2, Jianying Hu3,Joel Dudley1\*

1Icahn School of Medicine at Mt. Sinai, New York, NY, 10029, 2Courant Institute of Mathematical Sciences, New York University, New York, NY, 10012, 3IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598, 4Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University, New York, NY, 10065

\*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

**Motivation:** Gene expression profiling of *in vitro* drug perturbations is useful for many biomedical discovery applications including drug repurposing and elucidation of drug mechanisms. However, limited data availability across cell types has hindered our capacity to leverage or explore the cell-specificity of these perturbations. While recent efforts have generated a large number of drug perturbation profiles across a variety of cell types*,* many gaps remain in this combinatorial drug/cell space. Hence, we asked whether it is possible to fill these gaps by predicting cell-specific drug perturbation profiles using available expression data from related conditions--i.e. from other drugs and cell types.

**Results:** We developed a computational framework that first arranges existing profiles into a three-dimensional array (or tensor) indexed by drugs, genes, and cell types, and then uses either local (nearest neighbors) or global (tensor completion) information to predict unmeasured profiles. We evaluated the results using cross-validation and drug repurposing metrics, as well as comparison to external drug profiles. The two methods had generally comparable performance, with some distinctions depending on the input data dimensions. Predictions achieved correlations of 0.54 between true and predicted expression values, and in many cases had stronger correspondence with independently measured drug profiles than their measured counterparts. Further, the predicted profiles maintain biologically and therapeutically relevant properties such as accurate differentially expressed genes (AUC 0.81), gene correlation structure, and connection to disease signatures for drug repurposing.

**Contact:** joel.dudley@mssm.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

**Introduction**

Genome-wide expression profiling of *in vitro* drug perturbations has proven to be useful for many aspects of drug discovery and development 1. Applications include elucidation of drug mechanisms 2, lead identification 3, and drug repurposing 4, 5. Despite this success, the capacity to leverage cell-specific responses has been hindered by limited data availability across cell types 6, 7. To address this limitation, the Library of Integrated Cellular Signatures (LINCS) program 8, 9 has greatly expanded the publicly available data to nearly one million profiles characterizing thousands of drugs exposed to dozens of cell types. However, this combinatorial space of drugs and cell types is vast, and many gaps remain in this space (see Figure 7A). These gaps present difficulties both for large-scale analysis as well as for making cell-resolved comparisons, e.g. between two drugs or between drug and disease. Therefore, we asked whether it is possible to leverage existing expression profiles to predict the remaining unmeasured profiles.

Expression responses to drug exposure are often highly cell-specific, e.g. due to differences in expression of drug targets. Indeed, we observe a high degree of cell-specificity for many drugs in the LINCS data (see Figure 1). The utility of such cell-specific gene expression has previously been demonstrated for a variety of applications. For example, 10 showed that using transcriptional similarity to predict drug-target interactions works better when comparing drug profiles in the same cell line. Another study 6 identified cell-specific expression responses with potential therapeutic relevance, e.g. anti-estrogen drugs inducing a common response that was specific to breast cancer cells.

Prior studies have described methods to predict expression profiles using outside information. For example, (Gamazon, et al., 2015) predict tissue-specific expression profiles from genetic variants, but are limited to basal-state expression. Conversely, (Lagunin, et al., 2013) predict drug-induced expression responses from a drug’s chemical structure, but are agnostic to cell type. There are also many techniques to impute missing entries of a gene expression matrix, generally using either local (e.g. nearest neighbors) or global (e.g. low-rank matrix approximation) information (Brock, et al., 2008; Troyanskaya, et al., 2001). However, most of these methods are not directly applicable to our setting, as they rely on having at least some measurements available in the target experimental setting.

Here, we draw inspiration from this prior work to solve a new problem: predicting *entire* expression profiles for cell-specific drug perturbations. Our two approaches are complementary in their use of local vs. global information. The “local” algorithm, *Drug Neighbor Profile Prediction* (DNPP) is inspired by K-nearest neighbors but adapted to this *de novo* prediction setting. The “global” algorithm, the *Fast, Low-Rank Tensor Completion* (FaLRTC) algorithm (Liu, et al., 2013) fills in the missing entries of a tensor (see Figure 2A) using the observed entries. The underlying assumption behind this and many other tensor completion algorithms is that the data are low-rank, i.e. some small set of underlying factors (e.g. drug targets) explain most of the variation in the data.

We evaluate our methods, along with two simple baselines, using several approaches. We use cross-validation (CV) to measure performance readouts including correlation of predicted and true expression values, accuracy of differentially expressed genes (DEGs), preservation of gene correlation structure, and preservation of cell-specificity. We analyze accuracy per drug, per gene and per cell type. We also study the dependence of accuracy on the amount of input data. In addition to these internal, CV-based evaluations, we perform external evaluation against an independently measured drug profile database. Finally, we perform an application-oriented evaluation of method performance by assessing preservation of connection to disease signatures in a drug repurposing framework.

**The LINCS drug expression data** (herein, the “L1000 data”) is measured on a targeted expression profiling platform called L1000. The platform measures the expression of 978 “landmark” genes (roughly 1000, hence the name), selected to be maximally predictive of the other genes while being widely expressed across many cell and tissue types.

**Differential expression** computed from the level 3 L1000 data were downloaded from <amp.pharm.mssm.edu/public/L1000CDS_download>. The dataset was generated using the Characteristic Direction (CD) method (Clark, et al., 2014) and is validated and described more fully in (Duan, et al., 2016). Briefly, a CD was calculated for each replicate using linear discriminate analysis, to find the direction in gene space that best separates cases and controls. Replicates were averaged and normalized to unit length. Average cosine distance (ACD), i.e. the mean pairwise cosine distances between an experiment’s CD replicates, was used to estimate significance. The null distribution of the ACDs was calculated per batch using random sampling (*n* = 10,000) of replicates in the same batch. A *p*-value for each profile (ACD *p*-value) was computed by comparing its ACD to the null.

**Tensor construction:** The 201,484 CD profiles (20,413 drugs, 72 cell types) were filtered to 34,716 profiles (6928 drugs, 72 cell types) with ACD *p* ≤ 0.1 in order to remove the most unreliable data. Drugs and cell types with < 3 remaining experiments were removed, as well as duplicate drug id’s corresponding to the same drug, for a final count of 25,672 profiles (2130 drugs, 71 cells, 12.7% of all CDs). Profiles were averaged across all available concentration and time points, renormalized, and then arranged into a tensor (see Figure 2A). Of the 151,230 possible drug-cell pairs, the tensor contained 15,855, corresponding to 10.5% observation density. Herein we refer to this as *Tlarge* or ‘the large tensor.’ Three subtensors were also defined (see Figure 7A) by varying the drug and cell-type dimensions while keeping all genes, to define: *Tmany‑drugs* (2000 drugs x 15 cell types, 39.5% observation density); *Tmany‑cells* (300 x 50, 38.5%) and *Tsmall* (300 x 15, 71.4%), also called ‘the small tensor.’ Note that *Tsmall* is the intersection of *Tmany‑drugs* and *Tmany‑cells*. The tensor element *Td,g,c* is the *g*th coordinate of the CD vector for drug *d* in cell *c.* All values lie in the range [‑1,1] after normalization, where a positive [negative] value corresponds to up- [down-] regulation. The 10 cell lines with the most data in *Tlarge* are listed in Table 1, along with the corresponding tissue of origin and number of profiles (i.e. drugs) present. Most of the 71 cell lines are cancer cell lines, and represent a range of human tissues including skin, lung, brain, kidney, and prostate. A list of all drugs, genes and cell lines in this tensor, along with some additional information (including prediction accuracy for each entity, see Section 3.6) is provided in Table S1.

**REFERENCES**

1. Qu XA, Rajpal DK. Applications of Connectivity Map in drug discovery and development*.* *Drug discovery today* 2012, 17:1289-1298.

2. Wei G, Twomey D, Lamb J, Schlis K, Agarwal J, Stam RW, Opferman JT, Sallan SE, den Boer ML, Pieters R. Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance*.* *Cancer cell* 2006, 10:331-342.

3. Hassane DC, Guzman ML, Corbett C, Li X, Abboud R, Young F, Liesveld JL, Carroll M, Jordan CT. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data*.* *Blood* 2008, 111:5654-5662.

4. Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning*.* *Briefings in bioinformatics* 2011:bbr013.

5. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles*.* *PLoS One* 2009, 4:e6536.

6. Khan SA, Virtanen S, Kallioniemi OP, Wennerberg K, Poso A, Kaski S. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis*.* *Bioinformatics* 2014, 30:i497-i504.

7. Parkkinen JA, Kaski S. Probabilistic drug connectivity mapping*.* *BMC bioinformatics* 2014, 15:113.

8. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, Rouillard AD, Tan CM, Chen EY, Golub TR. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures*.* *Nucleic acids research* 2014:gku476.

9. NIH BI. NIH LINCS Program. Available at: <http://www.lincsproject.org/>.

10. Iwata M, Sawada R, Iwata H, Kotera M, Yamanishi Y. Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics*.* *Scientific Reports* 2017, 7:40164.