Prediction of Compound Cytotoxicity Based on Compound Structures and Cell Line Molecular Characteristics

Takumi Nakano and J.B. Brown*

Kyoto University Graduate School of Medicine, Laboratory of Molecular Biosciences, Life Science Informatics Research Unit Kyoto, Sakyo, Yoshida, 606-8501 Japan, +81-75-753-9515

(Received December 13, 2019; Accepted January 14, 2020)

In parallel to developments in Next-Generation Sequencing for cancer patient therapy decision making, personalized approaches to chemotherapy selection are also becoming desired. In an ideal situation, an individual’s genomic, transcriptomic, and tumor-specific in-vitro response to chemical perturbation would be combined, and the US National Cancer Institute NCI-60 project has systematically screened a large chemical library against a variety of cell lines from various tumor types. Therefore, chemoinformatics approaches to make effective use of this data and identify the chemical and biological factors are of value. In this work, we investigate the impact of both chemical and biological descriptions of tumor response to chemical inhibition, and assess how well modeling approaches can predict tumor inhibition response on external datasets. We find that external datasets in both the classification and regression problems are reasonably well addressed, with the impact of chemical description outweighing the contribution from transcriptome or genome descriptions of tumors.

Key Words: Cytotoxicity, Chemoinformatics, Machine learning, Random forest, Molecular representation

1 Introduction

Malignant neoplasms are one of the leading causes of death in developed countries. While many compounds have been developed as anti-cancer drugs, their effects can vary significantly from cell line to cell line.

The NCI-60 cell line project [1], which cultures 60 different patient-derived cancer cell lines and their response to chemical perturbation, has been serving as a reference to measure compounds’ anticancer properties, which has then provided insights for the development of new therapies, as well as for the identification of proteins that play critical roles in protein signaling networks.
modulated by anticancer compounds [2]. Virtual screening based on this public resource has the potential to complement high-throughput experiments which are financially inefficient. Thus, it is desirable to systematically extract patterns from these cell line-inhibitor-response datasets, and the result could be used to either make a prediction on the effect of existing compounds against new cell lines or also to filter a set of candidate compounds for improved enrichment factors during in-vitro testing.

We applied a machine learning method to build a prediction model for chemical cellular toxicity, both for a single compound and a compound combination, where compounds were described in silico based on structural information and cell line molecular characteristics. The objectives include building more accurate prediction models by adding molecular information, evaluating the model’s applicability to external compounds or cell lines, as well as assessing how much impact each gene’s expression level influences the selection of genes that can potentially have corresponding biological meanings.

2 Methods

2.1 Cell line and anticancer response data

GI<sub>50</sub> endpoint tables and compound structures were downloaded from the NCI public dataset website. A compound’s GI<sub>50</sub> endpoint value was classified into active or inactive using a threshold of 1µM and 10 µM, respectively. The original GI<sub>50</sub> dataset had 3,054,004 data points. The compound library used is provided in standard SDF format by the NCI-60 project’s homepage.

First, replicate entries were reduced to single entries having averaged endpoint values. We also removed cell lines for which molecular information were unavailable. After these filters, approximately 2.5 million data points remained.

The related ALMANAC [4] project has curated the phenotypical response of pairwise drug combinations on NCI-60 cell lines. In short, this score is defined as the gain of efficacy from a compound combination compared to the expected efficacy that is based on two single-agent endpoint measurements, summed over experiments on multiple concentration settings.

We also retrieved the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, which comprises over 1000 human tumor cell lines and more than 200 drugs [3]. Activities in this dataset are recorded as IC<sub>50</sub> values.

Compound bioactivity tables from the GDSC and ALMANAC datasets were all retrieved from Cell Miner-CDB (URL=https://discover.nci.nih.gov/cellminercedb/). While there are two versions of the GDSC dataset based on different assay platforms, Cell Miner-CDB provides these as a merged resource.

2.2 Molecular descriptors

The MACCS organic chemistry-based substructural fingerprint was used as a compound descriptor. We primarily focused on MACCS descriptors as they can be interpreted as substructures in the compound and can provide compound structural insights with respect to cytotoxicity.

We also tested 99 DRAGON descriptors corresponding to polarizable surface area (PSA), estimated LogP (water-octanol partition coefficient), and other physicochemical properties, specifically the Molecular Properties, Constitutional, and Ring Descriptor blocks. While PSA and LogP represent key physicochemical parameters, they cannot be easily reverse-mapped back to structures. In some cases, these descriptors could not be computed for a compound; we adopted a policy to discard the compound and its bioactivity annotations from the training or prediction data in such situations.

For cell line descriptors, mRNA expression data restricted to DNA repair-related genes was used. DNA repair genes were as defined by a reference information service (URL=https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-genes.html). In addition to the gene expression data, we also tested the use of protein quantity (RPPA). mRNA and RPPA were also provided by Cell Miner-CDB.

Finally, one-hot encoding, a method to guarantee a numeric representation for each unique item in a collection, can differentiate each cell line without providing any molecular information and served as a cell line reference descriptor in control experiments.

2.3 Training

Random forest (RF, composite decision tree) models [5], support vector models (SVM), and K-nearest neighbors (KNN), all as implemented in the Python-language module scikit-learn [6], were used.

Classification models were evaluated using the Matthews Correlation Coefficient [7], also shown in a recent analysis to be less subject to misinterpretation in the case of data imbalance [8], and is computed from a standard classification confusion matrix as follows:

\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}
\]
50% of available data was randomly selected for training, and the predictive MCC on the remaining half was evaluated. Where given, confidence scores correspond to probabilities associated with predictions, as computed by scikit-learn RF models. For each of train-test split, training was repeated three times using different random seeds. For both classification and regression, train-test split generation was repeated 5 times to minimize artifacts from any one randomized trial.

For regression models built by RF, Pearson correlation (\(R\)) as well as a coefficient indicating the strength of correlation such that the linear fit of the data is constrained to pass through the origin, known as the \(R^2\) statistic [9], were used. We considered the Spearman rank-correlation coefficient as well, using the implementation provided by the SciPy library [10].

2.4 Biological pathway analysis

When employing transcriptome data in compound inhibition modeling, we also performed a downstream analysis of single-sample gene set enrichment analysis [11] [12] (ssGSEA) on the genes whose feature weights were high. ssGSEA is a technique which compares the ranks of individual gene expression levels (in a single sample) to a defined collection of genes associated with a biological function, and derives a continuously-valued score to suggest the strength of a gene set or pathway’s relevance. For annotation in this report, we used the collection of pathways defined in the Molecular Signatures Database, version 5.0 [13].

3 Results

3.1 Prediction of compound GI\(_{50}\) in cell lines

3.1.1 Cell line similarities based on biological description

As an initial investigation of the cell lines available, mRNA expression and protein quantity were used independently as initial cell line descriptors. Figure 1 shows similarities of cell lines. Heatmaps are colored based on the Pearson correlation of descriptors of each cell line. While melanoma (ME) cell lines and Leukemia (LE) cell lines exhibit relatively similar gene expression profiles, cell lines from breast cancer (BR) have diverse gene expression.

3.1.2 Comparison of descriptors

Out of different machine learning models tested, the highest MCC was achieved using the RF algorithm. From models trained using all DNA repair gene expressions, we identified the genes BLM, LIG1, TDG, POLD1, and SHPRH as highly influential factors based on their feature weights. We trained a model using only the expression of these 5 genes as a cell line descriptor, and it improved the prediction performance over the use of the full DNA repair gene panel employed. Models using protein quantities as a cell line descriptor suffered from poor prediction compared with models trained using gene expression (data not shown).

In addition, we manually picked 5 genes with well characterized biological roles (KRAS expression, MYC expression, SLFN11 expression, BRAF mutation, TP53 mutation). By analyzing feature weight parameters for the trained model, we found that predictions largely relied on specific compound substructural presence. This is in accordance with the previous research where Xia et al [14] reported that the compound descriptors provided the largest contribution to predictive ability of deeply-layered neural networks (deep learning). Figure 2 shows compounds containing such “high-importance” substructures, where substructures are highlighted in red. Proper interpretation of “high-importance substructures” using the RF estimator means that the fingerprint bits corresponding to these chemical substructures were frequently used as decision nodes in the decision trees comprising the forest model.

Note that the prediction of whether a compound is active or inactive will be made based on these substructures as well as the other biological descriptors mentioned above, and this result does not necessarily suggest that compounds with these substructures tend to be active or inactive.

Switching from the MACCS substructural descriptors to the DRAGON physicochemical descriptors, experiments were repeated. While averaged predictive MCC values were marginally higher for the DRAGON descriptors, the difference was not significant in (data not shown). Further, given that the DRAGON dataset was slightly reduced in size compared to the MACCS dataset (see Methods), direct comparison of MCCs is illogical.

3.1.3 Prediction for missing values

The original GI\(_{50}\) endpoint table contains a substantial amount of missing values, that is, there are pairs of cell lines and compounds having GI\(_{50}\) values of "NA". We used the trained model to make predictions on such missing values. Figure 3 shows several predictions, where predictions are based on MACCS keys and mRNA expression. In most cases, the prediction model simply classified compounds that are uniformly (in)active against
Figure 1 Cell line similarities (Pearson correlation) based on (a) gene expression (mRNA) or (b) protein quantity (RPPA).
Figure 2 The highlighted substructures were particularly highly weighted among the 166 MACCS compound fingerprints. Feature weights for substructures are shown under each figure, where weights of the whole model have been scaled such that they sum to unity, in accordance with their original description [15].

all profiled cell lines with the same label in the untested cell lines (rows in Figure 3). However, there were several cases where the prediction model predicted a compound to be active against a cell line when the compound was active against some cell lines and inactive against others. Examples include compound 380856 predicted to be active against line MALME-3M and compound 619029 predicted to be active against SR. These two examples are untested compound-line pairs and therefore represent potentially new experiments. More examples of untested compound-line pairs are provided in Table 1.

In addition to (in-)active categorical prediction, regressive models to predict numeric pGI$_{50}$ values were trained as well. Figure 4 shows the correlation between predicted pGI$_{50}$ and the actual pGI$_{50}$. Pearson, Spearman and R$^2$ correlations are strong, with a least-squares fit of the actual-predicted values having an offset of only half of a log unit. As the plot indicates, several compounds had negative pGI$_{50}$ values in cell lines, meaning molar or higher concentrations of compounds were required for inhibition, including compound 626674 against HOP62 (pGI$_{50}$ of -3.87) and compound 624589 against MALME-3M (pGI$_{50}$ of -4.00).

To evaluate the applicability of the resulting models, a prediction model was used to make predictions on an
external dataset containing new compounds. Note importantly that the GDSC dataset does not curate pGI\textsubscript{50} values but rather pIC\textsubscript{50} values, and further that there is a difference in response ranges (pGI\textsubscript{50} \sim [-4,12], pIC\textsubscript{50} \sim [3,7]), indicating that these two endpoints are not equal. Thus a perfect correlation is not to be expected, due to different biological parameters, including incubation time and normalization conditions. For example, the NCI60 experimental dataset documentation explains that cell lines were incubated for 48 hours after drug addition, while cell lines were incubated for 72 hours for the GDSC dataset protocol [3]. Further, there can be high variance even within one dataset (Figure 5(b)).

Figure 5 shows the resulting prediction on the external data. We confirmed that there was a moderate Pearson correlation of approximately 0.5 between the pGI\textsubscript{50} and pIC\textsubscript{50} on compounds that are present in both datasets. This indicates that the general direction of the trend is present. The Pearson correlation between predicted pIC\textsubscript{50} and actual pGI\textsubscript{50} was 0.233 with a corresponding p-value of 0.003 (2-tailed Pearson correlation test). While it is statistically significant and again the positive Pearson correlation indicates a general trend, we re-emphasize that caution against over-expectation is necessary when applying the model to untested compounds due to the high pGI\textsubscript{50} variance that underlies the regression model fit (Figure 5(b)).

3.2 Modeling of combination compound efficacy

We also applied fingerprint-based compound features and cell line gene expression to predict the pairwise combination compound effectiveness (see Methods). For the continuously-valued effectiveness score, we first used the thresholds of -100 and 100 to discretize compound pairs into inactive and active pairs, respectively.

![Figure 3 Cell line-compound activity prediction for unannotated pairs. Rows correspond to compounds and columns correspond to individual cell lines. Grey cells indicate that the recorded GI\textsubscript{50} values are between 1\mu M and 10 \mu M, and white cells are compound-line pairs missing in the data. A number of missing compound-line pairs are predicted to be active (red) or inactive (green) based on related annotations, where predictions are shown only when models report high confidence of a prediction.](image-url)
Table 1 Benchmark validation for predicted inhibitory activities in compound-line pairs that are not experimentally tested.

| Compound | Primary tumor type | Cell Line   | Confidence score |
|----------|--------------------|-------------|------------------|
| 380856   | Melanoma           | MALME-3M    | 0.935            |
| 619029   | Leukemia           | SR          | 0.925            |
| 731981   | Breast             | T-47D       | 0.900            |
| 681642   | Breast             | T-47D       | 0.900            |
| 771549   | Melanoma           | MDA-N       | 0.895            |

Figure 4 Numeric prediction using a random forest regressor model, evaluated with an external test dataset. The correlation between predicted pGI$_{50}$ and actual pGI$_{50}$ was evaluated using the Pearson and Spearman Correlations (annotated as R(P) and R(S), respectively) and origin-constrained best fit regression correlation (annotated as RO2). Dotted lines indicate actual or predicted values of 0. Green and red lines indicate actual-predicted best-fit equations computed passing through the origin or not, respectively, with equation coefficients given above the scatterplot.

3.2.1 Comparison of descriptors for activity classification

As cell line descriptors, we tested transcriptome expression of all DNA repair genes, one-hot encoding of DNA repair genes, and expression of a profile of 5 biologically important genes, in a fashion similar to Section 3.1.2. CLK2, NEIL3, SHPRH, DUT, and BLM were selected systematically by analyzing feature weights of a prediction model built on all DNA repair genes. Notably, models trained using the 5 systematically selected genes showed a higher prediction performance than the knowledge-based profile of 5 genes that are discussed in Section 3.1.2 (Table 2), indicating the possibility that they could be used as novel biomarkers for cancer therapy selection.

ssGSEA analysis was then applied to question if the genes selected were associated with pathways that are relevant from a biomedical perspective. First, POLD1 (obtained in Section 3.1.2), is the DNA polymerase delta catalytic subunit, and present as a trimer or tetramer; it exhibits both DNA polymerase and 3'- to 5'-exonuclease activities [16]. It combines with polymerase kappa to account for half of the repair activity in the nucleotide excision repair pathway [17].

BLM (Bloom Syndrome gene product) is a DNA
Table 2 Comparison of combination efficacy external dataset MCC with different cell line descriptors. Note that the cell line descriptors are combined with the chemical descriptors for model construction and external prediction.

| Compound descriptor | Cell line descriptor | test MCC         |
|---------------------|----------------------|------------------|
|                     |                      | Mean  | Standard Deviation |
| MACCS               | One hot encoding     | 0.72  | 0.0061             |
|                     | Full DNA repair gene panel expression | 0.71  | 0.0067             |
|                     | Knowledge-based gene set (5) | 0.72  | 0.0076             |
|                     | CLK2/NEIL3/SHPRH/DUT/BLM | 0.74  | 0.0088             |
|                     | CLK2/NEIL3/SHPRH/DUT/BLM and site of primary tumor | 0.74  | 0.0094             |
| pChem               | One hot encoding     | 0.73  | 0.0076             |
|                     | Full DNA repair gene panel expression | 0.74  | 0.0066             |
|                     | Knowledge-based gene set (5) | 0.73  | 0.0046             |
|                     | CLK2/NEIL3/SHPRH/DUT/BLM | 0.75  | 0.0043             |
|                     | CLK2/NEIL3/SHPRH/DUT/BLM and site of primary tumor | 0.76  | 0.0037             |

Figure 5 (a) Numeric predictions for pGI\textsubscript{50} on the external GDSC dataset with comparison to experimentally measured IC\textsubscript{50}. Despite the difference in ranges of pIC\textsubscript{50} and pGI\textsubscript{50}, the actual-predicted trend is moderately strong. (b) Background measurement data to interpret results of (a); the pGI\textsubscript{50} endpoint mean and standard deviation across experiments for cell line-compound pairs.

3.2.2 Regression prediction of combination efficacy

Similar to modeling experiments with the NCI-60 dataset, we also executed a study of numeric predictions of Drug Combo scores. Results are shown in Figure 6. Once again, the trend was captured overall by the MACCS and mRNA expression descriptors.

4 Discussion

There is a clear demand to increase the precision of computational analyses of patient data in order to suggest the optimal compounds for a patient in a given physiological condition. As physiological state is a consequence of both genotypic background and environmentally-induced gene expression, we rationalized that gene expression could serve as an additional contextual descriptor to strengthen chemical structure-activity relationships. In this work, we addressed both discrete and continuous endpoint measurements of chemical activity.

While the benefit of adding gene information was not clear for single compound cellular toxicity prediction (Section 3.1), the resulting model can extrapolate patterns...
found in the NCI-60 dataset into compounds or cell lines that are not in the original dataset. One issue that must be considered is the reproducibility of compounds’ cytotoxic activity, as GI$_{50}$ values only moderately correlated with IC$_{50}$, thus limiting the direct comparison of endpoints in different datasets.

The combo score prediction model could be employed to identify potential pairs of drugs that are highly effective for specific cancer cell lines. Multiple drug therapy is a common chemotherapy treatment regimen against many cancers (e.g., cisplatin and tamoxifen), and it is essential that pairs of drugs are selected to maximize the effect of the therapy. While gene expression is highly individualized to the patient, cancer stage, and the tumor microenvironment, it is still reasonable to expect that there is some degree of information transfer between patients with similar genome-transcriptome backgrounds. Thus the research herein to clarify the ability to classify or estimate drug combination efficacy on an untested patient by utilizing existing patient-derived reference cell line efficacy data provides insight for computational methods to assist personalized therapy selection.

Acknowledgements

This work was supported by Japan Society for the Promotion of Science grants 16H06306 and 17K20043 (to JB). Research exchange supported by JSPS Core-to-Core A: Advanced Research Networks also contributed to viewpoints. The authors wish to thank Professors Masatoshi Hagiwara and Shunichi Takeda of the Kyoto University Graduate School of Medicine for scientific support and philosophical discussions impacting this manuscript. TN also wishes to thank Dr. Fathi Elloumi, Dr. Sudhir Varma, Mr. William Reinhold, and Dr. Yves Pommier of the National Cancer Institute (USA) for discussions during the development of this work.

References and Notes

[1] R. H. Shoemaker, *Nature Reviews Cancer*, 6, 813 (2006).
[2] G. Zoppoli, M. Regairaz, E. Leo, W. C. Reinhold, S. Varma, A. Ballestrero, J. H. Doroshow, Y. Pommier, *Proceedings of the National Academy of Sciences*, 109, 15030-15035 (2012).
[3] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barborpe, H. Lightfoot, others, *Cell*, 166, 740-754 (2016).
[4] S. L. Holbeck, R. Camalier, J. A. Crowell, J. P. Govindharajan, M. Hollingshead, L. W. Anderson, E. Polley, L. Rubinstein, A. Srivastava, D. Wilsker, others, *Cancer research*, 77, 3564-3576 (2017).
[5] L. Breiman, *Machine learning*, 45, 5-32 (2001).
[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J.
Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *Journal of Machine Learning Research*, 12, 2825-2830 (2011).

[7] B. W. Matthews, *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405, 442-451 (1975).

[8] J. B. Brown, *Molecular informatics*, 37, 1700127 (2018).

[9] A. Golbraikh, A. Tropsha, *Journal of molecular graphics and modelling*, 20, 269-276 (2002).

[10] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. Mulbregt, S. I. 0. Contributors, *arXiv e-prints*, 2019, arXiv:1907.10121, 7.

[11] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *Proceedings of the National Academy of Sciences*, 102, 15545-15550 (2005).

[12] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, others, *Nature genetics*, 34, 267 (2003).

[13] N. B. Ivanova, J. T. Dimos, C. Schaniel, J. A. Hackney, K. A. Moore, I. R. Lemischka, *Science*, 298, 601-604 (2002).

[14] F. Xia, M. Shukla, T. Brettin, C. Garcia-Cardona, J. Cohn, J. E. Allen, S. Maslov, S. L. Holbeck, J. H. Doroshow, Y. A. Evrard, others, *BMC bioinformatics*, 19, 486 (2018).

[15] L. Breiman, J. Friedman, R. Olshen, C. Stone, *Group*, 37, 237-251 (1984).

[16] X. Meng, Y. Zhou, S. Zhang, E. Y. C. Lee, D. N. Frick, M. Y. W. T. Lee, *Nucleic acids research*, 37, 647-657 (2008).

[17] T. Ogi, S. Limsririchaikul, R. M. Overmeer, M. Volker, K. Takenaka, R. Cloney, Y. Nakazawa, A. Niimi, Y. Miki, N. G. Jaspers, others, *Molecular cell*, 37, 714-727 (2010).

[18] J. K. Karow, R. K. Chakraverty, I. D. Hickson, *Journal of Biological Chemistry*, 272, 30611-30614 (1997).

[19] G. Langland, J. Elliott, Y. Li, J. Creaney, K. Dixon, J. Groden, *Cancer research*, 62, 2766-2770 (2002).