**STRUCTURAL DYNAMICS AND FUNCTIONAL ROLE OF DOMAINS IN RNA-BINDING PROTEINS: A FOCUS ON DISORDERED REGIONS**

Thesis submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment of the requirements

for the award of the degree

of

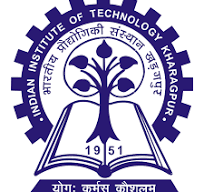
**Interdisciplinary Dual Degree M.Tech. in Biotechnology and Biochemical Engineering/Engineering Entrepreneurship**

*by*

**ABHINAY KUMAR PANDEY**

Roll No.: 21BT3EP11

*under the guidance of* **Prof. Ranjit Prasad Bahadur**

****

**DEPARTMENT OF BIOSCIENCE & BIOTECHNOLOGY**

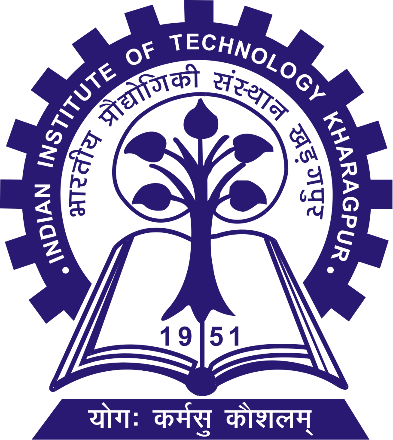
**INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR**

**KHARAGPUR 721302**

**November 2024**

**DEPARTMENT OF BIOSCIENCE AND BIOTECHNNOLOGY**

**INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR**

**KHARAGPUR- 721302, INDIA**

**CERTIFICATE**

This is to certify that the project entitled “**Structural Dynamics and Functional Role of Domains in RNA-Binding Proteins: A Focus on Disordered Regions’’** submitted by **Abhinay Kumar Pandey** (Roll No. 21BT3EP11) to the Indian Institute of Technology Kharagpur towards partial fulfillment of requirements of the award of Interdisciplinary Dual Degree Master of Technology in Biotechnology and Biochemical Engineering/Engineering Entrepreneurship is a record of bona fide work carried out by him under my supervision and guidance during Autumn Semester, 2024-2025.

Date: November 26, 2024

Place: Kharagpur

Prof. Ranjit Prasad Bahadur

Department of Bioscience and Biotechnology

Indian Institute of Technology Kharagpur

Kharagpur- 721302, India

**ACKNOWLEDGEMENT**

I would like to express my heartfelt gratitude to **Prof. Ranjit Prasad Bahadur**, my supervisor, for his expert guidance, invaluable insights, and continuous support throughout this research. His encouragement has been a driving force behind the successful completion of this project.

I am profoundly thankful to **Mr. Swarnadip Mitra**, my mentor, for his exceptional dedication, patience, and extensive assistance throughout my research. His invaluable feedback, constructive suggestions, and hands-on guidance have been pivotal in shaping this work and overcoming challenges during the project.

I also wish to thank the **Department of Bioscience & Biotechnology, IIT Kharagpur**, for providing the resources and academic environment essential for this research.

Finally, I am deeply grateful to my family and friends for their unwavering support and motivation throughout this journey.

**Abhinay Kumar Pandey**Roll No.: 21BT3EP11

**CONTENTS**

1. **LIST OF ABBREVIATIONS**
2. **SUMMARY**
3. **INTRODUCTION**
4. **OBJECTIVES**
5. **MATERIALS & TOOLS**
6. **METHODS**
7. **RESULTS & CONCLUSION**
8. **DISCUSSION**
9. **REFERENCES**

# **LIST OF ABBREVIATIONS**

**RBP**: RNA-Binding Protein

**RBD**: RNA-Binding Domain

**IDR**: Intrinsically Disordered Region

**PDB**: Protein Data Bank

**UniProt**: Universal Protein Resource

**FASTA**: Fast Alignment Search Tool

**IUPred3:** Intrinsically Unordered Protein Regions Prediction

**XML**: Extensible Markup Language

**RRM**: RNA Recognition Motif

**KH:** K Homology

**mRNA**: Messenger RNA

# **SUMMARY**

This project aimed to explore the intrinsically disordered regions (IDRs) within human RNA-binding proteins (RBPs) and understand their structural and functional significance. We started with the dataset, which initially contained 1,433 UniProt IDs of RBPs. After filtering the dataset to focus only on proteins with solved 3D structures (those with associated PDB IDs), we narrowed it down to 658 unique UniProt IDs. These proteins were selected for further analysis due to their availability of experimentally resolved structures, which are essential for studying the relationship between sequence and structure.

Using Python scripting, we automated the retrieval of FASTA sequences for each protein, followed by the prediction of disordered regions using the IUPred3 tool. The tool provided long and short disorder scores, which helped identify regions of flexibility and disorder within each protein. These scores were then used to analyze the presence of IDRs, crucial for understanding the dynamic behavior of these proteins.

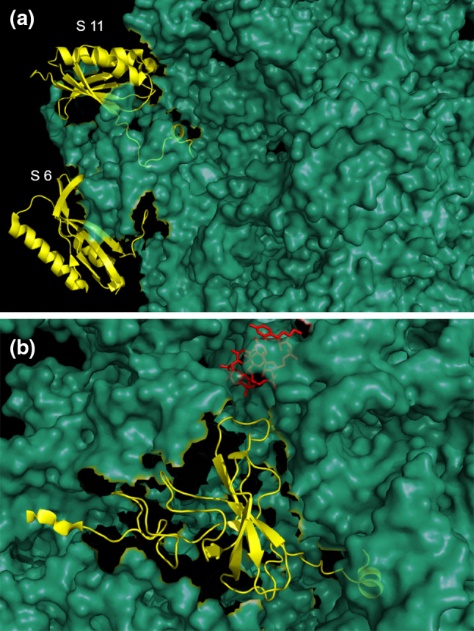
In addition to predicting disordered regions, we extracted domain information from XML files provided by UniProt for each protein. This allowed us to annotate the disordered regions with functional domain data. In total, we identified 246 unique domain files, shedding light on how disordered regions intersect with known protein domains. This analysis revealed the pervasive presence of disorder in RBPs and its potential functional roles.

Looking ahead, the project sets the stage for deeper exploration into how these disordered domains interact with RNA, especially in the context of disease mechanisms. By leveraging solved PDB structures, we aim to model these interactions and understand how the misregulation of protein-RNA interactions can lead to diseases like neurodegenerative disorders and cancers. Understanding these dynamics will not only enhance our knowledge of protein functionality but could also open doors for therapeutic strategies targeting misregulated RBPs.

# **INTRODUCTION**

## **RNA-Binding Proteins (RBPs):**

RNA-binding proteins (RBPs) are vital players in regulating RNA processes like splicing, transport, translation, and stability. They influence post-transcriptional gene regulation by interacting with various RNA forms, such as mRNA, non-coding RNA, and ribosomal RNA, thus ensuring proper cellular functioning and adaptability to environmental changes.

**** *Fig.1:*

Ribosomal proteins of 30S subunit (PDB ID: 1N34).

1. Small subunit proteins S6 and S11 are shown in *yellow cartoon* and other small subunit proteins are shown in *green surface*.

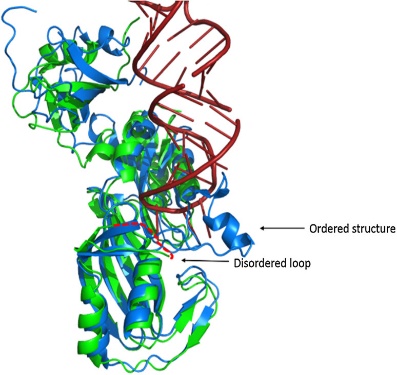
**b.** Small subunit protein S12 (shown in *yellow cartoon*) is interacting with mRNA (shown as a *red fragment*) through its disordered extension, the other small subunit proteins are shown in *green surface*.

## **Domains in RBPs:**

RBPs typically contain RNA-binding domains (RBDs), specialized regions that allow for precise recognition of RNA sequences. Common RBDs, like RNA recognition motifs (RRMs) and K homology (KH) domains, are essential for the specific interactions RBPs have with RNA molecules, guiding RNA processing events.

## **Instrinsically Disordered Regions (IDRs) in RBPs:**

In contrast to structured domains, RBPs can also feature intrinsically disordered regions (IDRs), which lack a fixed three-dimensional structure. These regions offer flexibility, enabling RBPs to engage with multiple RNA targets and other proteins. This adaptability is crucial for regulating cellular functions, but it can also lead to problems, especially in disease states.

*Fig.2:*Superposed structure of TruB with its partner RNA in bound (*coloured in blue cartoon*, PDB ID: 1R3E) and unbound (*coloured in green cartoon*, PDB ID: 1R3F) conformations.The disordered thumb loop (*red dashed lines*) of TruB undergoes conformational transitions and become ordered upon binding with its partner RNA

## **Structure of Normal Domains and Disordered Regions:** The structures of RNA-binding domains, like RRMs, have stable secondary structures like β-sheets and α-helices that enable specific RNA binding. IDRs, on the other hand, lack a defined structure, that allows dynamic interactions but increases susceptibility to instability.

*Fig.3* In RBP having uniport id A0A0A0MR66 which has been merged with P91875 at domain RRM 1 which is from residue 129-209, residue position 133 Leu(*red coloured marked in left image*) is disordered and position 181(*red coloured marked in right image*) is disordered.

## **Impact of IDRs on Cellular Function and Disease:** Intrinsically disordered regions (IDRs) in RBPs provide flexibility but can also cause dysfunction. Mutations in these regions can impair protein-RNA interactions, leading to cellular stress and diseases like ALS, SMA, and fragile X syndrome. This highlights the fine balance between flexibility and stability in maintaining proper cellular function.

## **OBJECTIVE AND SCOPE OF STUDY**

The objective of this project is to identify and analyze domain-disordered sequences within RNA-binding proteins (RBPs), with a focus on understanding how these intrinsically disordered regions (IDRs) contribute to RNA-protein interactions. These disordered sequences play critical roles in RNA regulation and processing, including splicing, transport, and translation. The project aims to uncover how disruptions in these IDRs can affect RNA interactions and contribute to diseases like neurodegenerative disorders and cancers.

The long-term goal is to identify how mutations or misfolding in these disordered regions lead to disease by disrupting RNA binding and regulatory functions, ultimately contributing to disease mechanisms. This insight could aid in developing targeted therapeutic strategies to restore proper RNA-protein interactions.

# **MATERIALS & TOOLS**

## **Datasets:**

A curated listing of human RNA-binding proteins (RBPs) as reported by Gesrtberger et al with UniProt IDs and PDB IDs, ensuring selection of proteins with experimentally resolved structures. This served as the primary reference for analyzing intrinsic disorder.

UniProt Database: Provided FASTA sequences and XML files for each UniProt ID, including annotated sequence data and domain information for human RBPs.

Protein Data Bank (PDB): Source of 3D structural data. Only proteins with associated PDB IDs were considered, allowing insights into sequence-structure relationships within RBPs.

## **Tools and Software:**

**Python:** A versatile programming language that enabled automation, tool integration, and efficient handling of complex datasets.

**Libraries:**pandas:Facilitated efficient data manipulation, filtering, and organization in tabular formats.  
openpyxl:Enabled reading and writing Excel files for structured data storage and analysis.  
wget: Automated downloading of FASTA sequences and XML files from UniProt and PDB databases.os: Managed folder structures and file paths for systematic organization of project data.

IUPred3: IUPred3 predicts intrinsically disordered regions by estimating **residue-specific free energy (ΔG)** changes. It evaluates how amino acids interact within a structured protein environment, identifying regions with destabilized interactions as disordered. This approach relies on pairwise interaction potentials derived from the sequence rather than full structural data, making it efficient for disorder prediction.

**Long disorder scores:** These scores are suited for detecting extended regions of disorder that might play roles in protein-protein or protein-nucleic acid interactions.

**Short disorder scores:** These scores provide a finer, more localized view of disorder, useful for identifying small flexible regions.

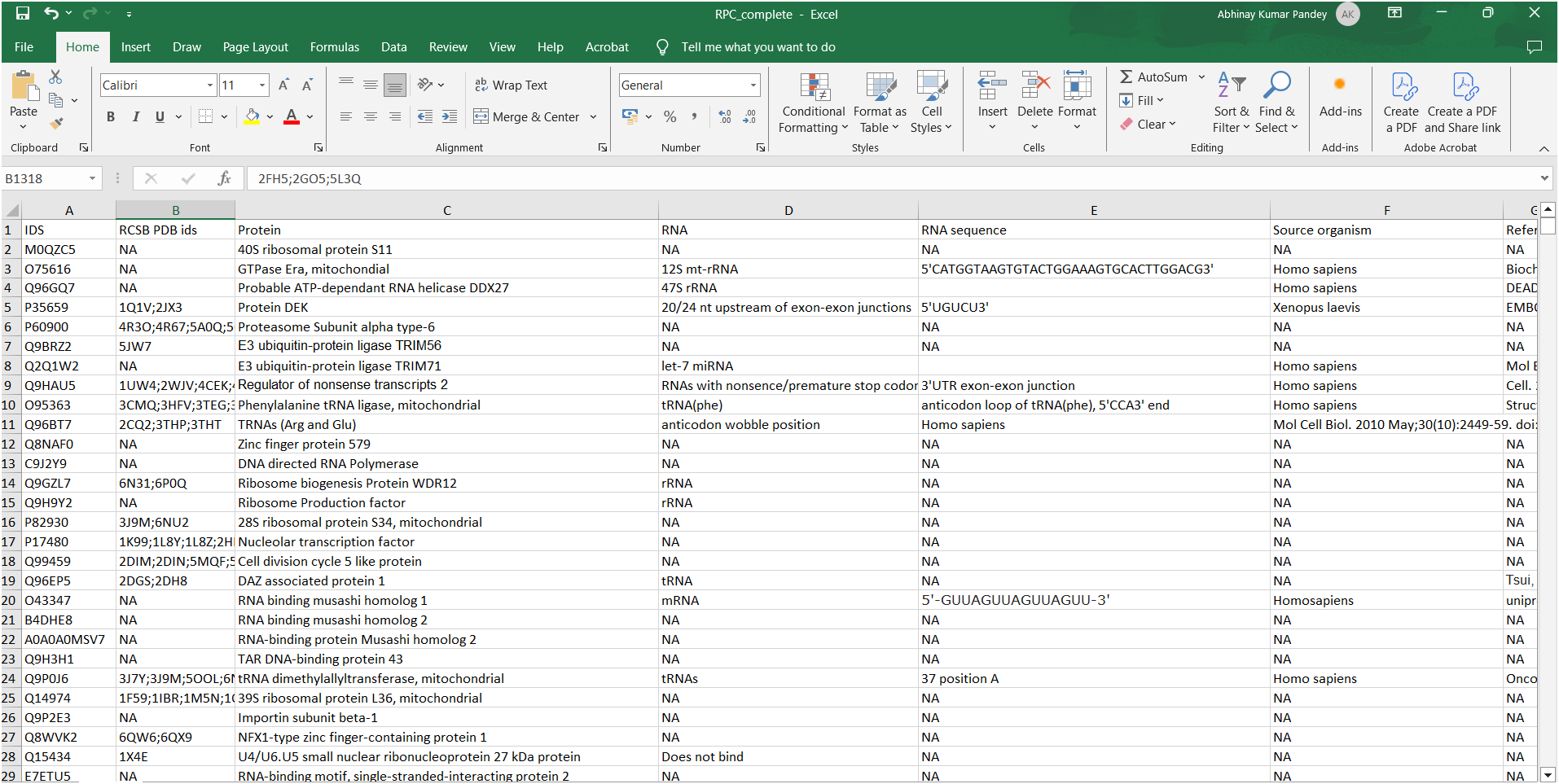
## **Annotation and Data Organization:**

XML Files from UniProt: Provided domain annotations, including start/end positions and descriptions. These annotations were linked to disorder predictions.

Domain Information Excel Sheets: Generated for each unique domain, these sheets consolidated UniProt IDs, domain positions, sequences, and IUPred scores, enabling clear visualization of disorder patterns across proteins.

## **METHODS**

### 1. Dataset Preparation and Filtering

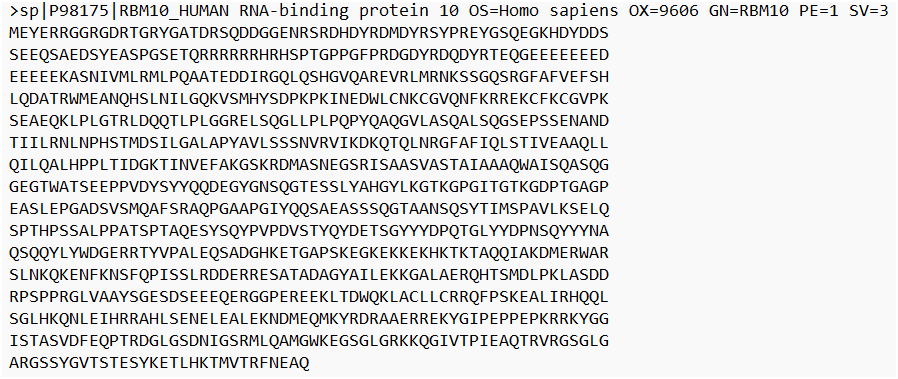
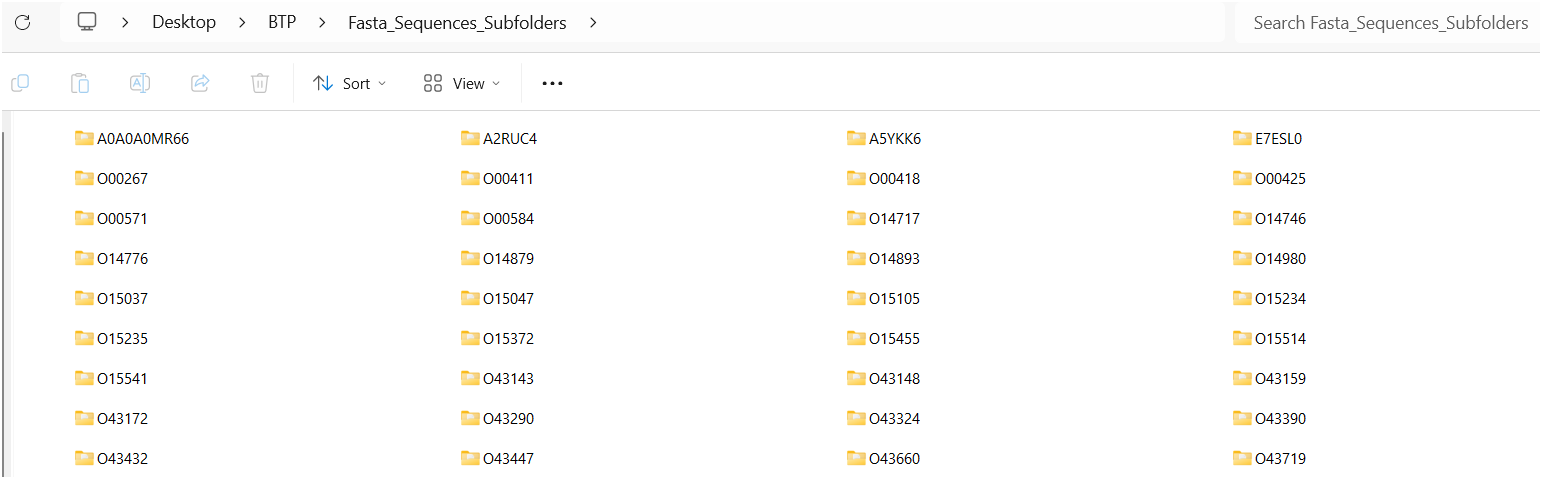
We started with a dataset containing 1,433 UniProt IDs of RNA-binding proteins (RBPs). Using Python with the panda**s** and openpyxl libraries, we filtered the dataset to include only proteins with associated PDB IDs, focusing on those with solved structures for further analysis..  


### Proteins with unique UniProt IDs and associated PDB IDs (indicating experimentally resolved structures) were selected, resulting in a refined dataset of 658 unique UniProt IDs. This filtering step ensured the dataset focused exclusively on proteins with reliable and experimentally validated structural data.

### 2. Retrieving Protein Sequences

Using the filtered UniProt IDs, protein sequences were systematically downloaded:

UniProt fasta sequences: Fetched using the command:  
wget [https://rest.uniprot.org/uniprotkb/{uniprot\_id}.fasta](https://rest.uniprot.org/uniprotkb/%7buniprot_id%7d.fasta).

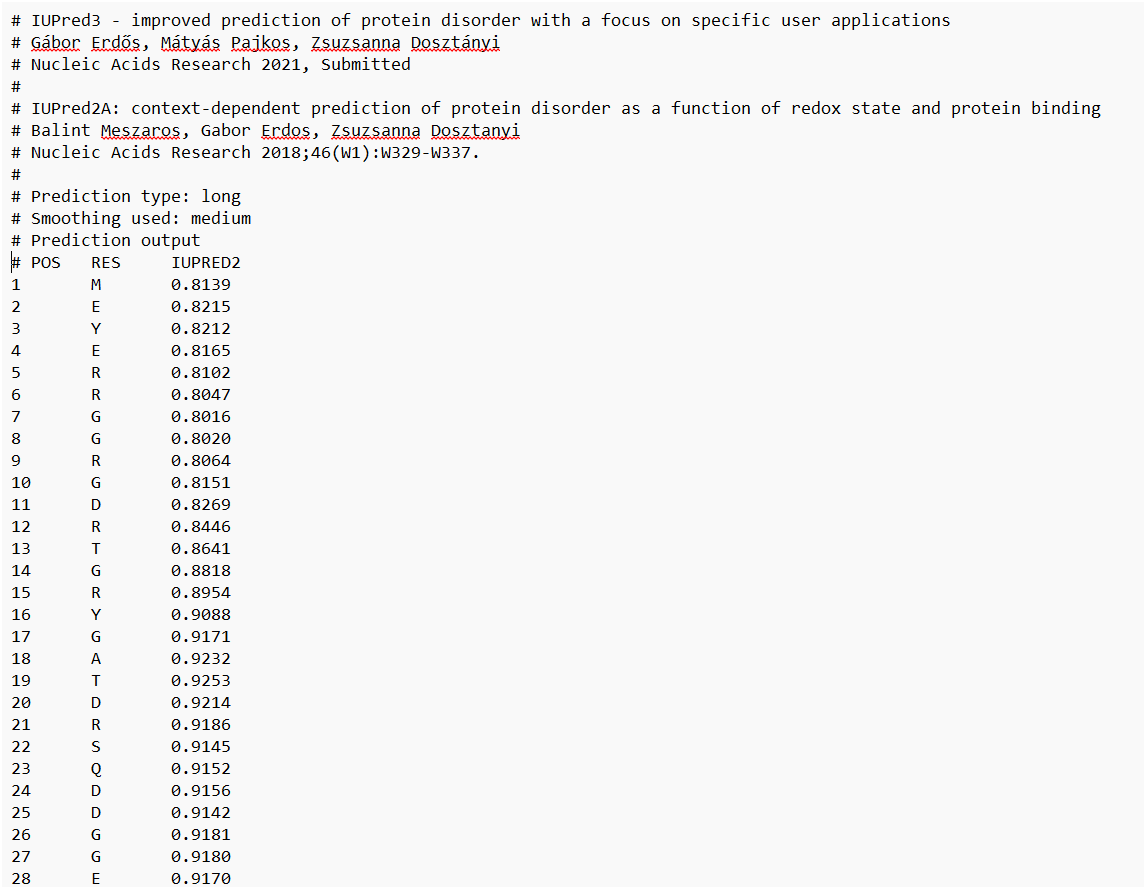
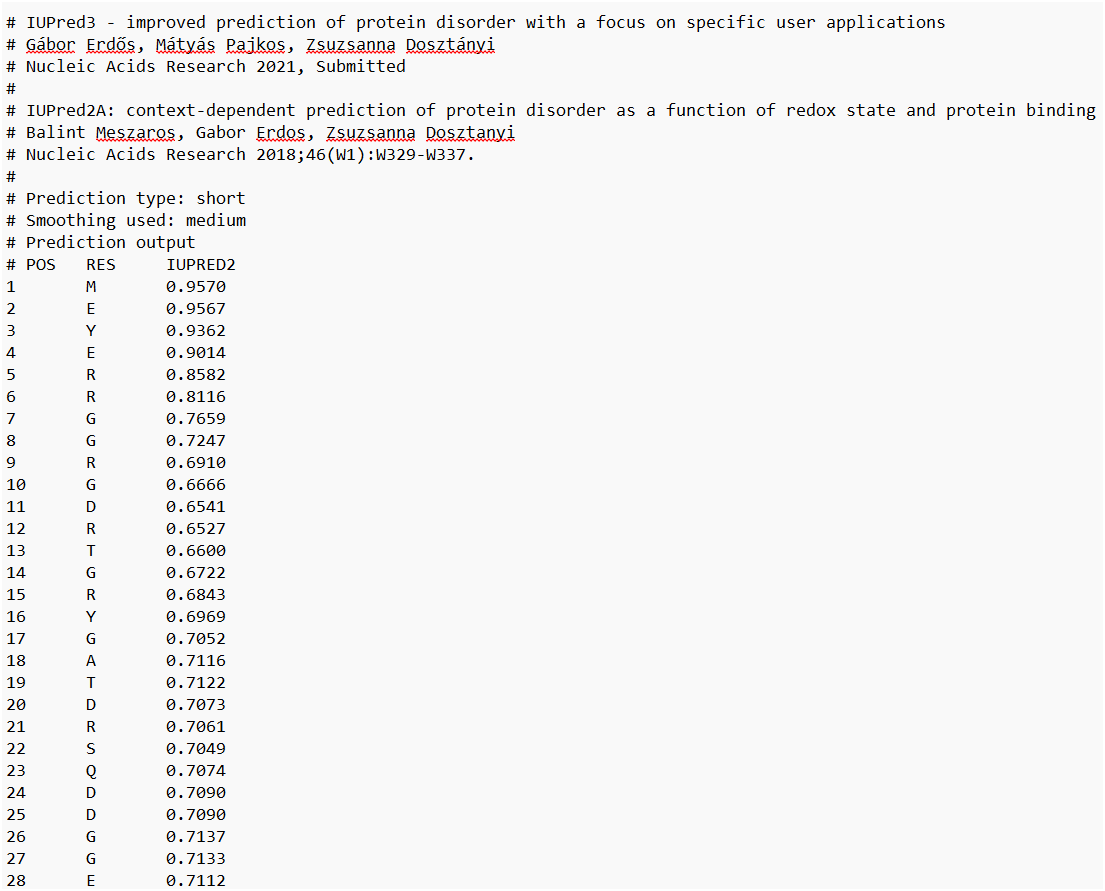
PDB fasta sequences: Retrieved using:  
wget [https://www.rcsb.org/fasta/entry/{pdb\_id}](https://www.rcsb.org/fasta/entry/%7bpdb_id%7d)  
Below is the example of one fasta sequence for uniport id A0A0A0MR66: A total of 658 unique UniProt IDs were identified. A dedicated folder was created for each uniport ID, containing its canonical protein sequence and corresponding structural sequences: 

### 3. Intrinsic Disorder Prediction

IUPred3 was employed to predict regions of intrinsic disorder across all protein sequences.

Disorder scores were calculated using:

python3 iupred3.py {fasta\_file} long/short.  
Separate files for long and short disorder scores were generated for each sequence.  
eg. For uniport id A0A0A0MR66 , for each amino acid we have short and long run iupred scores as:



#### **Insights:**

Scores > 0.5 indicate disordered residues.

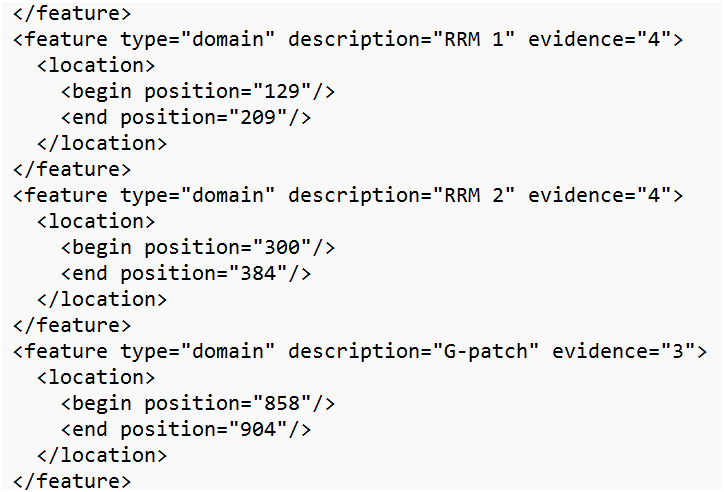
Scores < 0.5 correspond to ordered residues.  
This provided a detailed residue-level disorder profile.

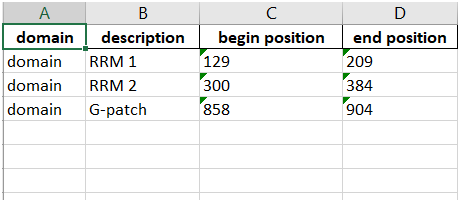
### 4. Domain Annotation Extraction

UniProt XML files were downloaded using:  
wget [https://rest.uniprot.org/uniprotkb/{uniprot\_id}.xml](https://rest.uniprot.org/uniprotkb/%7buniprot_id%7d.xml).

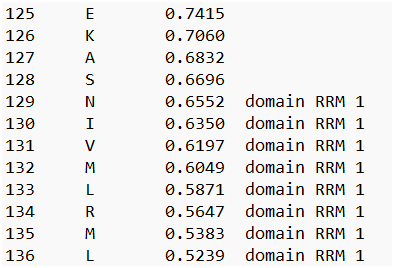
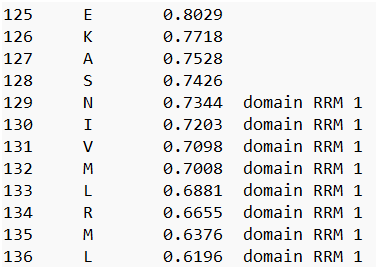


#### **Domain Details:**

To extract domain descriptions and boundaries from XML files, a Python script parses lines starting with <feature type="domain". It retrieves description, <begin position>, and <end position> details from the respective tags, providing essential information for identifying functional regions in RNA-binding proteins.

For each UniProt ID, domain data (name, start, and end positions) were saved into domain-specific Excel sheets for easy reference.  


### 5. Integrating Disorder and Domain Data:

The IUPred score files were enriched with domain information by appending a column that annotated residues within domain regions. For example:

This step connected structural disorder predictions to functional annotations.

### 6. Unique Domain Analysis

All unique domains across the dataset were identified. For each domain, a dedicated sheet was created summarizing:

UniProt IDs where the domain appears.

Start and end positions of the domain in each protein.

Domain sequence.

# Long and short IUPred disorder scores. eg. RRM1 domain is found in multiple RBPs:

# **RESULTS & CONCLUSION**

The project began by filtering 1,433 human RNA-binding proteins (RBPs) to focus on 658 UniProt IDs with experimentally solved structures, ensuring that our analysis was based on proteins with reliable 3D structural data. From this subset, we identified 246 unique domains that are likely involved in RNA interactions, a hallmark of RNA-binding proteins. These domains play critical roles in cellular processes, and disruptions in their function may contribute to cellular dysfunction and various diseases, particularly those involving the misregulation of RNA-binding proteins.

By linking intrinsically disordered regions (IDRs) to specific functional domains, we’ve established a valuable resource for understanding how these domains contribute to RNA-binding protein function. This foundational dataset can guide future studies aimed at elucidating the precise molecular mechanisms by which these domains interact with RNA and how their dysregulation leads to diseases. Our work lays the groundwork for exploring potential therapeutic targets, offering insights into the broader implications of RNA-binding protein dysfunction in cellular health and disease.

## **DISCUSSION**

In this study, we investigated the relationship between intrinsically disordered regions (IDRs) and functional domains in RNA-binding proteins (RBPs). By analyzing domain annotations and disorder prediction scores, we observed that a substantial proportion of domain regions in RBPs exhibit disordered behavior. This highlights the flexibility and dynamic nature of these regions, crucial for their interactions with RNA and involvement in cellular processes.

Our findings suggest that disordered regions play a significant role in the functionality of RBPs, as they facilitate protein-RNA interactions, particularly in RNA processing and gene regulation. On average, a substantial portion of these domains demonstrates disordered behavior, supporting the hypothesis that IDRs are key drivers of molecular recognition in RBPs.

Looking ahead, we plan to leverage solved PDB structures to explore how these disordered domains interact with RNA. This will provide deeper insights into the molecular mechanisms of RNA metabolism and its regulation. Furthermore, this research could help us better understand the implications of IDR misregulation in human diseases, such as neurodegenerative disorders and cancers, where RBPs and their disordered regions are often implicated. By integrating structure-function relationships, we aim to contribute to the development of therapeutic strategies targeting these dynamic protein regions, offering potential avenues for intervention in diseases linked to RBP dysfunction.

## **Future Works:**

1. A multi-parameter correlational analysis between domain versus non-domain structural and functional residue level and motif level characteristics with degree of intrinsic disorder.

2. Ranking of aforementioned parameters in order of significance in order to determine necessary features that will prove to be effective in a potential curation of a intra-domain protein intrinsic disorder predictor.

3. Obtaining representative protein structures for each interactive domain complexed with their most evident and frequent binding partner and observation of binding dynamics in such complexes.

4. Monitoring of the influence of aforesaid significant physico-chemical domain and non-domain parameters on above-mentioned protein-substrate binding interactions and analysis of incremental and decremental changes in the intrinsic disorder landscape.

5. A comparative follow up study of the stated parametric behaviour in protein interactions from the perspective of inter-domain protein regions as opposed to its previously carried out and currently undergoing intra-domain counterpart.

**REFERENCES:**

1. Basu S, Bahadur RP. A structural perspective of RNA recognition by intrinsically disordered proteins. Cell Mol Life Sci. 2016 Nov;73(21):4075-84. doi: 10.1007/s00018-016-2283-1. Epub 2016 May 26. PMID: 27229125; PMCID: PMC7079799.

2. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014 Dec;15(12):829-45. doi: 10.1038/nrg3813. Epub 2014 Nov 4. PMID: 25365966; PMCID: PMC11148870.

3. IUPred3. (2023). Predicting intrinsically disordered regions in proteins. Retrieved from http://iupred3.elte.hu/

4. Dunker AK, Oldfield CJ, Yang J, et al. Intrinsically disordered protein. J Mol Graph Model. 2008 Sep-Oct;26(2): 314-27. doi: 10.1016/j.jmgm.2007.12.003. PMID: 18433723.

5. Chavali PL, Guharoy M, Chakrabarti P. DisProt: A database of protein disorder. In: Watterson D, editor. Computational Biology. Berlin: Springer; 2017. p. 75-88. doi: 10.1007/978-3-662-49734-15

6. Huang S, Liu F, Chen Y. The functional significance of protein disorder in the context of RNA-protein interactions. Int J Mol Sci. 2019 Jan;20(4):867. doi: 10.3390/ijms20040867. PMID: 30834537; PMCID: PMC6470630.