DATA MANAGEMENT PLAN

Project Title: FINDING, FOLDING, AND ELUCIDATING FUNCTIONS OF RNA STRUCTURES IN THREE VIRAL TARGETS OF CLINICAL IMPORTANCE

Abstract:

RNA plays a crucial role in the life cycles of cells and viruses. This versatile molecule is involved in gene expression, carrying instructions for protein production, and even directly catalyzing reactions. To perform these tasks, RNA adopts complex shapes determined by both its sequence and how it folds upon itself. RNA folding is governed by specific base-pairing interactions: guanine pairs with cytosine, adenine with uracil, and occasionally, guanine with uracil (G-U). Stems formed from these pairings are the foundational building blocks of secondary structures. RNA's true versatility arises from the intricate way stems combine into higher-order three-dimensional motifs. These motifs, like the kink-turn, provide precise surfaces for interacting with other molecules. Further complexity arises from long-range and even crossing interactions, called pseudoknots, adding to the potential for RNA to create unique binding pockets and active sites. Understanding an RNA molecule's function requires knowing its structure. However, predicting RNA structure accurately remains a major challenge in bioinformatics. This is due to the hierarchical, dynamic nature of RNA folding and the vast landscape of possible conformations a single RNA sequence can adopt. Despite this, recent progress in computational methods allows us to explore how different environmental cues or interactions might alter an RNA molecule's folded state. Viruses often utilize RNA as either their primary genetic material or as critical components of their replication machinery. Due to this reliance, RNA structure offers a rich source of potential therapeutic targets. Disrupting viral RNA with drugs could prevent crucial steps in virus infection with fewer side effects than targeting human host molecules. This project aims to identify novel RNA structures within three clinically significant viruses and elucidate how those structures contribute to viral biology.

This project aims to identify novel RNA structures within three clinically significant viruses and elucidate how those structures contribute to viral biology. To achieve this, we'll employ a multi-pronged approach. First, advanced computational algorithms will be used to predict secondary and tertiary structures of viral RNAs. These predictions will incorporate any available experimental data to increase their accuracy. Furthermore, we'll move to experimental validation and characterization. Predicted RNA structures will undergo rigorous testing using cutting-edge biophysical and biochemical techniques. This will provide real-world measurements of how these structures fold under different conditions. Finally, once an RNA structure is validated, we'll conduct functional analysis to determine its precise role within the virus lifecycle. This will include assessing whether the structure is critical for processes like viral replication, packaging, or even evading the host's immune system.

The project will expand our knowledge of viral RNA biology, potentially leading to the development of new therapeutics that target RNA structures directly. These therapeutics could be more effective and less prone to viral resistance than traditional antiviral drugs. This work also has important implications for understanding RNA function in general, contributing to fundamental scientific discoveries. Effective data management is crucial for the success of this project. The following plan outlines the strategies for data generation, storage, preservation, accessibility, and security.

\

Data Generation and Types

This project will produce a diverse set of data, including:

- **FastQ files**: These will contain the raw RNA sequencing reads obtained from conducted chemical probing experiments. These fastQ files will be analyzed and the reactivity profile will be examined before been used as a constrain in the scanfold program.
- **Processed sequencing data**: This comprehensive set will comprise not only alignment files but also reactivity profiles.
- **Structural models**: This project will generate accurate 2D of secondary RNA structure predictions alongside intricate 3D tertiary structures. These models will be derived from state-of-the-art cryo-electron microscopy techniques coupled with advanced computational modeling.
- **Cryo-EM datasets**: These datasets will feature meticulously captured 2D projection images and intricately reconstructed 3D representations. They will serve as invaluable tools for visualizing and understanding the complex architecture of vital RNA complexes.
- **Analysis Scripts**: Our team will develop bespoke analysis scripts tailored to each stage of data processing, structure prediction, and analysis. These meticulously crafted codes will ensure robustness and efficiency in exploring the intricacies of the generated datasets.

Data Storage and Preservation

FastQC: the project involves the generating of Fastq files resulting from the chemical probing of RNA and subsequent sequencing using Illumina technology (iSeq). These files containing raw sequences are analyzed to determine the reactivity profile before being used in the scanfold program. To accommodate the large data set that will be generated from this program, the team has been allocated a dedicated space on the Iowa State University High-Performance Computing (HPC) system. By centralizing the storage of Fastq files within the lab's designated area on the HPC cluster, accessibility, security, and scalability are ensured. This strategic decision facilitates seamless access to the primary data for downstream analyses. The HPC system will employ automated data integrity checks and regular backups to safeguard the data against loss or corruption.

Cryo-EM data: During the course of the project, structures of significant importance will be visualized using the Cro EM facility in the institute. The availability of Cryo-EM visualization facilities at Iowa State University further enriches the research environment, facilitating real-time exploration and interpretation of complex structural data. These high-resolution structural data provide invaluable insights into the spatial arrangement and conformational dynamics of selected RNA structures, shedding light on their functional roles and interactions. To manage the voluminous Cryo-EM datasets effectively, the images will be stored on encrypted hard drive for easy access and on the HPC cluster alongside the Fastq files. By consolidating all research data within a unified infrastructure, synergy is fostered, enabling interdisciplinary collaborations and holistic data-driven investigations.

Analysis Script: the development of analysis scripts constitutes an integral component of the research workflow. These scripts, tailored to specific analytical tasks and computational algorithms, serve as the backbone of data processing and interpretation will be saved on the lab github page, moss-lab (github.com). A decentralizable approach not only promotes transparency and reproducibility but also fosters knowledge dissemination and community engagement. Integration of code testing, validation procedures, and usage examples will be included within the GitHub repository to ensure quality and usability. The lab will encourage external contributions to promote collaboration and wider adoption of their scripts

Laboratory/Assay Result: A version control of all lab result (qPCR, Dual Luciferase assay reporter, gel images, PAGE etc), and cell line passages will be documented and saved on cybox using the version control system and that is accessible to all laboratory members. This will ensure that the data generated are updated regularly. Lab result documentation will follow standardized formatting, ensuring consistency across experiments. Archiving outdated versions will provide a historical record for reference.

Accessibility and Sharing

Secure shared access will be established within our research team. To promote scientific discovery, we plan to share our findings with the broader research community. Raw FastQ files will be deposited in public repositories like the Sequence Read Archive (SRA). Structural models will be submitted to the Protein Data Bank (PDB) or other relevant databases. Analysis scripts will be made publicly available on platforms like GitHub.

Data Security

We will adhere to all security protocols established by the HPC and handle any sensitive data according to specific ethical guidelines. Primary storage will utilize Iowa State University's High-Performance Computing (HPC) cluster, with space allocated specifically for our lab. To ensure data integrity and accessibility, we will meticulously organize our data using clear directory structures, descriptive naming conventions, and appropriate metadata standards. Furthermore, version control systems will be used for scripts and analysis pipelines, enabling reproducibility, and tracking of changes. Regular backups to secondary locations will provide redundancy and safeguard against potential data loss.

In conclusion, effective data management is the very important to achieve a successful and reproducible research, thus, by implementing robust strategies for handling diverse data types, including Fastq files, Cryo-EM structures, and analysis scripts, the I will ensure that the integrity, accessibility, and scalability of research data has been maintained. Through the synergistic integration of computational resources, visualization tools, and collaborative platforms, such as the Iowa State University HPC cluster and GitHub repository, the research project exemplifies a holistic approach to data-driven discovery. By embracing these principles of data management excellence, the research team advances scientific knowledge and contributes to the collective quest for biomedical innovation and societal impact.