Probabilistic Graph Models - CS 452

Project Proposal

Structured Learning on Bayesian Networks for Guide Efficiency Prediction for CRISPR-CAS13 System

Instructor: Dr. Saleha Raza

Syed Muhammad Ali Naqvi - sn07590
Ali Muhammad Asad - aa07190
Musab Kasbati - mk07811

1 Project Idea

In the world of molecular biology, CIRSPR (Clustered Regulatory Interspaced Short Palindromic Repeats) is a recent advancement in gene-editing technology. A recent variant, CRISPR-Cas13, targets RNA molecules rather than DNA, using programmable guide RNAs (gRNAs) to selectively bind and cleave target RNA sequences. This makes CRISPR-Cas13 a powerful tool for antiviral treatments, where targeting viral RNA can inhibit infections. However, determining the efficacy of a specific gRNA in targeting RNA remains a challenge. Currently, gRNA effectiveness is largely assessed through manual testing and empirical methods, requiring extensive lab work. There is no comprehensive model to predict gRNA efficiency based on known features.

In this project, we aim to explore how Bayesian Networks (BNs), coupled with structure learning, can provide insights into the features that determine gRNA efficacy. By creating a model capable of predicting gRNA efficiency, we hope to reduce reliance on manual experiments and offer a computational alternative that accelerates the development of effective gRNAs for antiviral treatments.

2 Literature Review

The CRISPR-Cas system provides a versatile and powerful tool for genome editing, with wide applications in fields such as functional genomics, immunotherapy, synthetic lethality, drug resistance, metastasis, and RNA targeting [2]. While the most well-known variant, CRISPR-Cas9, is primarily used for DNA editing, challenges remain in maintaining gene regulation efficiency. To address this, previous work employed a Bayesian Network to model the relationship between sequence features and the efficacy of the CRISPR-Cas9 system [5]. This study used structure learning and inference to predict system efficiency and applied D-Separation to analyze causal relationships, concluding that the active site of Cas9 and the location of scissile bonds corresponded to the findings from the Bayesian Network model.

In contrast, CRISPR-Cas13 is a more recent technology that targets RNA instead of DNA, offering high sensitivity and specificity for the detection of microorganisms through programmable RNA guides [4]. Despite its potential, there is little existing research that utilizes computational models, particularly Bayesian Networks, to predict the efficacy of gRNAs in CRISPR-Cas13. Current methods for determining gRNA efficiency rely heavily on empirical testing in lab settings.

Our project aims to bridge this gap by employing Bayesian Network structured learning and inference to predict the efficiency of gRNAs in the CRISPR-Cas13 system. Additionally, we seek to develop explainable models to gain a deeper understanding of the mechanisms behind CRISPR-Cas13s effectiveness. By integrating these models, we hope to move beyond purely empirical methods and provide a computational framework that can assist in the rapid development of effective gRNAs.

3 Contact Person / Domain Expert

Since this project requires knowledge in genetic engineering, especially CRISPR CAS 13 systems, Dr Faraz Ahmed from University of Melbourne, Australia is our point of contact for providing us with insights for initial modelling of the network. The domain expert would be able to provide information on some general characteristics of the guide RNAs of CRISPR Cas 13 system which

in-silico lab experiments suggests have a direct impact on the efficacy of the guide. Regardless, since the causal relationships between guide RNA characteristics and its efficacy is not clearly known, we ought to find some insights through learning a Bayesian network and share it with the Domain expert.

4 Dataset

We have obtained a dataset comprising of 245 lab-tested guide crRNAs for CRISPR Cas 13b for this project through our domain expert [3]. This dataset would help us learn the parameters of the Bayesian network. Additionally, we plan to collect more data for CRISPR Cas 13, through our contact person, datasets released online like [1], and by contacting authors who have contributed to research in this domain, to improve the accuracy of our model.

5 Our Approach / Methodology

Our approach involves constructing a Bayesian Network (BN) to model and predict the efficacy of guide RNAs (gRNAs) in the CRISPR-Cas13 system. This approach draws on methods similar to those used in [5], adapted for RNA-targeting CRISPR variants. The process is broken down into the following steps:

1. Problem Representation: We begin by representing the CRISPR-Cas13 system as a probabilistic graphical model, where the nodes represent key variables such as gRNA sequence features and their observed efficacy against target RNAs.
2. Structured Learning: The next step is to determine the structure of the Bayesian Network, i.e., the dependencies between nodes. For this, we will explore several optimization algorithms such as Repeated Hill Climbing - a heuristic algorithm that incrementally adjusts the network to minimize a scoring function, typically improving the network's fit to the data - and Evolutionary Algorithms - use a population-based search to explore a wider range of possible network structures which can be useful in avoiding local optima.

   The structure-learning process will help us identify the causal relationships between the gRNA sequence features and their efficacy.
3. Parameter Learning: Once the structure is determined, the next step is parameter learning, where we estimate the conditional probabilities between the nodes using available data. We might use the "PGMPY" library, which provides implementations for various algorithms. The learned parameters will quantify the influence of gRNA sequence features on its efficacy against target RNAs.
4. Inference / Evaluation: With the network structure and parameters in place, we will conduct inference tasks to predict the efficacy of new gRNAs. Additionally, we will use D-separation to verify the model's causal assumptions and to ensure that it aligns with current biological knowledge. If successful, this could offer new insights into the underlying mechanisms of gRNA efficacy in the CRISPR-Cas13 system.

6 Final Anticipated Outcome

We anticipate that the Bayesian Network structured learning will be able to predict the efficiency of guide RNA for the CRISPR-Cas13 system. By the end of the project we anticipate to have

an efficient BN Model, and a paper.

References

[1] X. Cheng et al. Modeling crispr-cas13d on-target and off-target effects using machine learning approaches. *Nature Communications*, 14(1):752, Feb 2023.

[2] Medina Colic and Traver Hart. Common computational tools for analyzing crispr screens. *Emerging Topics in Life Sciences*, 5, 12 2021.

[3] W. Hu et al. Single-base precision design of crispr-cas13b enables systematic silencing of oncogenic fusions. Jun 2022.

[4] Zhanchao Huang, Jianhua Fang, Min Zhou, Zhenghua Gong, and Tianxin Xiang. Crispr-cas13: A new technology for the rapid detection of pathogenic microorganisms. *Frontiers in Microbiology*, 13, 2022.

[5] Yi Yan. Efficiency prediction and mechanism discovery for the crispr-cas9 system. 05 2018.