Benton Capstone Model Documentation
Spring 2024

Harrison Tognela, Guy Wood, Henry Burch, Yihan Zhang, Katrina Nowak

Introduction

This documentation provides a comprehensive overview of the development and assessment of a novel approach to enhancer-to-gene target mapping. The project was born out of the need to better understand the gene targets of individual enhancers across tissues, a critical aspect when interpreting the effects of genetic variants occurring within enhancer sequences.

Enhancers, a class of gene-regulatory sequences, play a pivotal role in the tissue-specific activation of genes. Disruptions within these sequences have been linked to numerous diseases, including cancer and neurodevelopmental disorders. Despite their importance, linking enhancers to their target genes remains a significant challenge in genomics due to uncertainties about the exact biological mechanisms enhancers use to target genes, limited functional data across different cell types, and variability imposed by data generation and integration processes.

Our project aimed to address this gap by developing a new statistical model to link enhancers to target genes, integrating features from existing methods used by EnhancerGenie, and other relevant features. The features considered include the distance from the enhancer to the gene, similarity between the enhancer sequence and the gene's promoter sequence, correlation between epigenetic marker presence or strength at the enhancer and gene, and colocalization in the same 3D genome loop or connection based on chromatin interaction data.

In addition to method development, we also conducted a comprehensive tool assessment. This involved evaluating the similarity of gene sets linked by different approaches using gene ID/name, functional annotations, and pathway analysis. We also conducted a case study using a set of enhancers with experimentally validated gene targets (Bengi) to identify the strengths and weaknesses of existing approaches.

The following sections will delve into the specifics of our method development and tool assessment activities, providing a detailed account of our findings, challenges encountered, and the solutions we devised. Our hope is that this documentation will serve as a valuable resource for future researchers in the field of genomics.

Dealing with Imbalanced Data: The Importance of Resampling and Generative Approaches

In the field of machine learning, the quality and balance of the dataset play a crucial role in the performance of the models. Our project utilized the Benchmark of Candidate Enhancer-Gene Interactions (BENGI) dataset, a comprehensive bedfile encompassing various tissues and gene-enhancer pairs. This dataset includes details such as the enhancer start and end, gene position, chromosome, and whether or not they interact.

However, a significant challenge we faced was the imbalance in the data. Approximately only 5% of the data was labeled as correctly paired or interacting. This imbalance led to a bias in the models we initially tried, resulting in them rarely predicting that the gene and enhancer interacted. This is a common issue in machine learning, where the model becomes biased towards the majority class, in this case, the non-interacting pairs.

Balancing the data is crucial to ensure that the model does not become biased and can accurately predict both classes. To address this, we explored resampling techniques and generative approaches.

Resampling techniques, such as oversampling the minority class or undersampling the majority class, can help balance the classes in the training data. However, these techniques come with their own set of challenges. Oversampling can lead to overfitting, while undersampling can result in loss of information.

On the other hand, generative approaches can create synthetic examples of the minority class, thereby balancing the dataset without losing information or causing overfitting. These methods, such as the Synthetic Minority Over-sampling Technique (SMOTE), have shown promise in dealing with imbalanced datasets.

In the following sections, we will delve into the specifics of the resampling and generative approaches we tried, the challenges we encountered, and how these methods improved the performance of our models.

Data Balancing Techniques: SMOTE, Generative Adversarial Networks, and Diffusion Models

In this project, we explored several techniques to address the issue of class imbalance in our dataset. Here's a brief overview of each method:

SMOTE (Synthetic Minority Over-sampling Technique): This is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

Generative Adversarial Networks (GANs): GANs are composed of two deep neural networks competing against each other in a game. Given a training set, this technique learns to generate new data with the same statistics as the training set. For example, a GAN trained in photographs can generate new photographs that look at least superficially authentic to human observers. However, in our case, GANs were not ideal due to the low dimensionality of our data.

Diffusion Models: These are generative models that generate samples from a simple distribution via a sequence of small diffusion steps. They have been used successfully in various applications, including image synthesis and denoising. However, we encountered issues with the versioning of CUDA during our project, which prevented us from fully exploring this method. It remains a promising approach to try again in the future.

SMOTE (Synthetic Minority Over-sampling Technique) emerged as the most effective method for our project due to its ability to generate synthetic examples of the minority class, thereby balancing the dataset without causing overfitting or loss of information.

For instance, consider a dataset where the instances of the minority class are sparse and scattered. In such a case, a model trained on this data might struggle to correctly classify the minority class due to the lack of representative data. However, with SMOTE, synthetic examples are created by interpolating between several minority class instances that lie together. These new instances increase the representativeness of the minority class, enabling the model to make better and more accurate predictions.

In our project, where only 5% of the data was labeled as correctly paired or interacting, SMOTE helped to balance the classes, leading to a significant improvement in the performance of our models. It allowed the models to better understand the characteristics of the interacting pairs, thereby enhancing their ability to correctly predict interactions. This is why SMOTE was the best method for our project.

In the following sections, we will delve into the specifics of each method, the challenges we encountered, and how these methods improved the performance of our models.

Machine Learning Models: SVM, Random Forest, and Naive Bayes

In this project, we explored several machine learning models to predict enhancer-gene interactions. Here's a brief overview of each model:

Support Vector Machines (SVM): SVM is a powerful and flexible class of supervised algorithms for both classification and regression. It is effective in high dimensional spaces and best suited for problems where the number of dimensions is greater than the number of samples.

Random Forest: Random Forest is an ensemble learning method that operates by constructing multiple decision trees at training time and outputting the class that is the mode of the classes of the individual trees. It is effective for dealing with overfitting and works well with a large range of features.

Naive Bayes: Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. They are highly scalable and able to handle large amounts of features.

In the context of our project, Naive Bayes outperformed SVM and Random Forest. A few reasons could explain this:

- Assumption of Independence: Naive Bayes classifiers assume that the presence of a particular feature in a class is unrelated to the presence of any other feature. This assumption often holds true in genomic data where different features (like enhancer length, gene length, interaction distance, etc.) can often be considered independent.

- Performance with Small Datasets: Naive Bayes classifiers can perform well even if the model's independence assumption is broken, which makes it a robust model. It also performs well with small datasets, which might be the case in the project given the imbalance in the data.

- Probabilistic Approach: Unlike SVM or Random Forest, which are deterministic models, Naive Bayes provides a probabilistic framework. This allows for a more nuanced understanding of the data and can better handle the uncertainties inherent in predicting enhancer-gene interactions.

In the following sections, we will delve into a novel approach that transcends traditional machine learning methods, focusing instead on leveraging biological domain knowledge to predict enhancer-gene interactions.

Spatial Proximity Threshold: A Novel Approach

In our quest to predict enhancer-gene interactions, we introduced a novel approach that deviates from traditional machine learning methods. This approach, which we refer to as the Spatial Proximity Threshold Concept, leverages the biological domain knowledge that spatial proximity plays a significant role in enhancer-gene interactions.

The concept is simple yet powerful. We establish a spatial proximity threshold, representing the maximum distance between enhancer and gene midpoints where interactions are most likely to occur. This threshold was determined by analyzing the distribution of enhancer-gene distances for both interacting and non-interacting pairs and employing statistical methods to identify a distance where the probability of interaction significantly exceeds that of non-interaction. Our analysis pinpointed an optimal threshold at approximately 185,400 base pairs with an accuracy of 98.4%.

This approach offers several advantages:

- Screening Efficiency: The threshold serves as a powerful filter to prioritize enhancer-gene pairs for detailed analysis, reducing computational load and focusing efforts on the most promising candidates.
- Biological Insights: It emphasizes the role of spatial proximity in enhancer-gene interactions, corroborating existing genetic theories and providing a quantitative basis for further investigations.
- Generalizability and Overfitting: Unlike machine learning models, which may overfit to the training data, this approach is more generalized and less prone to overfitting. It is based on a fundamental biological principle, making it broadly applicable across different datasets.
- Simplicity: The threshold concept is straightforward and easy to understand, making it accessible to researchers from various backgrounds.
-

This thresholding approach opens new avenues for refining predictive models in genomics, suggesting that similar methods could be applied to other types of genomic interactions to improve prediction accuracy and interpretability.