Data-driven Experimental Prioritization via Imputation and Submodular Optimization

Jacob Schreiber [1] **William S. Noble** [2]

Abstract

An unfortunate reality is that modern science is often limited by the number of experiments that one can afford to perform. When faced with budget constraints, choosing the most informative set of experiments sometimes requires intuition and guess-work. Here, we describe a data-driven method for prioritizing experimentation given a fixed budget. This method involves first predicting the readout for each hypothetical experiment and, second, using submodular optimization to choose a minimally redundant set of hypothetical experiments based on these predictions. This approach has several strengths, including the ability to incorporate soft and hard constraints into the optimization, account for experiments that have already been performed, and weight each experiment based on anticipated usefulness or actual cost. Software for this system applied to the ENCODE Compendium can be found at https://github.com/jmschrei/kiwano.

An unprecedented number of genomic assays have been developed in the past decade, dramatically expanding the types of biochemical measurements one can make of the genome. Some of these assays provide higher quality readouts than traditional assays, such as ChIP-exo and CUT&RUN improving upon ChIP-seq. Other assays extend the readouts to the single-cell and spatial dimensions. Some are capable of making new measurements, such as measuring chromatin architecture or enhancer activity directly. Together, these assays are invaluable for developing a complete understanding of how the genome works across complex tissues and organisms

Unfortunately, despite the clear value of thse assays, they are expensive and require expertise to perform well. Because budgets are limited, the application of these assays is unlikely to be comprehensive. Thus, investigators and consortia generally must choose a subset of potential experiments to perform. This selection task is challenging because one cannot know exactly how informative an experiment will be before it is performed, and so selection is sometimes guided by intuition and informed guesses.

In a recent publication (Schreiber et al., 2021), we proposed a data-driven method for directing experimentation given a vast search space and a limited budget. The method relies on submodular optimization, which has been described as a discrete analog to convex optimization, for choosing a minimally redundant subset of elements from a collection in a principled manner. In this setting, the elements are combinations of assays and cell types, some of which have not yet been performed ("hypothetical experiments"). A challenge in applying submodular optimization to this selection problem is that the strategy we use relies on similarities being calculated between each hypothetical experiment—difficult because one cannot calculate similarities between data that does not exist. Our proposed procedure overcomes this challenge by first predicting the readouts for all hypothetical experiments using a machine learning model and then calculating similarities between these predicted readouts.

Our work demonstrated this approach with an application to the ENCODE Compendium. We focused on a subset of the compendium that spanned 400 cell types and 77 assays and asked "given the 3,510 experiments that have been performed, what are the next experiments that should be performed to maximize information content?" Notably, the selected experiments should minimize not only redundancy with each other, but also redundancy with the experiments that have already been performed. Instead of building our own predictive model, we downloaded imputation for all 30,800 experiments made by Avocado spanning the ENCODE Pilot Regions. Then, we calculated a similarity matrix across all experiments based on the Pearson correlation. A UMAP projection of this similarity matrix (Figure 1A) shows several distinct clusters that mostly align with different forms of biochemical activity (see paper for more details). Finally, we used submodular optimization,

[*] Equal contribution [1] Department of Genetics, Stanford University, Stanford, California, USA [2] Department of Genome Science, University of Washington, Seattle, Washington, USA. Correspondence to: Jacob Schreiber <jmschreiber91@gmail.com>.

The 2021 ICML Workshop on Computational Biology. Copyright 2021 by the author(s).