Ben-Gurion University of the Negev
Faculty of Engineering Science
Department of Information Systems Engineering
Data Science in Cell Imaging S2 2021

Treatment Compound Prediction From Treated Cells Imaging

Gil Shamay gshamay@post.bgu.ac.il

1 Introduction

Using microscopy Image based screening of single cells, researches can generate rich information such as morphological profile with multiple features about the cells that can help analyzing and understanding their biological state and condition. High-performance computational resources are helpful in the cells profiling process and increase variety of options used by researches in cell biology and drug discovery, where machine learning approaches are used for the analysis of high dimensional profiling data and for improving many steps in their work. Mining those profiles can find patterns that detect unexpected biological activity, help in identifying disease that are associated screenable phenotypes, understanding disease mechanisms and predicting a drug's activity, toxicity or mechanism of action and etc.

In their dataset [1], Bray et al. produce such profiling data which we used for mining, and using artificial neural network, we built an application that predict the atoms that are required to build a treatment that will cause a cell to have a morphological change as required.

From the experiments that were done we can see that some level of prediction can be achieved, however this can not yet be used for any real usage or production, but can only be a beginning of a more complex research that may be done in that area.

2 Background

In [5], Scheeder et al. reviewed different methods that use deep neural networks that use image based screening and profiling, with a focus on chemical genetics in drug discovery. Simm et al. showed in [6] that the biological activity of cell compounds can be predicted using large-scale imaging screening of cells. Chen et al. [2] review a few works that used different methods of machine learning to predict molecular design. In one of those presented methods [3] the authors used deep learning methods on chemical data that was represented as SMILES to generate new molecules. Taking in consideration that optimizations in molecular space are extremely challenging, this technique can be used for efficient exploration and optimization through and allows to generate open-ended spaces of chemical compounds. In our work, we use machine learning method

on the image based screening and profiling joined with chemical information supplied as SMILES.

3 Objectives

In this work we will try to predict the treatment molecular compound, that should be used for perturbing a given cell, according to the the cell morphological profile parameters, that we would like to archive when using the predicted treatment. We would like to be able to predict which atoms, and how many of each atom type, should be used in order to generate the treatment formula that would affect the cell and cause the required morphological change.

This can help researchers to design the treatment that they need to use, in order to get some specific morphological change on cells.

4 Methods

4.1 The Data

The data used for this work is "A dataset of images and morphological profiles of 30000 small-molecule treatments using the Cell Painting assay" by Bray et al. It is a microscopy dataset [1] includes 919265 five-channel fields of view, representing 30616 tested compounds, available at "The Cell Image Library"(CIL) repository. In this dataset, human U2OS cells were plated in 384 well plates in a total of 406 plates and treated with each of 30616 compounds in quadruplicate. The 5 channels are:
(1) Hoechst 33 342, which provides the ability to capture the image of the Nucleus (DNA CellProfiler)
(2) Concanavalin A/Alexa Fluor 488 conjugate, which provides the ability to capture the image of the Endoplasmic reticulum (ER CellProfiler)
(3) SYTO 14 green fluorescent nucleic acid stain , which provides the ability to capture the image of the Nucleoli, and cytoplasmic RNA (RNA CellProfiler)
(4) Phalloidin/Alexa Fluor 594 conjugate, wheat germ agglutinin (WGA)/Alexa Fluor 594 conjugate, which provides the ability to capture the image of the F-actin cytoskeleton, Golgi,plasma membrane (AGP CellProfiler)
(5) MitoTracker Deep Red, which provides the ability to capture the image of the Mitochondria (Mito CellProfiler)
Each plate contain both 320 perturbed wells and 64 control wells (that were not perturbed with any treatment). The dataset also includes data files containing morphological features derived from each cell in each image, both at the single-cell level and population-averaged level(i.e., per-well). The image analysis workflows that generated the morphological features are also provided.

In addition, chemical annotations are supplied for the compound treatments applied. The treatment chemical annotations are provided as SMILES formulas The population-averaged morphological data is provided in a total of about 4 gigabytes csv files, where each plates data is stored in a separated csv file.

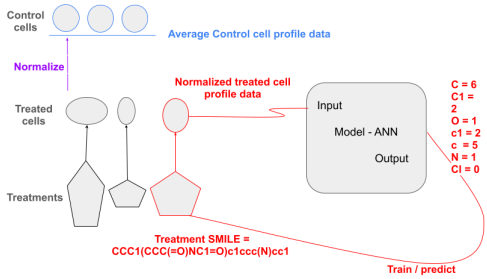

Figure 1: Model Illustration

The supporting data for this dataset is public and accessible at github. More information and links to all plates data and morphological tables can be found at gigadb.The data is also available by FTP.

In this work, we used only the data provided for the population-averaged (average well profiles, i.e. a single line of data is used for each well) with the chemical annotations data. Each well has 1783 feature values.

5 Research Plan

We will build a model that get as an input all 1783 features that exists in the dataset for the well average cell morphological profile, and will output the prediction of the number of atoms that should be used in the treatment that will cause the cell to transform to be as the input morphological description.

5.1 Preprocess Chemical Annotations

We first read all SMILES formulas from the Chemical Annotations csv file. parse the SMILES formula according to the atoms that it's built from. In the SMILES parsing process, we count the number of times that each atom type appears in each formula. This will be used later as the output of the model. We also count the number of formulas that use each atom type. This will be used for generating random predictions that are build with atoms distribution that is close to those that exists in the data.

5.2 Plates Data

5.2.1 Data Reading

We first read the list of available plates (the list of available *mean well profiles.csv* files) Due to the size of the data, we read each plate separately and process it either for training or validating the model.

5.2.2 Split Train and Validation Data

To avoid bias, we use each complete plate either for training or for validations. we do not use wells from a single plates, for both training and validation. The train and validation split rate will be discussed later in the 'Evaluation' section.

5.2.3 Normalize

Since each plate data may be biased non-biological factors, known as the *'batch effect'* [4], we normalize the wells (population-averaged) data, according to the control wells that are provided within every plate. For each plate, we first split the control wells from the perturbed wells. Then we calculate average of each feature value among all the control wells in the plate. We normalize each of the perturbed wells by subtracting the average control value from each of the treated wells values. We normalize both the training and the validation plates data.

5.3 Prediction Model

In order to predict the required target, we build an Artificial Neural Network model (ANN). We build a fully connected network that has 1783 input neurons, one for each morphological profile value from the average well values, and two inner layers: one with a size of 891 neurons, and one with the size of the expected output, 14 neurons (as the number of atoms that are used across all treatments chemical compounds). The activation method of both layers is *sigmoid*, and the loss function is *Mean Squared Error*.

5.4 Evaluation

We evaluate the model in 10 cross validation method, where in each validation step, we choose 10% of the plates, and use them as the validation set. We fit the model using the left 90% of the data, and calculate the RMSE (Root Mean Square Error) of the predicted formula compound, with the actual formula compound that was used in the predicted well. In each validation cycle, we chose a different validation set, without overlapping.

As baseline we generate a *random prediction*. For each atom type that was used in any of the known treatments in the dataset, we first select if the Atom will be used in the randomized formula. We select it randomly based on the distribution of the usage of the atoms in the given treatments compounds. Then, if the given atom type was choose to be part of the new formula, we randomly select the number of atoms of this type, based on the maximum number of atoms that were used from this atom, across all known treatments. Then we calculate the RMSE of the 'random prediction' and compare it with the RMSE we got by the model prediction. We calculate the average RMSE values for each cross validation cycle and for the overall validation iterations.

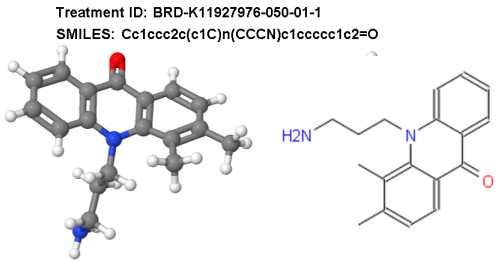

Figure 2: Future Work: Use and predict treatment molecule image representation of SMILES

6 Results

The mean RMSE that was given by the predicted formula was **4.383**, and the one provided by the random values produce RMSE of **9.4185**. We can see that the predicted results are almost 50% better then the random.

7 Discussion

From domain experts, we know that most treatments that were used does not have a significant biological affect on the cells. Surprisingly, from the given results we can see that there is probably some morphological affect on the cells even for treatments with no proven biological affect. However, the result provide an error that is probably not usable for any real application but the fact that it's better then the random, gives us a hint that there may be some co relation between the treatment compound and the affect it has on cells.

There may have been a large lose of information, in the way we parsed the SMILES and used the treatment representation in a too simple way. We did it for simplicity reasons and we assume that using a structure that will hold more information (like the order of the atoms, the isotopes and the type of connections) may produce better results. There is a known bias with the random treatment, as it is calculated by using treatments from all wells, including the validations wells. Fixing this is expecting to increase a bit the distance between the prediction and the random RMSE values.

7.1 Future Work

Following the lose of chemical information data as discussed before, we suggest to use the SMILES image representation as a better method to describe the treatment formula. As in [3], we find it as a representation that may keep the complete molecule structure, without loosing precious information. There

are available applications that generate image from SMILES (see figure 2), and CNN networks should be able to handle such images naturally.

We can also suggest developing a better normalize method in the prepossess step and produce a more clever random treatment, with more precise atoms distribution

The code used for this project, and the output of the experiment run are public and published @ https://github.com/gshamay/DSCI

References

[1] Mark-Anthony Bray, Sigrun M Gustafsdottir, Mohammad H Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L Sokolnicki, Joshua A Bittker, Nicole E Bodycombe, Vlado Dančík, Thomas P Hasaka, et al. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the cell painting assay. *Gigascience*, 6(12):giw014, 2017.

[2] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 2018.

[3] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[4] J Luo, M Schumacher, Andreas Scherer, Despoina Sanoudou, D Megherbi, T Davison, T Shi, Weida Tong, Leming Shi, Huixiao Hong, et al. A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data. *The pharmacogenomics journal*, 10(4):278–291, 2010.

[5] Christian Scheeder, Florian Heigwer, and Michael Boutros. Machine learning and image-based profiling in drug discovery. *Current opinion in systems biology*, 10:43–52, 2018.

[6] Jaak Simm, Günter Klambauer, Adam Arany, Marvin Steijaert, Jörg Kurt Wegner, Emmanuel Gustin, Vladimir Chupakhin, Yolanda T Chong, Jorge Vialard, Peter Buijnsters, et al. Repurposing high-throughput image assays enables biological activity prediction for drug discovery. *Cell chemical biology*, 25(5):611–618, 2018.