Exploiting cellular CRISPR-Cas9 and chemogenomics platforms for drug discovery

Andrew Leach (EBI) and Mathew Garnett (Sanger)

This research project aims to bring together two of the Genome campus' key "molecular" resources: the large-scale drug sensitivity screening platforms of the Garnett group with the chemogenomics resources and chemo-informatics expertise in the Leach team. The high-level goals of our proposed project are:

- To characterise and understand at a molecular level cellular drug sensitivity screens
- To identify compounds for further iterations of screening that will add value to the current knowledge base
- To characterise and understand the output from genome-wide CRISPR-Cas9 synthetic lethal screens in the context of chemogenomics datasets
- To build computational models that can be used to predict the effects of drug administration on cells

Introduction

The Garnett group has established a panel of more than 1000 human cancer cell lines and organoids representing diverse histological and genetic sub-types. These cell culture models are highly annotated at the level of the genome, transcriptome, and epigenome. The cell models are used to profile potential drug candidates and other molecules of interest in order to understand and predict drug sensitivity in patients. To date, over 600 compounds have been profiled for their differential sensitivity across the cell line panel, and both single agent and two-drug combination screens are on-going. This cellular screening platform is now being complemented by a genome-wide CRISPR-Cas9 synthetic-lethal screening platform that will systematically inactivate genes across the genome to nominate new oncology drug targets.

The latest release of EMBL-EBI's core chemogenomics resource ChEMBL includes data on more than 1.7m unique chemical entities across more than 14m assays. ChEMBL is widely used to answer key questions in drug discovery, such as the identification of tool compounds, understanding trends in SAR and selectivity, deconvoluting phenotypic screens and to build computational models that can be used to predict target-drug interactions including potential off-target effects.

Our proposed project aims to capitalise on the expertise and synergies in our respective groups to derive further understanding and insights into the experimental platforms, to develop methods for predicting compound behaviour and to enhance the use of the combined platforms and resources for drug discovery applications. We have divided the work into four sub-projects, each aligned to one of the high-level goals of the overall collaboration, though it should be noted that there is significant synergy between each of these areas.

Leveraging compound pharmacology to interpret cellular sensitivity

Many of the compounds that have been profiled in differential drug sensitivity screens have a rich published pharmacology which is not currently utilised, but is curated and captured in ChEMBL. This provides an exciting opportunity to seek to understand and rationalise drug sensitivity data in the context of the known compound pharmacology. Such studies can be performed by considering the responses of multiple compounds across individual cell cultures, single compounds across multiple cellular data and whole-matrix analysis. Of particular interest will be the identification of structurally similar compounds that show different cellular responses, to determine whether the known pharmacology and compound SAR can rationalise such variation and gain fundamental insights that

have broader application. Where appropriate, computer models that predict compound pharmacology will be used to “fill gaps” in the available experimental data.

Identifying informative compounds for further screening iterations

The current compound set has evolved in a somewhat “ad hoc” fashion, driven by compound availability, research interests and external requests. As a consequence, the current data set is potentially biased or imbalanced relative to a more informative compound set. Moreover, no consideration has been taken of the known molecular pharmacology of the compounds when making selections. The goal in this sub-project is to identify additional compounds for sensitivity screening that will provide significant value to the existing set. For example, we will identify additional compounds which maximise coverage of proteins and pathways implicated in carcinogenesis through the mining of patient genomic datasets. We will explore a number of potential sources for such compounds, including marketed drugs or compounds in clinical trials, public compound sets such as the GSK PKIS (Published Kinase Inhibitor Set), validated chemical probes (see for example http://www.chemicalprobes.org) and through our contacts with the academic, biotech and Pharma communities.

Characterising genome-wide CRISPR-Cas9 synthetic lethal screens

The CRISPR-Cas9 screening platform offers the exciting prospect of being able to efficiently generate data at the genome scale. Of particular interest in the context of the proposed project will be the correlation between the CRISPR data relative to that from the drug sensitivity screens, and to seek to understand such data in the context of the known compound pharmacology. An overall goal will be to establish links between the CRISPR-modified cellular systems, their response to compound administration, and the underlying molecular recognition events that compound is involved in. By comparing screening and bioactivity data across multiple compounds we will be able to generate testable hypotheses for subsequent follow-up, for example by identifying additional compounds to screen and to guide more detailed mechanistic studies.

Construction of computational models to predict the effects of compound administration

The large data sets generated by the cellular and gene inactivation experimental platforms will provide an excellent opportunity to create computational models that can be used to predict the effects of hitherto unscreened (or even yet-to-be-synthesised) compounds. A number of methodologies will be explored for the construction of such models. One approach will be based on descriptors derived from the molecular structures, analogous to a “traditional” QSAR approach. We will also explore how the molecular pharmacology data from ChEMBL can be used as a descriptor (a so-called “bioactivity fingerprint”) to predict the more complex cellular/CRISPR output. Further studies will explore how subsets of the cellular or CRISPR data could be used to build models of related but unseen cellular/CRISPR endpoints. A variety of model-building approaches will be explored; of particular interest will be the potential application of the newer deep-learning methods which will benefit from the large volume of data available from these experimental platforms

Resources and timescales

A strength of this project is the availability of existing large datasets and the closely coupled, highly iterative nature of the interactions between the experimental and computational work. It will therefore be important to identify a candidate who not only has excellent computational and analytical skills but also has a robust understanding of the underlying biological aspects of the screening platforms. Further, he/she will need to be a good team player, able to interact effectively with scientists and colleagues from multiple disciplines and to forge links as needed to the external community.