Isabelle Mieling
Harvard T.H. Chan School of Public Health
Boston, MA
January 4, 2018

Thesis Advisors: Dr. Paul Clemons, Dr. John Quackenbush
Proposed Research Project: Predictive drug candidate toxicology using high-throughput multiplex imaging methods

Computational Biology & Quantitative Genetics Master's Thesis Proposal: 2017-2018

Background

Developing safe and effective drugs for treating human disease requires thorough investigation to determine the likelihood of toxicity or adverse side effects. Determining these effects can be slow, expensive, and require large amounts of compound. Many drug candidates fail in or before clinical trials due to high levels of toxicity or adverse side effects, further emphasizing the need for a way of testing compounds in a high-throughput fashion to predict future toxicity, ideally using chemical structure or calculable properties. Though work in predictive toxicology has been done, none has used high-dimensional cell-based profiling methods to achieve this goal. Small-molecule profiling, including some techniques developed at the Broad Institute, has been used as a tool to discover mechanisms of action and identify compound similarities, among other things. The aim of this project is to use existing and novel small-molecule profiling techniques to generate profiles of drug activity that can be used as predictive signatures of drug toxicity and adverse side effects.

Approach

We will begin by using data from two existing profiling methods, developed at the Broad Institute, "L1000" and "cell painting". L1000 was used to create a small-molecule connectivity map from gene-expression readouts of compound action, and cell painting produces cell-morphology profiles after compound treatment using automated microscopy [1,2]. These two profiles have been shown to reveal a greater diversity of screening activity than those based on chemical structure alone [3].

As a starting point for analysis, we will study a diverse collection of current and withdrawn drugs and drug candidates with varying degrees of adverse effects in humans. We will assemble this collection with compounds from the Broad Institute Repurposing Hub as well as two external public databases, SIDER and WITHDRAWN [4,5,6]. Upon assembly, compounds will be curated for chemical information, chemical structure properties will be computed, and for those compounds for which we have L1000 and cell-painting data, phenotypic profiling data from these assays will be collected. These will be appended to side-effect and toxicity data from SIDER and WITHDRAWN databases. Lastly, using a small subset of this data, for which we have L1000, cell-painting, SIDER, and WITHDRAWN data, initial modeling will begin to identify clusters of profiles

of groups of compounds known to cause adverse effects. This initial predictive modeling may guide future large-scale modeling and will provide insight into the data.

Resources

We will initially consider three databases containing current and withdrawn drugs and drug candidates with varying degrees of adverse effects in humans. The Broad Institute Drug-Repurposing Hub contains 2,350 FDA-approved drugs, 1,600 drugs that reached phases 1-3 of clinical development, and 95 compounds that were previously approved but withdrawn from use. Annotations include compound name, chemical structure, clinical trial status, mechanism of action, protein targets, and disease areas, among others. The SIDER Side Effect Resource contains information on 1,430 approved drugs and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. Annotations include side-effect frequency, and drug and side-effect classifications, among others. The WITHDRAWN database contains information on 578 withdrawn or discontinued drugs removed from the global market due to safety concerns. It contains information about the therapeutic targets, off-targets, toxicity types, and biological pathways associated with the drugs, among others.

Broad Drug Repurposing Hub: https://clue.io/repurposing
SIDER Side Effect Database: http://sideeffects.embl.de
WITHDRAWN Database: http://cheminfo.charite.de/withdrawn/

Specific Aims

1. Data harmonization: Download and integrate public side-effect and adverse-events databases
    a. Assemble small collection of 95 FDA-approved small-molecule drugs withdrawn from the US market, available through the Broad Repurposing Hub
    b. Download data from SIDER and WITHDRAWN databases
2. Data integration: Curate chemical information for compounds, compute chemical structure properties and features, integrate phenotypic profiling data already available
    a. Collect profiling data on available compounds from prior L1000 and cell-painting experiments
    b. Append high-dimensional data to vector of side-effect and toxicity information that can be extracted from SIDER and WITHDRAWN databases
3. Computational analysis: Begin application of existing or development of novel computational methods to predict side effects and adverse events from profiling data
    a. Select subset of compounds for which we have L1000 and cell painting data to begin modeling
    b. Use comparisons between cellular assays to identify clusters of profiles corresponding to groups of compounds known to cause these adverse effects by different mechanisms

References

[1] Subramanian, A., et al. “A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles.” *Cell*, vol. 171, no. 6, 30 Nov. 2017, pp. 1437–1452., doi:10.1016/j.cell.2017.10.049.

[2] Bray, Mark-Anthony, et al. “Cell Painting, a high-Content image-Based assay for morphological profiling using multiplexed fluorescent dyes.” *Nature Protocols*, vol. 11, no. 9, 25 Aug. 2016, pp. 1757–1774., doi:10.1038/nprot.2016.105.

[3] Wawer, MJ., et al. “Toward performance-Diverse small-Molecule libraries for cell-Based phenotypic screening using multiplexed high-Dimensional profiling.” *PNAS*, vol. 111, no. 30, 14 July 2014, pp. 10911–10916., doi:10.1073/pnas.1410933111.

[4] Corsello, Steven M, et al. “The Drug Repurposing Hub: a next-Generation drug library and information resource.” *Nature Medicine*, vol. 23, no. 4, 7 Apr. 2017, pp. 405–408., doi:10.1038/nm.4306.

[5] Kuhn, Michael, et al. “The SIDER database of drugs and side effects.” *Nucleic Acids Research*, vol. 44, no. D1, 19 Jan. 2015, pp. 1075–1079., doi:10.1093/nar/gkv1075.

[6] Siramshetty, Vishal B., et al. “WITHDRAWN—a resource for withdrawn and discontinued drugs.” *Nucleic Acids Research*, vol. 44, no. D1, 8 Nov. 2015, pp. 1080–1086., doi:10.1093/nar/gkv1192.