Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action

Maris Lapins[1] and Ola Spjuth[1]

[1] Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

Abstract

Profiling drug leads by means of in silico and in vitro assays as well as omics is widely used in drug discovery for safety and efficacy predictions. In this study, we evaluate the performances of machine learning models trained on data from gene expression and phenotypic profiling assays, with in vitro assays by means of chemical structure descriptors, for prediction of various drug mechanisms of action and target proteins. Models for several hundred mechanism(s) of actions and protein targets were trained using data on 1484 compounds characterized in both gene expression using L1000 profiles, and phenotypic profiling with cell painting assays. The results indicate that the accuracy of the three profiling technologies varies for different endpoints, and indicate a clear potential synergistic effect if these methods are combined. We also study the effect of predictive accuracy of data from different cell lines for L1000 profiles, showing that the choice of cell line has a non-negligible effect on the predictive accuracy. The results strengthens the idea of integrated approaches for predicting drug targets and mechanisms of action in preclinical drug discovery.

Introduction

Over the past decade, methods have been developed to systematically determine cellular effects of chemical compounds with the aid to improve fields such as drug screening and safety profiling [1, 2, 3]. Important objectives include to predict off-target effects and adverse drug reactions (ADRs), but also to offer insights into compound's Mode-of-Action (MoA) and the establishment of Adverse Outcome Pathways (AOPs).

Pharmaceutical profiling using ligand binding or enzyme assays is the most widely used in vitro methodology, and it is widely implemented in drug discovery safety platforms. Profiling using gene expression is relatively recent, and pioneering work includes Connectivity Map [4] that has been widely used and built upon [5, 6].

L1000 is a high throughput and low cost gene expression profiling method, based on representation of transcriptome by 978 “landmark genes”. Recently, datasets with L1000 profiles were made available in Broad LINCS L1000 Connectivity Map project, including profiles for a total of 20K small molecule compounds, of which over 2K compounds were studied systematically in nine human cancer cell lines [6, 7].

Multiparametric high-content imaging has also proven to be a highly useful and successful technique for understanding biological activity in response to chemical and genetic perturbations. The Broad Bioimage Benchmark Collection (BBBC) is an important publicly available collection of microscopy images. Some of the largest image sets obtained by Cell Painting assay comprise osteosarcoma cells treated by 1.6K known bioactive compounds [8] and by 30K compounds, most of which being derived from diversity-oriented synthesis [9].

It is hypothesized that chemical compounds with a similar mechanism of action (MoA), which act upon the same signaling pathways, will produce comparable phenotypes, and that analysis of phenotyping profiling data can predict compound mechanism of action [10]. Successful prediction examples include study by Ljosa et al. [11] where 37 compounds are classified to 12 MoA’s with 94% prediction accuracy and study by Warchal et al. where 24 compounds are classified to 8 MoA’s with over 80% accuracy in several cell lines [8]. On a large scale, predicting of results of particular biological assays on the basis of phenotyping profiling data have been recently undertaken [12, 13]. In particular, in study by Simm et al. information extracted from microscopy-based screen for glucocorticoid receptor translocation was able to predict assay-specific biological activity in two ongoing drug discovery projects, leading to a tremendous 60-fold and 250-fold increase of hit rates.

For transcriptomic data, models are reported by Aliper et al. [14] where several hundred compounds selected from Broad LINCS database are linked to 12 therapeutic use categories in breast cancer (MCF7), prostate cancer (PC3), and lung cancer cells (A549).

The aim of the current study was to compare the performances of descriptors derived from gene expression and phenotypic profiling assays with the performance of chemical structure based descriptors for prediction of various drug mechanisms of action and target proteins. To this end, models for several hundred mechanism(s) of actions and/or protein target(s) (MoA/Ts) were created using data for 1484 compounds characterized in both gene expression and phenotypic profiling assays.

As L1000 gene expression profiles have been collected systematically in several cell lines, we also aimed to investigate cell-context specificity of transcriptomic data for predicting MoA/Ts.

In phenotyping profiling, each compound is typically tested in quadruplicates or octaplicates on different plates and thus four or eight profiles per compound are obtained. The overall profile thus depends on the way the data are aggregated. In this study we therefore also investigated effects of data pre-processing on the prediction accuracy.

Methods

Datasets

Gene expression (Connectivity Map)

The Connectivity Map (CMap) dataset built using L1000 high-throughput gene-expression assay was downloaded from GEO (ascension GEO: GSE92742). The dataset comprises transcriptional responses (expression of 978 landmark genes) to perturbations of various cells by 19,811 small molecule compounds. 2,429 of the compounds are tested systematically across nine human cancer cell lines.

Phenotypic profiling (Cell Painting)

Dataset of images and morphological profiles of 30,616 small molecule treatments obtained by Cell Painting assay was downloaded from http://gigadb.org/dataset/100351. In this assay, human U2OS (human osteosarcoma) cells are stained for eight major organelles and sub-compartments, using a mixture of six fluorescent dyes. From five channel microscopy images, 1783 morphological features are generated by CellProfiler software [15].

Annotation of compounds with protein targets and/or mechanism of action

We used Touchstone data base (https://clue.io/touchstone) [6] and Drug Repurposing Hub (https://clue.io/repurposing) [16] to associate compounds to their mechanism(s) of actions and/or protein target(s) (MoA/Ts). From annotations to individual targets, we also derived labels for protein kinase groups.

For Phenotypic profiling dataset, we obtained annotations for 1759 compounds, where 257 MoA/Ts were shared by at least five compounds. In CMap dataset, the three cell lines with the highest number of annotated compounds were MCF7 (breast cancer, 2801 annotated compounds, 444 MoA/Ts shared by at least five compounds), PC3 (prostate cancer, 2775 annotated compounds, 435 MoA/Ts), and A549 (lung cancer, 2319 annotated compounds, 380 MoA/Ts). The intersection of Phenotypic profiling dataset and the largest of CMap datasets (MCF7) contained 1484 compounds and 234 MoA/Ts.

Data pre-processing

In Phenotypic profiling dataset, most of the compounds have been applied to cells eight times on different plates, thus giving eight sets of morphological features for each compound. In data

pre-processing, we first centered and (optionally) normalized the features on plate-to-plate basis, by subtracting the mean value and (optionally) dividing by standard deviation for the control samples on this plate. Thereafter we calculated the mean or the median values of each feature from the eight sets, and used them as descriptors for the compounds. Some of 1783 features were invariant in the present dataset, and were removed before the modelling.

Random Forest

Random Forest (RF) is a classifier that consists of multiple decision trees. A decision tree is made of nodes and branches. At each node the dataset is split based on the value of some attribute that is selected so that the instances of different classes are predominantly moved to different branches. Classification starts at the root node and is performed by passing the instances along the tree to leaf nodes. To introduce diversity between the trees of a random forest, a subset of all attributes is randomly selected to take decisions at each node of each tree. The class probability of an instance is estimated considering results of all trees. We here developed RF models with 500 trees using the randomForest package of R. Thus, for a test set instance the class probability was one of 500 numerical values in the range from 0 to 1.

Evaluation of modeling performance

For every MoA/T, 25 RF models were created, assigning 80% of compounds to the training set and 20% of compounds to the prediction set. The predictions from all models were aggregated to calculate Receiver Operating Characteristic (ROC) curve, which is plotted as the true positive rate versus the false positive rate at various discrimination threshold values. The area under the ROC curve (AUC) is a measure of the discriminatory power of a classifier, which is insensitive to class distributions and the costs of misclassifications; AUC = 1 indicates perfect classification, while AUC = 0.5 means that the classifier does not perform better than random guessing.

Results and Discussion

1. Models for CMap datasets in three cell lines

In CMap dataset, the three cell lines with the highest number of annotated compounds were MCF7 (breast cancer, 2801 annotated compounds, PC3 (prostate cancer, 2775 annotated compounds), andA549 (lung cancer, 2319 annotated compounds).

We created Random Forest models for mechanisms of action and targets (MoA/Ts) shared by at least five compounds, which gave 444, 435, and 380 models for MCF7, PC3, and A549,

respectively. For 20 MoA/Ts models with the area under the ROC curve (AUC) > 0.90 were obtained, for 55 MoA/Ts the AUC exceeded 0.80, and for 140 MoA/Ts the AUC exceeded 0.70. The results for the best-predicted MoA/Ts are presented graphically in Figure 1 (for full results with number of active compounds in each model, AUC, and confidence intervals see Supplementary Table 1.)

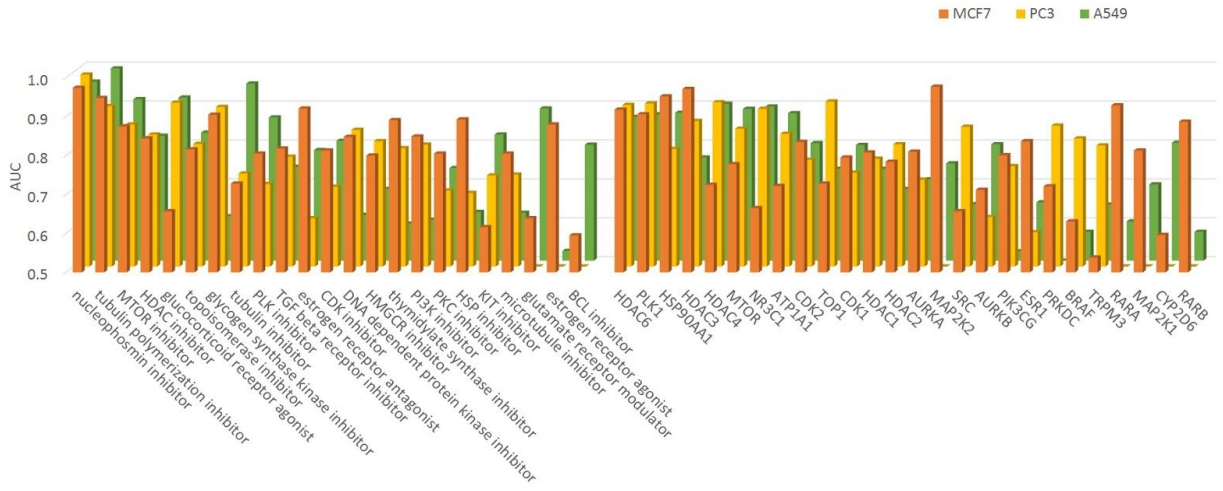

Figure 1. The areas under the ROC curve (AUC) for predicting MoA/Ts in three cell lines. AUC is a measure of the discriminatory power of a model. AUC=1 indicates perfect predictions, i.e. complete separation of all class members from all non-members, whereas AUC=0.5 indicates predictions not better than random.

In the presentation of CMap dataset, the authors noted that only 15% of compounds produced highly similar transcriptional profiles across the entire panel of cell-lines suggesting that transcriptional response is cell dependent [7]. For instance, it was found that glucocorticoid receptor antagonists shared similar profiles only in cell lines where the glucocorticoid receptor NR3C1 was highly expressed (i.e. A549, but not PC3 and MCF7). Our results confirm this finding for glucocorticoid receptor agonists, where the models for A549 and PC3 cell lines show much better predictive performance than the model for MCF7. Similarly, for glycogen synthase kinase inhibitors good models are obtained in MCF7 and PC3, but not in A549 cell line, but for glutamate receptor modulators only in MCF7, and for estrogen receptor antagonists and agonists only in A549.

An overall comparisons of the models does not reveal, however, large differences between the cell lines, the average AUC for top-50 models in MCF7 being 0.85 and in the two other cell lines 0.82. An overview of results for the broadest drug classes indicates that gene expression data is not suited for modeling of GPCR-targeted drugs (such as agonists and antagonists of dopamine, histamine, serotonin, and acetylcholine receptors). For these mechanisms of action, the models show AUC around 0.50, i.e., they do not perform better than random guesses. In

contrast, an overall model for kinase inhibitors (that constitute about 10% of all dataset compounds) possesses predictive performance of AUC = 0.70 in MCF7 cell line and 0.71 in A549.

2. Models for CMap/Cell Painting dataset

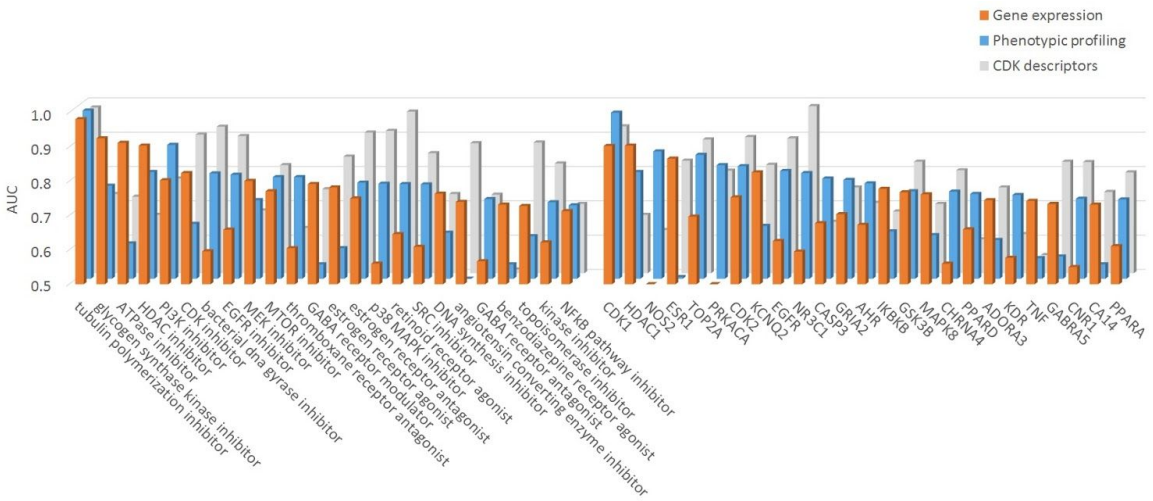

Figure 2. The areas under the ROC curve (AUC) for predicting MoA/Ts for models based on gene expression data, phenotypic profiling data, and structural (CDK) description of chemical compounds.

In the next step of the study we created models for a set of 1484 compounds that have been characterized in both gene expression and phenotypic profiling assays. For the sake of comparisons, we also created models using structural descriptors of molecules, calculated by Chemistry Development Kit package of R (rcdk). These descriptors include a variety of topological, geometrical, charge based and constitutional descriptors [17].

The results for MoA/Ts where AUC for either gene expression or phenotypic profiling based model exceeded 0.70 are presented graphically in Figure 2; results for all 234 MoA/Ts are given in Supplementary Table 2. In many cases, gene expression or phenotypic profiling models show comparable predictive performance. For some of the targets, however, only one of the two descriptions has produced a predictive model.

Similarly as with gene expression data, morphological profiling data has not given any predictive models for agonists/antagonists of most GPCR classes. This is in contrast to models for inhibitors of several protein kinases and protein kinase groups (such as non-receptor tyrosine

kinases with AUC = 0.71) and for agonists/antagonists of nuclear receptors (e.g. estrogen and retinoid receptors with AUC > 0.75).

Our negative results for GPCRs are in agreement with findings of Rohban et. al. [17] who estimated similarities of morphological profiles of pairs of compounds sharing the same MoA. For GPCR agonists/antagonists it was found that a very low fraction of the top most-similar profiles were profiles of compounds with the same MoA. Thus, for the four largest groups of compounds in the dataset, agonists and antagonists of dopamine and serotonin receptor, only 0 - 1% of top most-similar profiles belonged to another member of this group. (This can be compared to 5% for SRC inhibitors, where we got a predictive model with AUC = 0.78, 2% for tubulin EGFR inhibitor, where our model showed AUC = 0.80, and 96% for tubulin polymerization inhibitor, where our model showed AUC = 0.99).

In fact, for a multitude of MoA/Ts, the drug effect need not lead to profound morphological or transcriptional changes of cells. In profiling of 1600 known bioactive compounds by Cell Painting assay, Gustafsdottir et al. observed that only 13% of them could be deemed active, i.e. their profiles could be distinguished from the natural variation of profiles of untreated cells.

Another aspect that could be considered in phenotyping profiling is differences in pharmacokinetic/pharmacodynamic properties of chemical compounds. Because of these differences, imaging at one fixed data point may be suboptimal compared to temporal monitoring to observe maximum changes of cell morphology.

3. Models for Cell Painting dataset with different data pre-processing methods

We compared two pre-processing approaches: 1) centering of CellProfiler derived features on plate-to-plate basis by subtracting the mean value for the control samples on this plate and 2) centering and normalization by subtracting the mean and dividing by standard deviation for the control samples. In the latter case, use of some features was problematic because the values for control samples were invariant for part of the plates.

Thereafter we described the compounds by either the mean values or the median values from the eight feature sets (in the latter case, the three “weakest” and the three “strongest” changes in cell morphology are not considered).

Thus, four models were were created for each of 234 MoA/Ts. Overall, the results are very similar for most of MoA/Ts, the standard deviation calculated from the four AUC values being below 0.05, thus confirming reliability of the models. However, discrepancies can be observed for some MoA/Ts where the number of active compounds is low (see Supplementary Table 3).

It should be noted that calculation of CellProfiler features is not mandatory for analysis of cell imaging data. Use of raw images as inputs to pre-trained convolutional neural networks has in fact shown to give better results in some studies [13, 19].

References

1. Applications in image-based profiling of perturbations. Caicedo JC, Singh S, Carpenter AE. Curr Opin Biotechnol. 2016 Jun;39:134-142. doi: 10.1016/j.copbio.2016.04.003. Epub 2016 Apr 17. Review. PMID: 27089218
2. Mining data and metadata from the gene expression omnibus. Wang Z, Lachmann A, Ma'ayan A. Biophys Rev. 2019 Feb;11(1):103-110. doi: 10.1007/s12551-018-0490-8. Epub 2018 Dec 29. Review. PMID: 30594974
3. Quantitative high content imaging of cellular adaptive stress response pathways in toxicity for chemical safety assessment. Wink S, Hiemstra S, Huppelschoten S, Danen E, Niemeijer M, Hendriks G, Vrieling H, Herpers B, van de Water B. Chem Res Toxicol. 2014 Mar 17;27(3):338-55. doi: 10.1021/tx4004038. Epub 2014 Feb 5. Review. PMID: 24450961
4. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. Science. 2006 Sep 29;313(5795):1929-35. PMID: 17008526
5. A review of connectivity map and computational approaches in pharmacogenomics. Musa A, Ghoraie LS, Zhang SD, Glazko G, Yli-Harja O, Dehmer M, Haibe-Kains B, Emmert-Streib F. Brief Bioinform. 2018 May 1;19(3):506-523. doi: 10.1093/bib/bbw112. Erratum in: Brief Bioinform. 2017 Sep 1;18(5):903. PMID: 28069634
6. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, Rouillard AD, Tan CM, Chen EY, Golub TR, Sorger PK, Subramanian A, Ma'ayan A. Nucleic Acids Res. 2014 Jul;42(Web Server issue):W449-60. doi: 10.1093/nar/gku476. Epub 2014 Jun 6. PMID: 24906883
7. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.. Subramanian A, Narayan R, Corsello SM, Peck DD et al. Cell 2017 Nov 30;171(6):1437-1452.e17. PMID: 29195078
8. Multiplex cytological profiling assay to measure diverse cellular states. Gustafsdottir SM, Ljosa V, Sokolnicki KL, Anthony Wilson J, Walpita D, Kemp MM, Petri Seiler K, Carrel HA, Golub TR, Schreiber SL, Clemons PA, Carpenter AE, Shamji AF. PLoS One. 2013 Dec 2;8(12):e80999. doi: 10.1371/journal.pone.0080999. eCollection 2013. PMID: 24312513
9. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. Bray MA, Gustafsdottir SM, Rohban MH, Singh S, Ljosa V, Sokolnicki KL, Bittker JA, Bodycombe NE, Dancík V, Hasaka TP, Hon CS, Kemp MM, Li K, Walpita D, Wawer MJ, Golub TR, Schreiber SL, Clemons PA, Shamji AF, Carpenter AE.Gigascience. 2017 Dec 1;6(12):1-5. doi: 10.1093/gigascience/giw014. PMID: 28327978
10. Evaluation of Machine Learning Classifiers to Predict Compound Mechanism of Action When Transferred across Distinct Cell Lines. Warchal SJ, Dawson JC, Carragher NO. SLAS Discov. 2019 Mar;24(3):224-233. doi: 10.1177/2472555218820805. Epub 2019 Jan 29. PMID: 30694704
11. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. Ljosa V, Caie PD, Ter Horst R, Sokolnicki KL, Jenkins EL, Daya S, Roberts ME, Jones TR, Singh S, Genovesio A, Clemons PA, Carragher NO, Carpenter AE. J Biomol Screen. 2013 Dec;18(10):1321-9. doi: 10.1177/1087057113503553. Epub 2013 Sep 17. PMID: 24045582

12. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Simm J, Klambauer G, Arany A, Steijaert M, Wegner JK, Gustin E, Chupakhin V, Chong YT, Vialard J, Buijnsters P, Velter I, Vapirev A, Singh S, Carpenter AE, Wuyts R, Hochreiter S, Moreau Y, Ceulemans H. Cell Chem Biol. 2018 May 17;25(5):611-618.e3. doi: 10.1016/j.chembiol.2018.01.015. Epub 2018 Mar 1. PMID: 29503208
13. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. Hofmarcher M, Rumetshofer E, Clevert DA, Hochreiter S, Klambauer G. J Chem Inf Model. 2019 Mar 6. doi: 10.1021/acs.jcim.8b00670. [Epub ahead of print] PMID: 30840449
14. Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. Aliper A, Plis S, Artemov A, Ulloa A, Mamoshina P, Zhavoronkov A. Mol Pharm. 2016 Jul 5;13(7):2524-30. doi: 10.1021/acs.molpharmaceut.6b00248. Epub 2016 Jun 8. PMID: 27200455
15. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. Genome Biol. 2006;7(10):R100. Epub 2006 Oct 31. PMID: 17076895
16. The Drug Repurposing Hub: a next-generation drug library and information resource. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, Johnston SE, Vrcic A, Wong B, Khan M, Asiedu J, Narayan R, Mader CC, Subramanian A, Golub TR. Nat Med. 2017 Apr 7;23(4):405-408. doi: 10.1038/nm.4306. No abstract available. PMID: 28388612
17. Chemical Informatics Functionality in R. Guha, R. Journal of Statistical Software 2007;6(18)
18. Capturing single-cell heterogeneity via data fusion improves image-based profiling. Mohammad Hossein Rohban, Shantanu Singh, E Carpenter. bioRxiv preprint, May 22, 2018. doi: https://doi.org/10.1101/328542
19. Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes. Kensert A, Harrison PJ, Spjuth O. SLAS Discov. 2019 Jan 14:2472555218818756. doi: 10.1177/2472555218818756. [Epub ahead of print] PMID: 30641024