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.

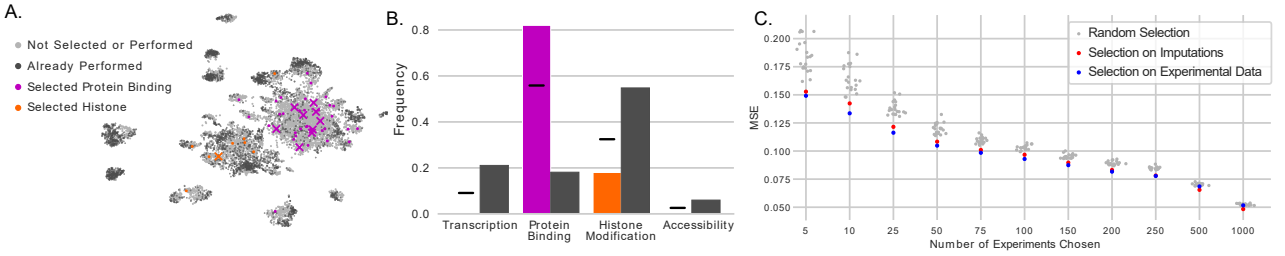

Figure 1. **Application of the proposed method to the ENCODE Compendium.** (A) A UMAP projection of the similarity matrix derived from 30,800 imputed ENCODE experiments where each dot is an experiment. The first 10 selected experiments are colored and marked as an X and the next 40 selected experiments are only colored. (B) The number of selected (colored) and already performed (dark grey) experiments by form of biochemical activity. The lines represent the number of experiments one would select by change. (C) Performance of a linear regression model that takes in signal from the chosen experiments at a genomic coordinate as input and predicts the signal at that coordinate for all 3,510 experiments as output. When the experiments are chosen using submodular optimization, the experiments used as input are more representative of the whole output space than random chance.

seeded with the 3,510 experiments that have already been performed, to rank all of the hypothetical experiments. We found that the first 50 chosen experiments are enriched for protein binding, which was unsurprising given that so many transcription and chromatin accessibility experiments have already been performed (Figure 1B) and that proteins can bind in very different patterns from each other.

A central premise of submodular optimization is that elements are chosen to minimize redundancy. Hence, one would expect that the chosen experiments are more representative of the entire space of hypothetical experiments than a randomly selected set. A fair way of measuring representativeness (in this context) is to use the chosen experiments to predict the entire set of experiments because redundant inputs would lead to a less accurate model than representative ones. When we trained linear regression models using experimental (not imputed) data, we found that those which used experiments selected using submodular optimization were much more accurate than those which used experiments that were selected randomly (Figure 1C).

An interesting finding from our work emerged when we applied our procedure to rank the order that experiments should be performed in each cell type or assay independently. For each cell type and assay, we measured how redundant the experiments were when selected using our approach as measured by a facility location function. We found that, as expected, the experiments selected using our approach were less redundant than those selected randomly. However, unexpected, we found that the actual set of experiments performed in almost every cell type and assay were *more* redundant than randomly selected experiments in almost every case. Although this can be partially explained by experiments being chosen to pursue a specific research agenda, such as comparing different forms of RNA-seq measurements, the redundancy highlights that current research agendas do not naturally result in the most globally informative experiments being performed.

A strength of our proposed approach is that it is extremely flexible. In our work, we demonstrated two constraints that could be incorporated into the optimization directly. First, we weighted each protein binding experiment by the expression of that protein according to an RNA-seq experiment. Reweighting in this manner emphasizes experiments measuring the binding of proteins that are expressed, while deemphasizing experiments measuring the binding of proteins that are not expressed. Second, we proposed a novel submodular-supermodular objective function that encourages the selected experiments to span a small set of assays or cell types. This constraint can help address the real-world difficulties in culturing many cell types or needing to perform many different types of assays.

Although our work focused on the application of this procedure to ENCODE data, the procedure can be applied to any collection of data where a predictive model can be built. Even if the data spans only a single assay, one could predict the readouts for that assay in all missing cell types and use submodular optimization to choose the next experiments to perform. We anticipate that our method will serve as a useful resource for groups that seek to make the most use of a limited budget.

References

Schreiber, J., Bilmes, J., and Noble, W. S. Prioritizing transcriptomic and epigenomic experiments using an optimization strategy that leverages imputed data. *Bioinformatics*, 37(4):439–447, May 2021.