University of Padova

Department of Mathematics

Master Thesis in Data Science

Assessment of Missense Intolerant Region in Intrinsically Disordered Proteins (IDPs)

Supervisor
Professor Emanuela Leonardi
University of Padova

Co-supervisor

Master Candidate
Sina Rasouli

Academic Year
2023-2024

Dedication.

This thesis is dedicated to my family and partner for their unwavering support.

To my mother, whose dedication to my education inspired me.

To my father, whose financial and logical support strengthened me.

To my brother, whose support means the world to me.

To my partner, whose patience and understanding made this journey possible.

I am eternally grateful for their love, encouragement, and belief in me.

Abstract

Genes that are crucial for the function of an organism are depleted of disrupting variants in natural populations, whereas non-essential genes tolerate their accumulation. Next-generation sequencing (NGS) of the general population has enabled comprehensive coverage of the human genome, identifying single nucleotide variants (SNVs) at an impressive density of two SNVs per three base pairs. This has demonstrated that regions intolerant to variations are important for gene function and usually map to structural domains. However, the presence and role of regions intolerant to variations in non-globular domains, such as intrinsically disordered regions (IDRs), remain to be investigated. The aim of this study is to determine the distribution of the Missense Intolerance Ratio (MIR), a measure of regional intolerance to missense variation, in intrinsically disordered proteins (IDPs) and to explore how these regions relate to protein functions.

We analyzed the content of missense intolerant regions (MIRs) in a set of human proteins retrieved from DisProt, the major manually curated repository of IDPs. The matched MIRs were then correlated with the presence of IDP features retrieved from MobiDB, a resource that integrates predictions and functional annotations of protein disorder and mobility. Additionally, the matched MIRs were analyzed for the presence of disease variants reported in the ClinVar database, which collects variants associated with human diseases.

Our results indicate that while MIRs are enriched in protein domains, a substantial proportion is also present within IDRs. Although no significant correlation was found between MIRs and other protein features, MIRs were frequently associated with disease-related variants. These findings highlight the functional importance of MIRs in both ordered and disordered protein regions. However, limitations in dataset coverage and methodological assumptions necessitate further investigation to fully elucidate the role of MIRs in IDPs.

Contents

Listing of figures

Listing of tables

Listing of acronyms

API	Application Programming Interface
CLT	Central Limit Theorem
DNA	Deoxyribonucleic acid
GATK	Genome Analysis ToolKit
IDP	Intrinsically Disordered Protein
IDR	Intrinsically Disordered Region
INDELs	Insertions or Deletions
IPMV	Identified Pathogenic Missense Variant
MIR	Missense Intolerant Region
MTR	Missense Tolerance Ratio
NMD	Nonsense-Mediated decay
PTM	Post Transnational Modification
SLiM	Short Linear Motif
VCF	Variant Call Format

1
Introduction

1.1 Motivation

As technological advancements have given rise to new methodologies for medicine prescriptions, exome sequencing is frequently used to guide more personalized diagnosis and treatment for many genetic diseases including cancer[1, 2, 3]. While this has led to more information about pathogenic variants[4, 5, 6], many variants still remain with uncertain significance. Many in silico predictors are used to prioritize likely candidates for each category , but it remains a major challenge to distinguish pathogenic variants from benign ones[7].

Large exome[8] and genome[9] sequencing projects have presented references of variation across the human genome providing the means to measure patterns of variability within genes [10, 11]. It has been demonstrated previously that measuring depletion of standing variation within genes can be used to identify novel disease-associated genes [10, 11]. With the current sample sizes of sequenced individuals, measurements in depletion of variation at a regional level within these genes has begun and many measures have been presented [12].

The Missense Tolerance Ratio (MTR) is a measure of regional intolerance to missense variation, and can capture this regional level of information [13]. ”The MTR is a direct measure of purifying selection of missense variation within a gene calculated as a ratio between the observed proportion of missense variants compared to an expected proportion, estimated under the assumption of no selection occurring in that sequence context. A sliding window summation is used to provide accurate regional measurements."[12] According to this study [13] the regions measured as intolerant to missense variation are significantly enriched for pathogenic missense variants in epilepsy genes.

According to previous studies, mechanisms associated with the pathogenesis of missense variants often correlate with the three-dimensional structure of proteins [14, 15, 16] and, for some disease-associated genes, mutations appear to cluster within specific regions [17, 18, 19, 20, 21, 22, 23, 24, 25]. More systematic analyses have identified coding DNA sub-regions intolerant to missense variants [26, 14] and domains in protein families enriched in variants associated with disease [27, 28].

Furthermore, within many genes, pathogenic missense variants tend to cluster within specific domains or regions of the encoded proteins, whereas most loss-of-function variants do not [23, 27] with the exception of the penultimate and last exons where premature termination codons can escape nonsense-mediated decay (NMD) [29]. Here in this research it is intended to find out the MIRs inside all the human proteins in Disprot Database [30], the major manually curated repository of IDPs, finding the characteristics of these MIRs and see if there is a correlation between MIRs with IDRs.

1.2 Genetic Variations

Deoxyribonucleic acid (DNA) is the molecule that carries the genetic instructions for the development, functioning, growth, and reproduction of all known organisms and many viruses [31]. It's essentially the hereditary material passed down from parents to offspring.

DNA is a complex molecule with a unique structure often referred to as a double helix. It resembles a twisted ladder with two long strands of sugar and phosphate molecules forming the sides, and pairs of nitrogenous bases forming the rungs that connect the sides [31]. These bases, adenine (A), thymine (T), guanine (G), and cytosine (C), are the essential components of the genetic code. The specific order of these bases determines the instructions for building and maintaining an organism (see figure 1.1).

The blueprint of life, DNA, is not a static instruction manual. It harbors a captivating diversity within individuals and across populations, known as genetic variations. These variations arise from a fascinating interplay of evolutionary forces and random events, shaping the biological tapestry of life [33]. Understanding the types, impact, and distribution of these variations is crucial in deciphering human health and disease. Genetic variations can arise through various mechanisms:

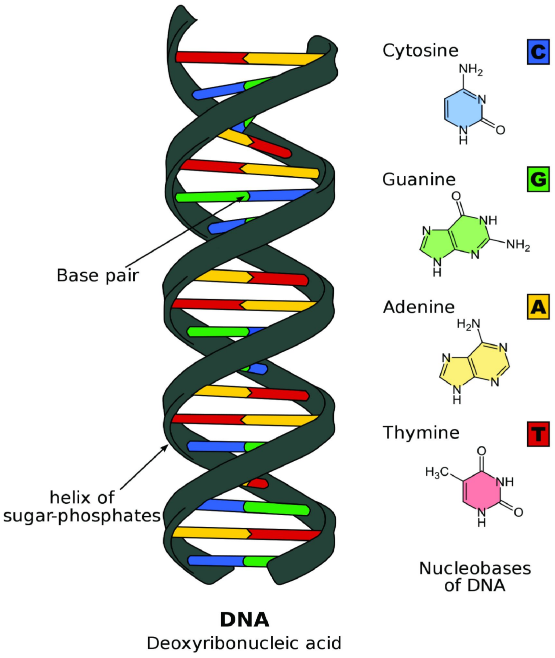

Figure 1.1: Deoxyribonucleic Acid Structure(DNA) is depicted as a double helix, resembling a twisted ladder. The ”rungs” of the ladder are formed by pairs of nitrogen-containing molecules called nucleobases(adenine (A), guanine (G), cytosine (C), and thymine (T))[32].

- **Mutations:** Spontaneous errors during DNA replication or damage from environmental factors like ultraviolet radiation can introduce changes in the DNA sequence. These mutations can be point mutations, involving a single nucleotide substitution, or larger insertions or deletions (INDELs) of nucleotides [34].

- **Recombination:** During sexual reproduction, chromosomes exchange genetic material, creating new combinations of alleles in offspring. This process shuffles existing variations and fuels genetic diversity within populations [33].

- **Horizontal gene transfer:** In some organisms, genetic material can be transferred directly between unrelated individuals, introducing novel variations not present in the ancestral lineage [35].

These mechanisms constantly introduce new variations into the gene pool, providing the raw material for evolution by natural selection. Beneficial variations that enhance an organism’s ability to survive and reproduce in a specific environment become more frequent over generations, leading to adaptation and the emergence of new species [36].

1.3 A Spectrum of Variation

Genetic variations can be broadly classified based on their location within a gene and the resulting changes they introduce:

- **Exonic vs. Intronic:** Variations can occur within the coding region of a gene (exon) or in the non-coding regions (introns) separating exons. Exonic variations often have a more significant impact on protein function compared to intronic variations, as they directly alter the instructions for protein synthesis [37].

- **Nucleotide Substitutions, Insertions, and Deletions (INDELs):** The most basic variations involve single nucleotide changes (point mutations), insertions of additional nucleotides, or deletions of existing sequences. These alterations can have varying effects depending on their location and type [33] (see Figure 1.2).

- **Missense vs. Synonymous vs. Frameshift:** Point mutations can further be categorized by their effect on the protein sequence. Missense mutations change a single amino acid, while synonymous mutations alter the DNA code without affecting the amino acid sequence. Frameshift mutations introduce or remove nucleotides, disrupting the reading frame and often leading to nonfunctional proteins [34] see Figure 1.3.

1.4 Databases of Variants

Two valuable resources play a central role in characterizing genetic variations which we are going to introduce them here.

1.4.1 ClinVar

This public database aggregates information on the relationship between human variations and human health. It provides classifications for variations based on their **clinical significance**, aiding researchers and clinicians in interpreting the potential impact of variants [39].

1.4.2 GnomAD

Genome Aggregation Database offers a global view of human genetic variation by capturing the frequency of variants across diverse populations. This information helps distinguish between

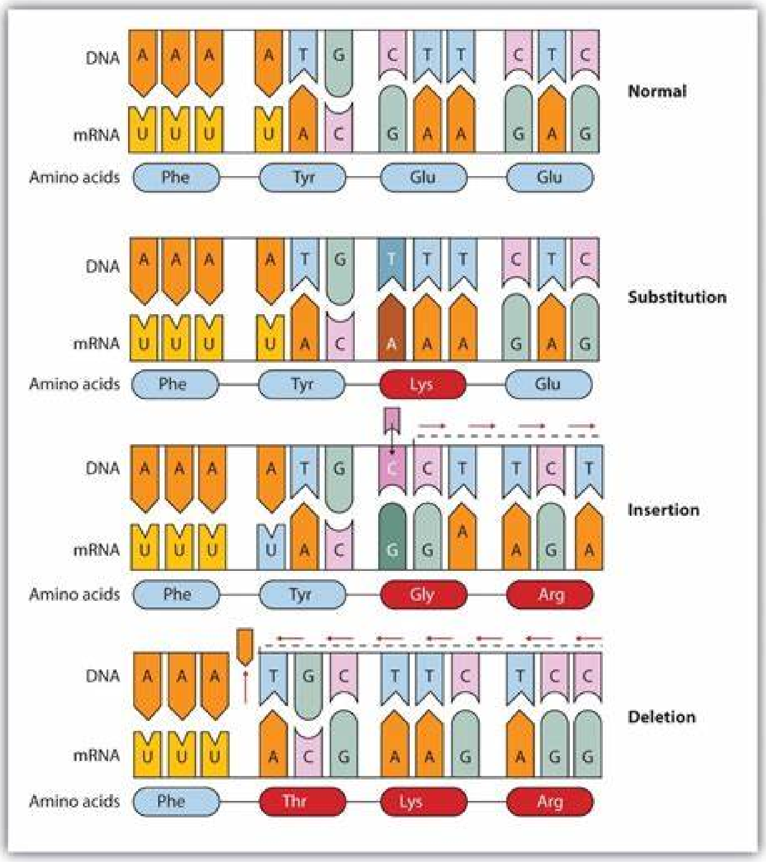

Figure 1.2: INDELs shows the 4 different scenarios which are: Normal (Shows the standard process), Substitution (Depicts a single nucleotide change), Insertion (Demonstrates the addition of an extra nucleotide) and Deletion (Shows the removal of a nucleotide)[38].

rare, potentially pathogenic variants and common polymorphisms with minimal or no effect on health.

GnomAD information has been gathered in Variant Call Format (VCF) files which normally has the following structure:

- **CHROM:** Chromosome name
- **POS:** Position on the chromosome
- **ID:** Variant identifier
- **REF:** Reference allele
- **ALT:** Alternate allele

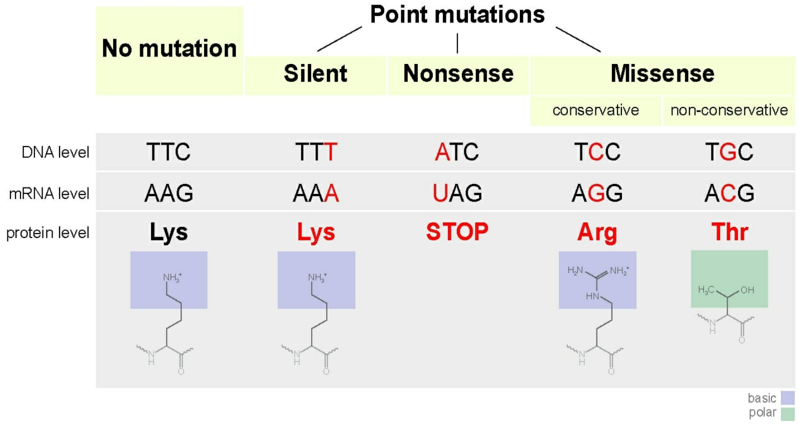

Figure 1.3: Missense vs. Synonymous vs. Frameshift illustrates the impact of different point mutations. No mutation (Shows the original DNA), Silent mutation (A change in the DNA with no change in the AA), Nonsense mutation (A change in the DNA cause premature stop codon) and Missense mutation (A change in the DNA with change in the AA which can be further classified as conservative or non-conservative).

- **QUAL:** Phred quality score for the variant call
- **FILTER:** Filter flags (**PASS** indicates the variant passed quality checks)
- **INFO:** Additional information about the variant
- **FORMAT:** Genotype format for samples

Much more information like Allele count, Allele frequency, Highest observed Allele Frequency across all populations in the dataset ect is inside INFO column [8].

It is worth mentioning that the GnomAd database has 4 different releases which have the following information:

- **ExAC:** The ExAC data set contains data from 60,706 exomes, all mapped to the GRCh37 / hg19 reference sequence.
- **gnomAD v2.1.1:** The gnomAD v2.1.1 data set contains data from 125,748 exomes and 15,708 whole genomes, all mapped to the GRCh37/hg19 reference sequence.
- **gnomAD v3.1.2:** The gnomAD v3.1.2 data set contains 76,156 whole genomes (and no exomes), all mapped to the GRCh38 reference sequence.
- **gnomAD v4.1.0 (released November 2023):** The gnomAD v4.1.0 data set contains data from 730,947 exomes and 76,215 whole genomes, all mapped to the GRCh38 reference sequence.

Before the latest release of GnomAD dataset (gnomAD v4.1.0) it was preferred to use the gnomAD v2.1.1 as it has more information especially for exonic variants. Gnomad v2.1.1 has also a Liftover version itself. All of these datasets are freely accessible in Gnomad Website although they are very heavy and it is strongly suggested to use the GnomAD public bucket on Google cloud.

1.5 Reference Genomes

Reference genomes are complete, or nearly complete, representations of an organism's DNA sequence. They serve as a foundational resource in genomics research, providing a standard for aligning, analyzing, and interpreting individual genomes [40].

The human genome has two major reference assemblies widely used: GrCh37 (also known as hg19) and GrCh38 (hg38). These assemblies are created through a multi-step process involving:

- **DNA Sequencing:**
  - Scientists first isolate DNA from a cell sample.
  - This DNA is then broken down into smaller fragments.
  - High-throughput sequencing technologies like Illumina or PacBio sequence these fragments, generating millions of short DNA reads [40].

- **Genome Assembly:**
  - Short DNA reads need assembly into a contiguous sequence representing the entire genome.
  - This complex process involves computational algorithms that overlap and align the reads to reconstruct the original DNA sequence [40].
  - Different assembly techniques exist, with advantages and limitations. Short-read assemblers are common but struggle with repetitive regions and gaps. Long-read sequencing technologies are emerging and offer more contiguous assemblies [40].

- **Gap Closure and Annotation:**
  - The assembled sequence may contain gaps where sequencing couldn't resolve the complete DNA sequence.

- Scientists use various methods like additional sequencing or computational prediction to close these gaps.
- Finally, the assembled sequence is annotated to identify genes, regulatory elements, and other functional regions.

HG37 (released in 2009) was an earlier reference genome assembly with gaps and inaccuracies but HG38 (released in 2013) is a significant improvement. It incorporates advancements in sequencing and assembly algorithms, resulting in a more complete and accurate representation of the human genome. HG38 has closed many gaps present in HG37, improved the assembly of complex regions, and provides a more reliable reference for gene annotation [41].

1.6 Genetic Intolerance

Understanding the potential consequences of genetic variations is essential for assessing their role in disease development. Various computational tools and databases have been developed to predict the impact of variations on protein function and potential pathogenicity. However, these predictions require careful interpretation as they are not always definitive [42].

As technological advancements have given rise to new methodologies for medicine prescriptions, Exome sequencing is frequently used to guide more personalized diagnosis and treatment for many genetic diseases including cancer[1, 2, 3]. While this has led to more information about pathogenic variants[4, 5, 6], many variants still remain with uncertain significance.

Many in silico predictors are used to prioritize likely candidates for each category, but it remains a major challenge to distinguish pathogenic variants from benign ones[7].

Large exome[8] and genome[9] sequencing projects have presented references of variation across the human genome providing the means to measure patterns of variability within genes [10, 11]. It has been demonstrated previously that measuring depletion of standing variation within genes can be used to identify novel disease-associated genes [10, 11]. With the current sample sizes of sequenced individuals, measurements in depletion of variation at a regional level within these genes has begun and many measures have been presented like Z-score[43], MPC[44], MTR[12] and ect.

1.7 MTR

The Missense Tolerance Ratio (MTR) is a measure of regional intolerance to missense variation, and can capture this regional level information [13]. "The MTR is a direct measure of purifying selection of missense variation within a gene calculated as a ratio between the observed proportion of missense variants compared to an expected proportion, estimated under the assumption of no selection occurring in that sequence context. A sliding window summation is used to provide accurate regional measurements." [12]

The MTR population variation was sourced from GnomAD v2 [8], the DiscovEHR dataset[45] and the UK Biobank[46] with 220,000 exome and genome sequences which were filtered for only single point variation with a quality control 'PASS' flag. Ensembl databases (v95)[47] were used for acquiring gene and protein sequences. In this study, transcripts were only used where they contained at least one single-point variant in gnomAD and had non-ambiguous sequences. Furthermore, Ensembl transcript ID's were queried for their matching HGNC gene symbols[48].

In order to calculate MTR score for each position, they have compared the observed proportion of missense variation to an expected proportion of missense variation and its calculation was performed under the absence of positive/negative selection[12].

For balancing the resolution and jitter, this study suggested using 31 codons as the sliding window and they have calculated MTR score for each residue based on the following formula:

$$MTR_i = \frac{\frac{\text{missense obsi}}{\text{missense obsi+synonymous obsi}}}{\frac{\text{missense expi}}{\text{missense expi+synonymous expi}}} \tag{1.1}$$

According to this study [13] the regions measured as intolerant to missense variation are significantly enriched for pathogenic missense variants in epilepsy genes.

According to previous studies, mechanisms associated with the pathogenesis of missense variants often correlate with the three-dimensional structure of proteins [14, 15, 16] and that, for some disease-associated genes, mutations appear to cluster within specific regions [17, 18, 19, 20, 21, 22, 23, 24, 25]. More systematic analyses have identified DNA sub-regions intolerant to missense variants [26, 14] and domains in protein families enriched in variants associated with disease [27, 28].

1.8 Intrinsically Disordered Proteins (IDPs)

For decades, the classical view of proteins positioned them as rigid structures with well-defined three-dimensional (3D) conformations essential for function. However, the discovery of functionally active proteins lacking a stable 3D structure challenged this paradigm. These proteins, known as intrinsically disordered proteins (IDPs), or intrinsically disordered regions (IDRs) within structured proteins, have emerged as a significant and functionally diverse class [49].

1.8.1 IDP Features

Unlike their structured counterparts, IDPs/IDRs lack a fixed 3D conformation, existing as an ensemble of rapidly interconverting conformations [50]. This inherent flexibility is primarily attributed to the amino acid composition of these regions, often enriched in disorder-promoting residues like glycine, proline, and they tend to be deficient in hydrophobic amino acids and enriched in polar and charged residues [49].

1.8.2 IDP Functions

The very feature that challenges the classical view of proteins – their lack of a fixed structure – plays a crucial role in IDP/IDR function. This inherent flexibility allows IDPs/IDRs to interact with multiple partners with high specificity and low affinity, a key aspect in cellular signaling and regulation [51]. Additionally, IDPs/IDRs are prime targets for post-translational modifications (PTMs) like phosphorylation and ubiquitination, further modulating their function and interactions [52]. Specific linear sequence motifs within IDRs, termed Short Linear Motifs (SLiMs), act as recognition elements for binding partners, highlighting the intricate code embedded within these seemingly disordered regions [53].

1.8.3 IDP Databases

Dissecting the universe of IDPs/IDRs necessitates robust informatics tools. Databases like DisProt provide comprehensive information on experimentally validated IDPs and predicted IDRs within protein sequences [30]. These resources empower researchers to identify and characterize IDPs/IDRs, paving the way for further functional studies.

"DisProt is the major manually curated dataset of Intrinsically Disordered Proteins, both for structural and functional aspects."[30] DisProt is based on three different ontologies to annotate intrinsically disordered regions:

- **The Intrinsically Disordered Proteins Ontology (IDPO)** which is used to describe structural aspects of an IDP/IDR, self-functions and functions directly associated with their disordered state and it is maintained by the DisProt consortium.
- **The Gene Ontology (GO)** which is used to describe functional aspects of an IDP/IDR
- **The Evidence and Conclusion Ontology (ECO)** which is used describes the technique or evidence associated with an annotation [30].

"MobiDB is a valuable resource for researchers and scientists studying intrinsically disordered proteins (IDPs). This database provides comprehensive information, analysis, and tools related to IDPs, aiding in the understanding of their structure, function, and dynamics." This database contains many information like:

- **IDP Entries and Annotations** MobiDB hosts a vast collection of entries on intrinsically disordered proteins. Each entry includes detailed annotations and information about the protein, such as its sequence, disorder predictions, functional regions, post-translational modifications, binding partners, and biological functions. These annotations are integrated from various reliable sources and can serve as a valuable reference for researchers.
- **Disorder Prediction Algorithms** The database incorporates state-of-the-art disorder prediction algorithms that assess the likelihood of disorder in protein sequences. These algorithms utilize different computational techniques and machine learning approaches to predict regions of intrinsic disorder within a protein sequence. MobiDB provides access to these prediction tools, allowing users to analyze their own protein sequences and obtain disorder predictions.
- **Structural Information and Visualization** MobiDB offers structural insights into intrinsically disordered proteins. It provides information on experimentally determined structures, such as NMR ensembles and X-ray crystallography data, as well as predicted structures based on computational modeling. Users can visualize and explore the structural information through interactive tools and viewers, enabling a deeper understanding of the conformational behavior and dynamics of IDPs.
- **Functional Annotations and Pathways** Understanding the functional aspects of intrinsically disordered proteins is crucial for unraveling their biological roles. MobiDB integrates functional annotations, including protein-protein interaction data, post-translational modifications, and binding sites.

- **Comparative Analysis and Cross-References** MobiDB facilitates comparative analysis by enabling users to compare multiple aspects within and between organisms. This feature allows for the identification of conserved regions, common binding partners, and shared functional annotations among different IDPs. The database also provides cross-references to other relevant protein databases and resources, enabling seamless integration with existing knowledge and facilitating interdisciplinary research [54].

1.9 Objectives

Here in this research it is intended to find out the MIRs inside all the introduced proteins for humans in Disprot Dataset[30] which are known to be IDPs and see if there is a correlation between MIRs with IDRs since there are no studies that clearly accept or reject the existence of any correlation between the two regions as far as this team knows.

2
Methods

In this Chapter it is intended to explain the steps that have been carried out to achieve the goals mentioned in the previous chapter. This chapter is divided in two sections: Section 2.1 and Section 2.2.

2.1 Finding MIRs

At first step, knowing the MIRs is required so to find these regions, finding a propitiate method is essential. Considering the previous studies, Selecting a method which serves the requirements of this study was pretty hard as there are many different methods presented for this purpose like Z-score[43], MPC[44], MTR[12] and ect.

Finally it is decided to use MTR [12] as a means to find MIRs considering that population variation was richer in comparison to other studies with 220 000 exome and genome sequences.

All of this information is stored in a flat file and freely accessible at the MTR-Viewer web server [55] for further analysis which this project aims to but just like any other prepared dataset it needs some adjustment to be used for our aims to.

2.1.1 Lifting over the MTR positions

The MTR flat file though reach and accessible, it uses GrCh37/hg19 build genomic coordinates[55] which is draw back for our purpose as the adjustment happened on 2013 every new information published is based on the latest genomic coordinate GrCh38/hg38 build so, to overcome this problem, Picard tool is used to lift the data from previous to the latest build version (hg38).

Picard tool is a java code generated by Genome Analysis ToolKit (GATK) to lift a VCF file from one build version to another having the new reference sequence."This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA." It is worth mentioning that using this tool may cause losing a part of the original dataset as it can not lift all the entries due the filter status[56] of them but in our case, in the lifting process less than 1 percent of the entries were lost.

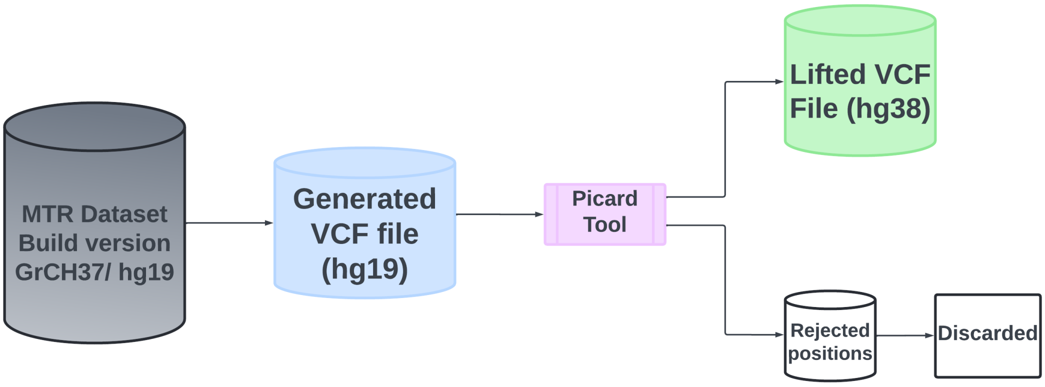
MTR Dataset Build version GrCH37/ hg19
Generated VCF file (hg19)
Picard Tool
Lifted VCF File (hg38)
Rejected positions
Discarded

Figure 2.1: The lifting process consists of steps to create the VCF file from the dataset and then pass this file to the picard tool for the lifting process. At the end the lifted position will be emerged in a new VCF file and the rejected positions will be emerged in a different VCF file which is discarded.

2.1.2 Finding the IDPs

Now that we have the MTR dataset which contains at least 85000 transcripts we have to narrow down our research to the aimed protein population which is IDPs. The list of IDPs that has been used for this research is driven from Disprot[30].

As MTR dataset use the Ensembl Transcript id for each protein according to Ensemble v95 [12] and Disprot uses only Uniprot accession for each protein we need to do a cross reference to finds the matches between two dataset using the gene name for which we have used the Uniprot Id mapping tool[57].

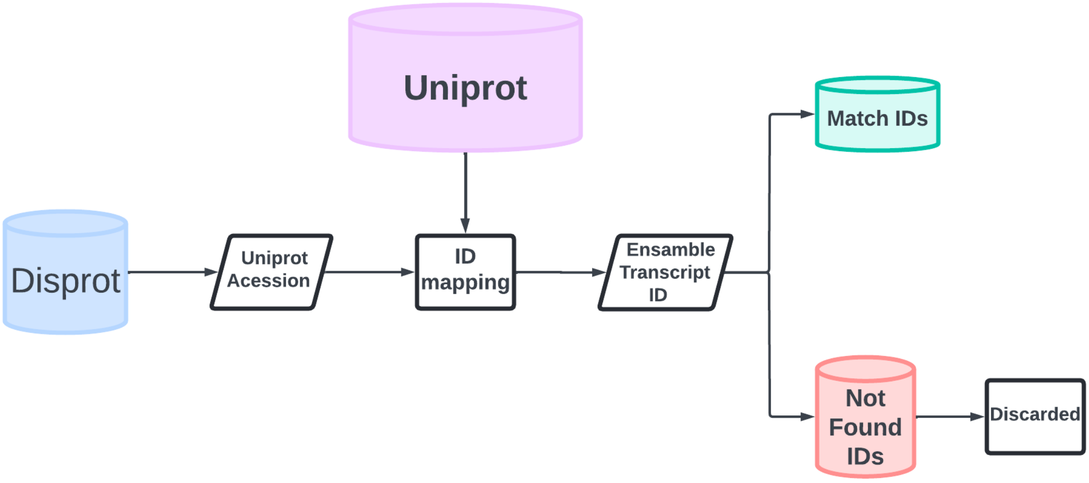

Figure 2.2: Uniprot accessions were used to find the gene name and the Ensemble transcript sequentially and any not-found ID was discarded.

2.1.3 Updating the MTR dataset

After Lifting the positions to the latest genome build version (GrCh38), it is necessary to update these coordinates in the MTR dataset for the target proteins(IDPs). In-order to do that we have extracted the nucleotide sequence from both build versions (hg19 and hg38), translated them and performed an alignment using NCBI pBlast [58] to validate the lifting process2.3. In this process, all the identical entries were kept and the rest were discarded.

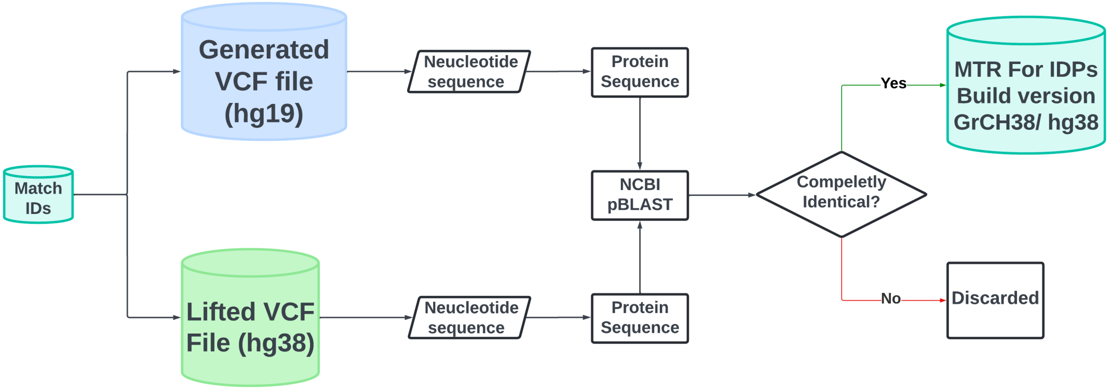

Figure 2.3: The positions in the MTR dataset were updated only for the matched IDs that preserve their sequence after lifting, using the 2 VCf files and the rest discarded.

2.1.4 MTR to MIR

After updating the position in the MTR dataset it is necessary to extract the intolerant regions inside each transcript. In-order to obtain this region, considering that each transcript represents a protein and each protein contains at least 100 amino acids and each amino acid consists of 3 nucleotides, we have at least 300 entries for each protein.

The Central Limit Theorem (CLT) tells us that as the sample size increases, the distribution of the average of independent and identically distributed random variables approaches a normal distribution regardless of the original distribution's shape.

For simplicity, if we consider the MTR score as a random variable for each residue from the same distribution and also considering the scores independent from each other, then according to the CLT we can use the represented MTR score for each residue to calculate the average and the standard deviation, we can make the following assumption:

$$\begin{aligned} H_0 : &\quad x > \mu - 2\sigma \\ H_1 : &\quad x <= \mu - 2\sigma \end{aligned}$$

Where mue is the average MTR score for the whole transcript and sigma is the standard deviation. In this case with these assumptions that we have made, we announce that if the MTR score is equal or less than the mean minus 2 standard deviation, then we have considered it as a hotspot and if it gets more than 3 residue we have considered them as an intolerant region.

2.2 MIRs vs IDRs

Now that we define and find out the MIRs it is time to find the important regions of IDPs in order to do that the following steps need to be followed sequentially 2.4.

2.2.1 IDPs Identified regions

Using the MobiDB database we can extract:

- **'Homology-domain-pfam':** Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models.Proteins are generally composed of one or more functional regions, commonly termed domains. Different combinations of domains give rise to the diverse range of proteins found in nature.

The identification of domains that occur within proteins can therefore provide insights into their function. These information is directly used in MobiDB.

- **'Prediction-disorder-mobidb-lite':** Disorder predictions are provided via the MobiDB-lite software. MobiDB-lite is a consensus method which is optimized to find long IDRs and to be extremely precise. MobiDB-lite reports also sub-regions which are particularly biased in terms of amino acid composition, some of the subregions follow Pappu's classification.

- **'Curated-disorder-priority':** MobiDB includes annotations from third-party manually curated databases which report disorder evidence from the literature. Integrated databases are: UniProtKB/SwissProtKB, DisProt and IDEAL.

- **'Derived-binding-mode-disorder-to-disorder-priority':** MobiDB calculates binding modes from PDB structures by analyzing the disorder content in monomeric form and in complex. Three different binding modes are derived: disorder-to-disorder, disorder-to-order, context-dependent.

- **'Prediction-lip-priority':** Those regions are called with different names, e.g. MoRFs, SLIMs, etc. In MobiDB a more general term is used which is Linear Interacting Peptides (LIPs) which embrace different subtypes. Interacting surfaces of IDPs exhibit a unique set of chemo-physical properties, e.g., a higher percentage of hydrophobic residues compared to the rest of the IDR, and a larger exposed interaction area per residue - even in their folded state induced by binding - that they use to contact their physiological partners. This information was predicted using the MobiDB software [59].

Although there are many more regions with different assigned names in this database, this team intends to use only these 5 regions which seems to be most relevant and accurate for the project purpose.

2.2.2 MIRs and Identified regions

After extracting the identified regions that was selected from MobiDB database it is time concatenate these information with the MIR information to create a unified data frame which carries all the necessary information to achieve the ultimate goal of this project which is visualizing and statistically trying to find the existence of any correlation between MIRs with any of these 5 named regions.

In the next chapter 3 all the achieved results will be discussed deeply.

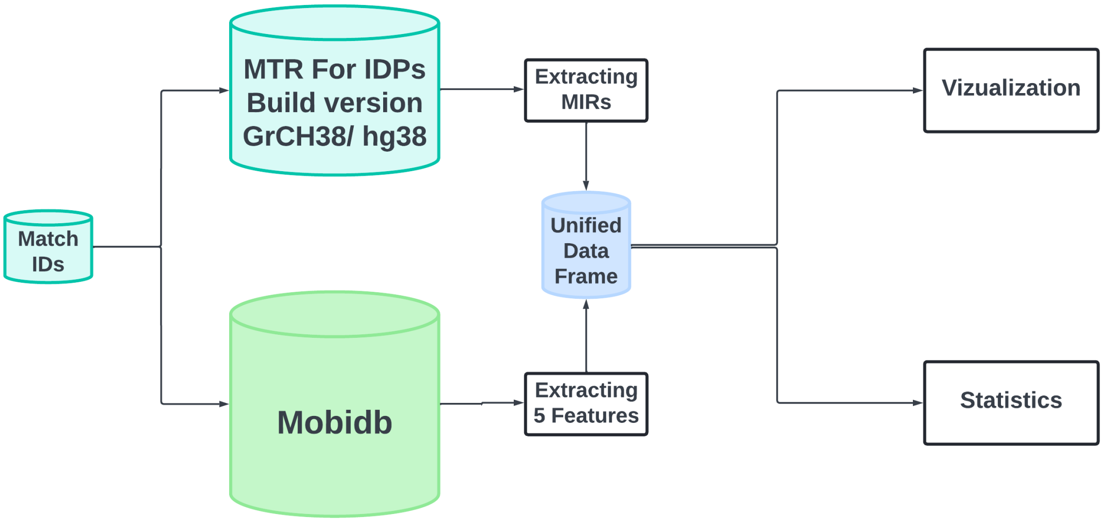

Figure 2.4: Analysis and visualization was performed based on the extracted information from The MTr dataset and the MobiDB selected features.

3 Results

In this chapter it is intended to analyze the achieved information from following the steps that has been discussed in the previous chapter. This chapter is divided in Five sections: Section 3.1, Section 3.2, Section 3.3, Section 3.4 and Section 3.5 .

3.1 MIRs Finding Results

As discussed before The lifting process was crucial to be able to carry on with the study purpose but before that it is intended to indicate how suitable the MTR dataset is for the target proteins. As indicated in the following figure 3.1, before lifting the dataset to the latest build version (hg38), 180 IDPs do not have information about their MTR score in the MTR database because the MTR datasets was build only for 85000 Ensembl Transcript ID which did not cover these 180 IDP IDs so, they had to be discarded from the study.

After performing the first two substeps of the Finding MIRs step (see figure 2.2 and 2.3), due the fact that 1 percent of the whole Positions from MTR datasets was lost and get rejected because the picard tool could not find the corresponding position in the HG38 build version, 30 more IDPs were also discarded. These limitations have decreased our IDPs from 1071 IDP to 861 IDP. At the end of this step, a new dataset that contains the MIR score for these 861 IDPs was built.

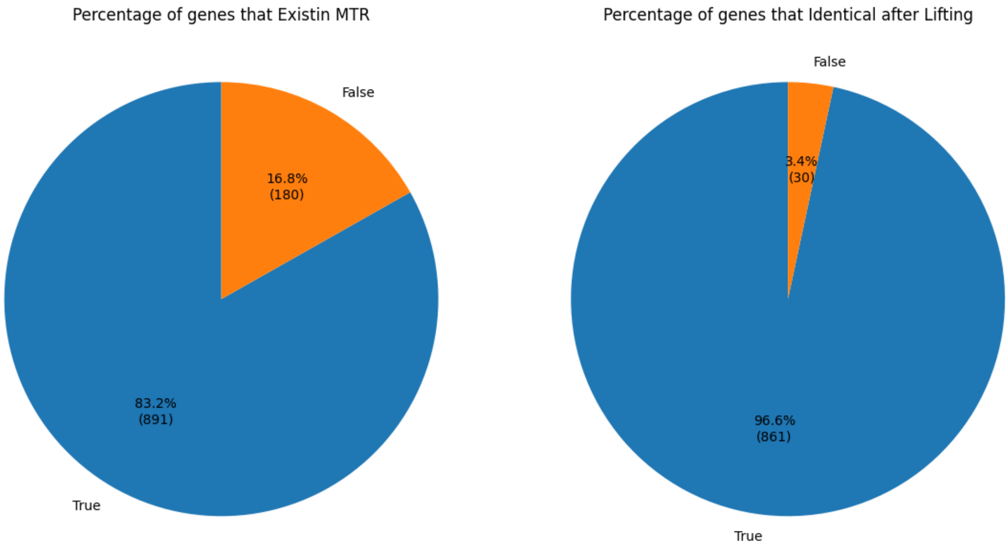

Figure 3.1: Percentage of proteins with curated disordered regions that have (in blue) or not have (in Orange) MTR score information on the left.Percentage of proteins which were identical (in blue) or not (in orange) on the right.

3.2 Dataset general information

According to the process that can be found in figure 2.4, the MIRs and the 5 selected features were extracted from the said datasets and the unified dataset was built. This dataset contains all the necessary data that this study tends to analyze.

According to the figure 3.2 from 861 IDPs that the scope of this study could cover due its limitations, 854 IDPs had at least one hotspot that was identified by this study and there were 7 IDP which they did not have even a single intolerant position according to this study which will be discussed deeply in the section 3.5. 762 IDPs had at least 1 MIR which were located in their Pfam Domains, this number for 'prediction-disorder-mobidb-lite', 'curated-disorder-priority', 'derived-binding-mode-disorder-to-disorder-priority' and 'prediction-lip-priority' are 244, 327, 154 and 475 respectively. These numbers suggest that many MIRs tend to happen more within the domain of the proteins and then in the LIPs rather than IDRs.

Observing the distribution of the MIRs from the position perspective it is indicated that the first 100 AA of the Disprot proteins tends to have MIR more than the other location (see figure 3.3 on the left). This could be due the fact that many of the proteins might have length around 100 AA but as the the figure 3.3 on the right indicates many of the genes has the length around 500 AA or more so it can be suggested the first 100AA of the Disprot IDPs tend to have more

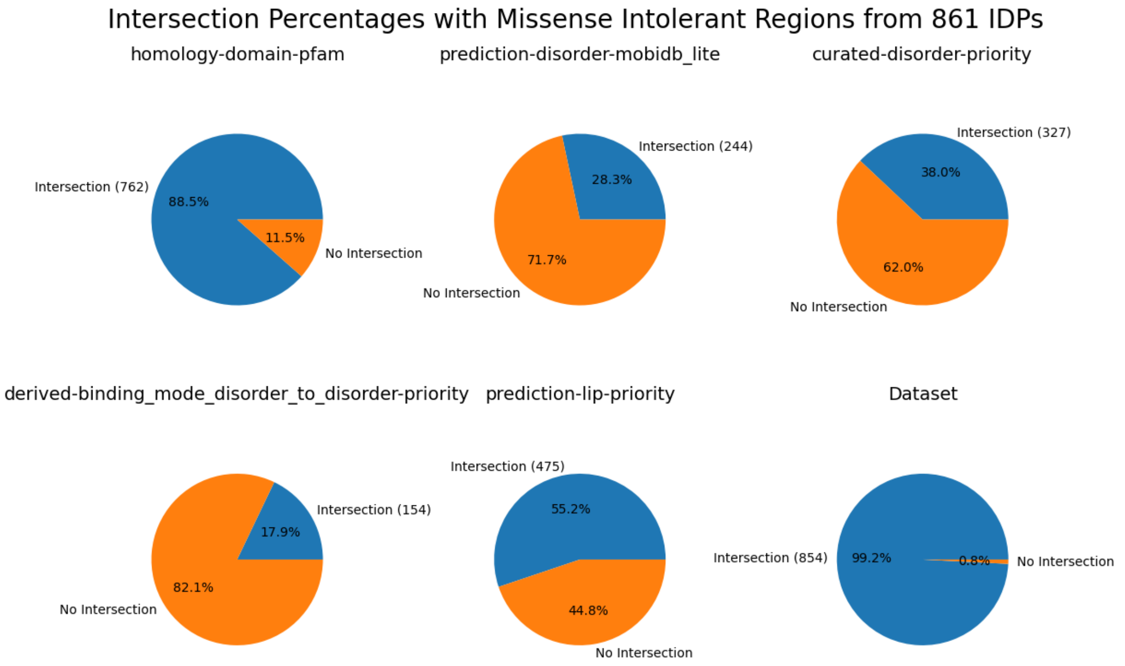

Figure 3.2: The number of genes that had (in blue) or had not (in orange) at least one MIR in them intersecting with other features. The last pie chart on the down right suggests that there were only 854 genes with at least 1 identified MIR.

importance in comparison to other positions.

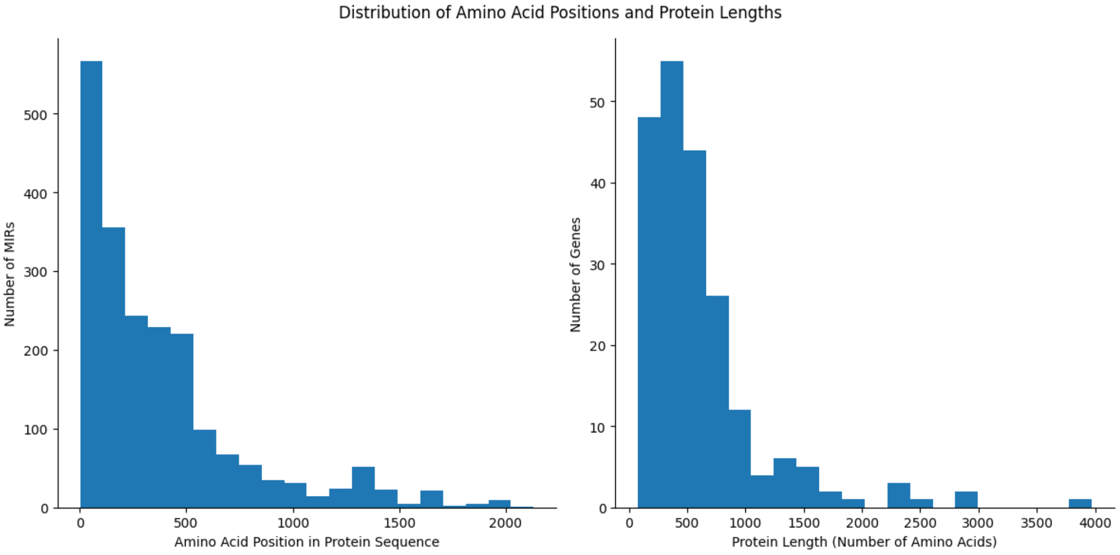

Figure 3.3: The MIR distribution inside the IDPs and the IDP length distribution.

3.3 Correlation of MIR with disorder related features

Considering the above information, assuming that the hotspots(single intolerant point to mutation) or MIRs have strong correlation with the Domains and LIPs might be true but considering the figure 3.4 it is obvious that there is no correlation between the MIRs with any selected features due the fact that MIRs are less populated than any other regions so their overlap can not be dependent to one another.

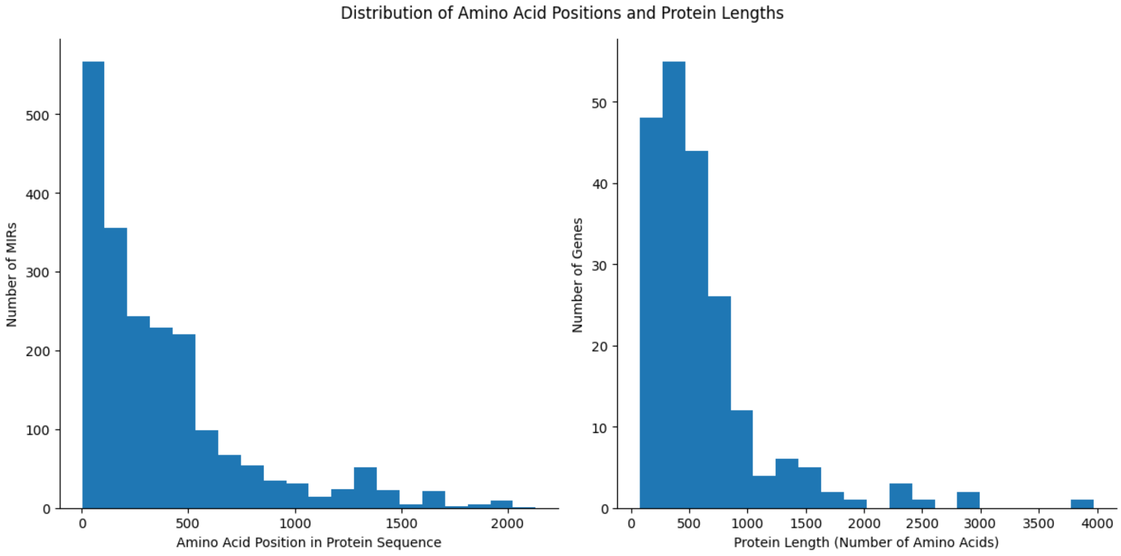

Figure 3.4: Regions Correlation Heatmap indicates the correlation between each pair of features that this study considered.

Although there are some correlation between some features like 'curated-disorder-priority' and 'derived-binding-mode-disorder-to-disorder-priority' but this matrix can not be very informative when it comes to the objective of this project so more analysis can enlighten the situation.

Observing the coincident of MIRs within each of the 5 feature, the bar chart (see figure 3.5) illustrate that almost 60% of hot-spots or MIRs tend to fall into the domains, almost 8 percent in the 'prediction-disorder-mobidb-lite' and almost 11 percent in the 'curated-disorder-priority'.

Furthermore almost 5 percent of the MIRs tend to happen in 'derived-binding-mode-disorder-to-disorder-priority' and more than 15 percent in 'prediction-lip-priority'. Looking at the numbers, something around 2 percent is missing which suggest the idea that MIRs can happen even if that location is not considered to be an important region in the IDP, this percentage might be higher than 2 percent since some MIRs might be located in a region that is labeled with more than 1 of the selected feature.

For example a hot spot might be located in a portion which can be inside the domain and disorder part of the protein so it is necessary to consider all the possible combinations of this event.

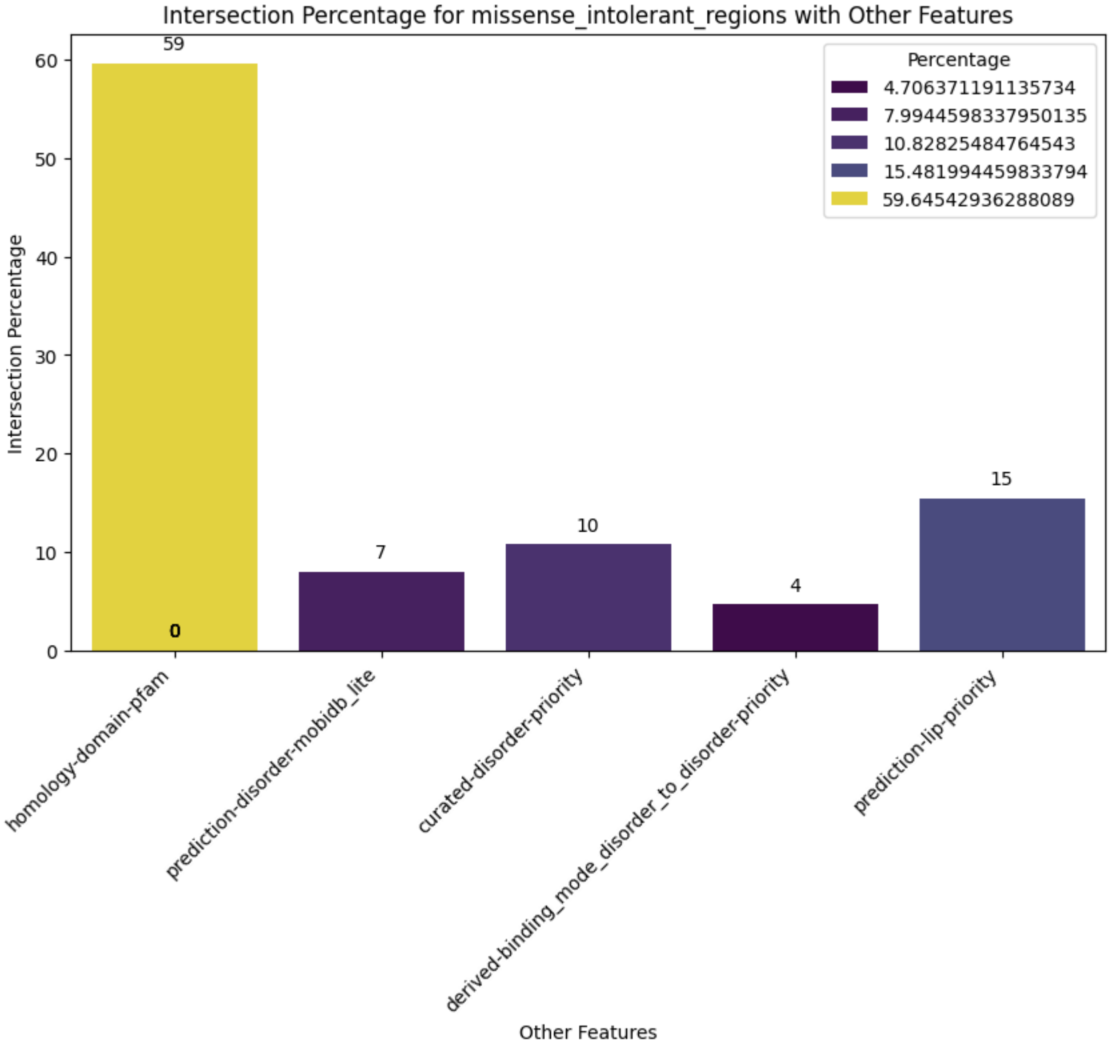

Figure 3.5: MIRs in MobiDB features indicate the probability of MIRs happening in any identified region.

3.4 MIR in IDRs

By far whatever has been achieved still supports the fact that the MIRs in the domains tend to happen more often as the domains known to be the functional part of the proteins but this study aims to find the importance of IDRs when it comes to MIRs and therefore the bar plot in figure 3.6 show that from that 8 and 11 percent which were the possibility of having MIRs in the 'prediction-disorder-mobidb-lite' and 'curated-disorder-priority' respectively, it is necessary to subtract 2 and 5 percent respectively which is the possibility of having MIRs in the location that is both the domain and also IDR.

Finally these results suggest that only 6 percent of the MIRs happen in the non-domain regions that are known to be either 'prediction-disorder-mobidb-lite' or 'curated-disorder-priority' regions. These results indicate that the possibility of having MIRs in the domain is almost 10 times more than the IDRs.

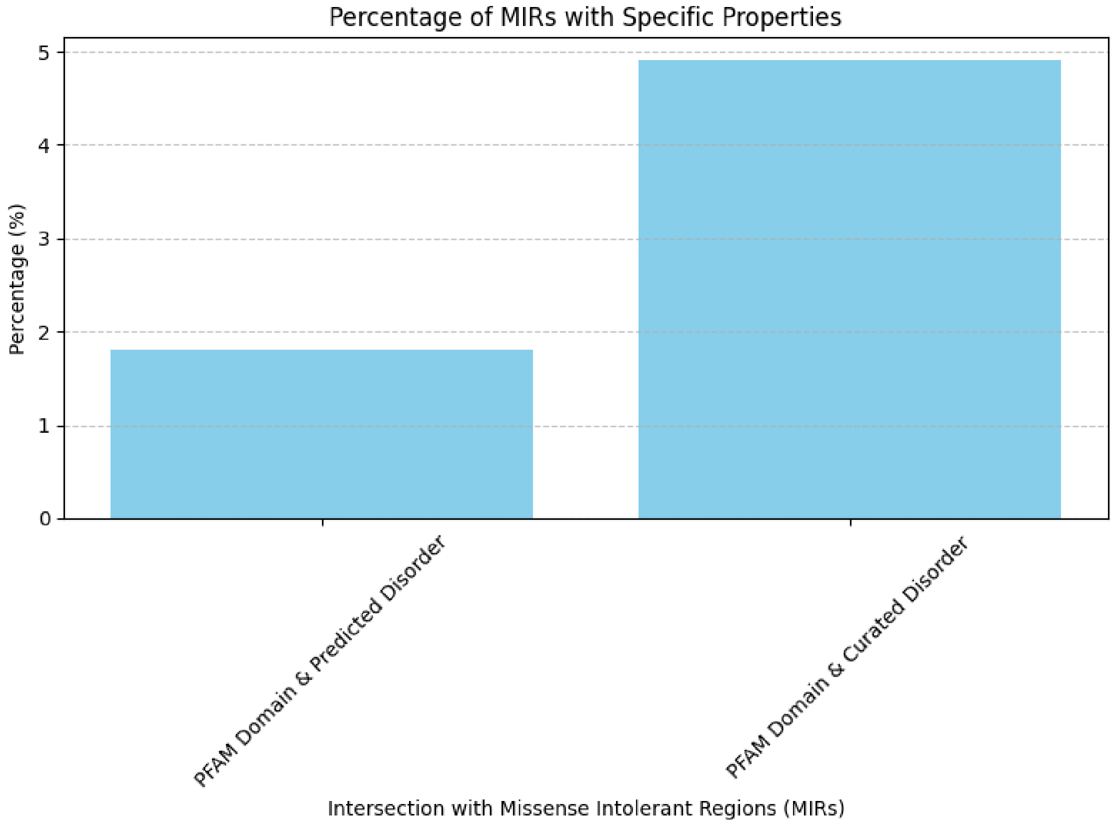

Figure 3.6: The probability of having MIR, IDR and Domain at the same time is indicated above.

3.5 Interesting Genes

In this section it is intended to discuss some interesting examples which this study achieved. According to what has been discussed the section 3.1 there were 7 genes that this study algorithm could not find even a single MIR for them and digging deep to these genes it was understood that the algorithm tends to fail finding the MIRs when the the selected threshold falls below the minimum MTR score (MTR score range[0, 1.418]) or when the MTR score has high value for all of the AA position. Below 3 selected genes out of that 7 genes are discussed.

- **U2AF2** This gene is considered as one of the genes that the algorithm with the threshold of 1.5 STD could not find even a single position as a missense intolerant position and the reason behind that is that the threshold falls below zero (see figure 3.7).

  Checking this gene variant from Clinvar dataset it was observed that at the position 149 and 249, 2 pathogenic variants have been identified which is indicated with red dots in figure 3.7. As this figure indicates the MTR score has high deviation and therefore the algorithm could not find any MIRs furthermore the identified variants have higher tolerance compared to other positions and that can explain the lack of performance for this specific gene.

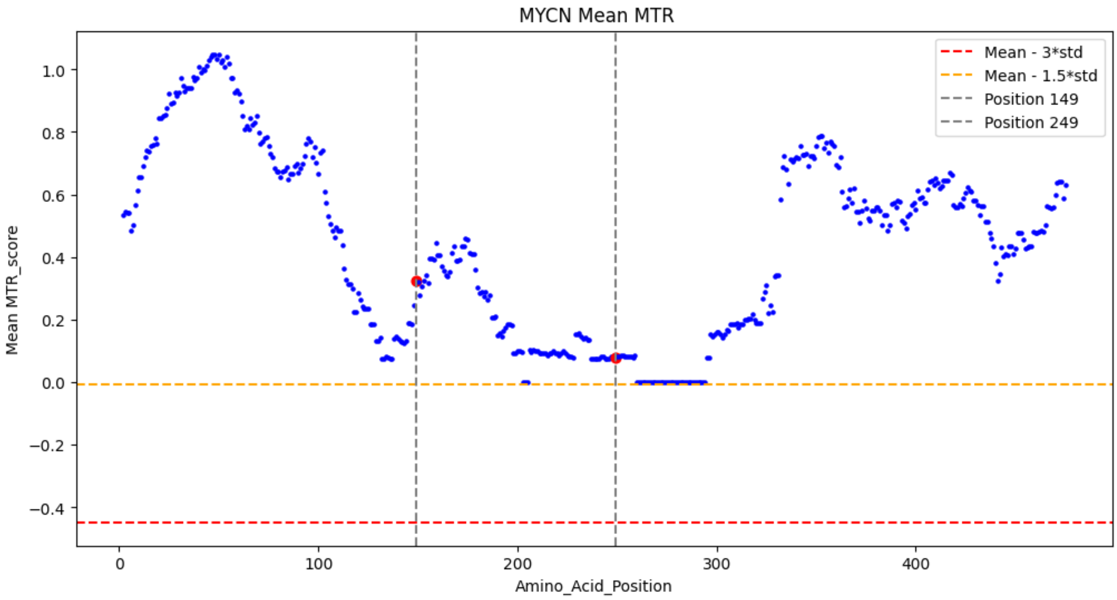

Figure 3.7: The scatter chart illustrates the deviation of MTR score through the protein length.

According to the figure 3.8 it is indicated that the variant at position 149 has not any overlap with any of the identified regions but the variant at position 249 is just located in the curated disorder part. This result can emphasize the importance of the IDRs due the fact that one of the known variants is located in a disordered region. It is worth mentioning that there is a region that can be considered as MIr due the fact that their MTR score is zero and this region is between 260 and 300 which overlap with the second domain of this protein.

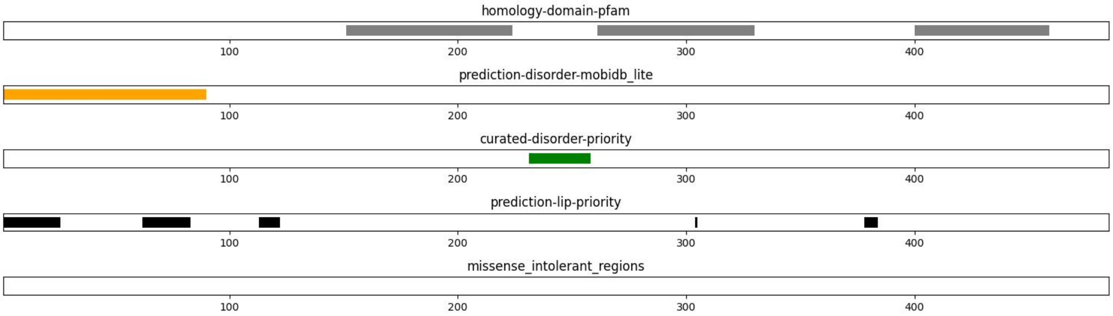

Figure 3.8: The plot illustrates the MobiDb features and the MIRs for the U2AF2.

- **EIF1AX** This gene like the one before was ignored by the algorithm due the high division and due the fact that the threshold falls way below the zero but again at position 99 there is a known pathogenic variant by Clinvar that brings the attention to that region of the gene.

  According to what can be seen in the figure 3.9 there are 3 regions, one between 20 and 30, the other between 70 and 80 and the last one between 80 and 110 which has a really low MTR score. These can be considered as MIRs and as it is indicated the known mutation is located in the third potential region.

  According to the figure 3.10 these potential regions have high overlap with the domain of this protein which lowers the importance of IDRs but the identified variant falls out of all the identified regions.

- **H3-3A** The last gene that it is intended to discuss is H3-3A which does not have any clinically proven pathogenic variant according to Clinvar, but the interesting par that according to figure 3.11 and 3.12 the region between 0 and 40 have MTR score equal to zero which pron them to be a MIR and the interesting part is that this region has overlap with the protein domain, curated and predicted disorder section.

  The fact that the protein seems to be all domain reduces the importance of the domain and as you can see the potentially identified MIR has almost 100 percent overlap with the disorder part that indicates the importance of the IDR in this protein.

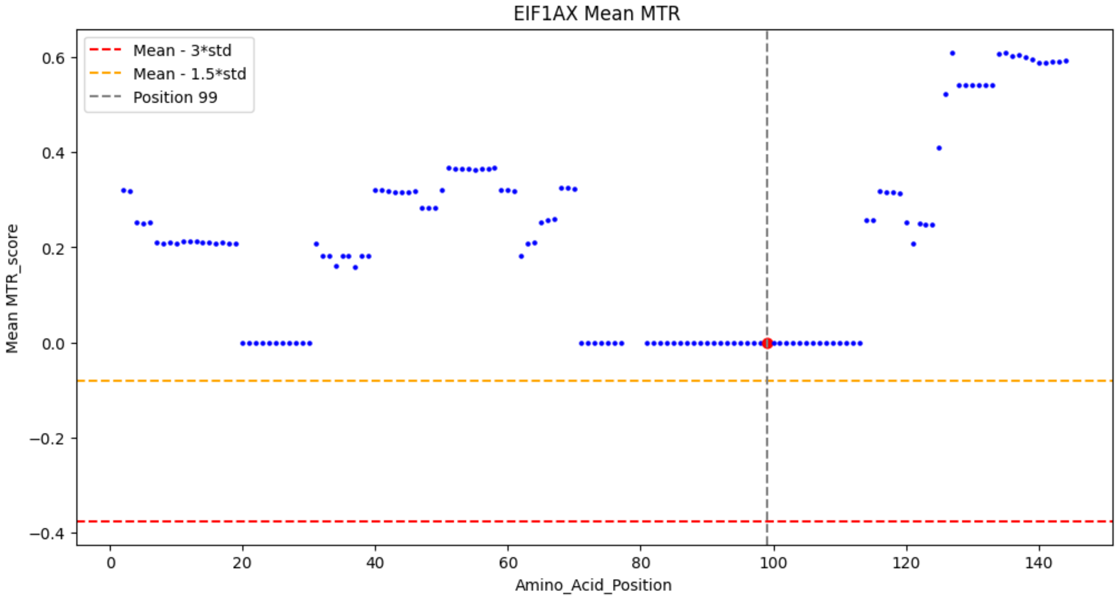

Figure 3.9: The scatter chart illustrates the deviation of MTR score through the protein length.

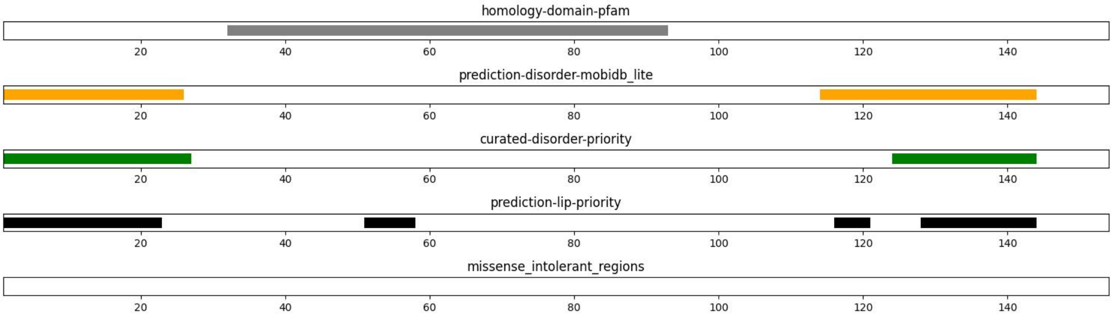

Figure 3.10: The plot illustrates the MobiDb features and the MIRs for the EIF1AX.

Furthermore There are 3 more gene that this study tends to discuss which know to have Variant in the IDRs:

- **MECP2** This gene is known to be an IDP in which pathogenic mutation causes Rett syndrome and intellectual disability. According to the figure indicates the 3.13 the division of MTR score is high in this gene and overally, most of the positions have high MTR score which means that most of its positions must be tolerant to the Missense mutations but the red dots in this picture indicates the known pathogenic missense variants.

  These results suggest that despite having a high MTR score still chances of having pathogenic missense mutation exist.

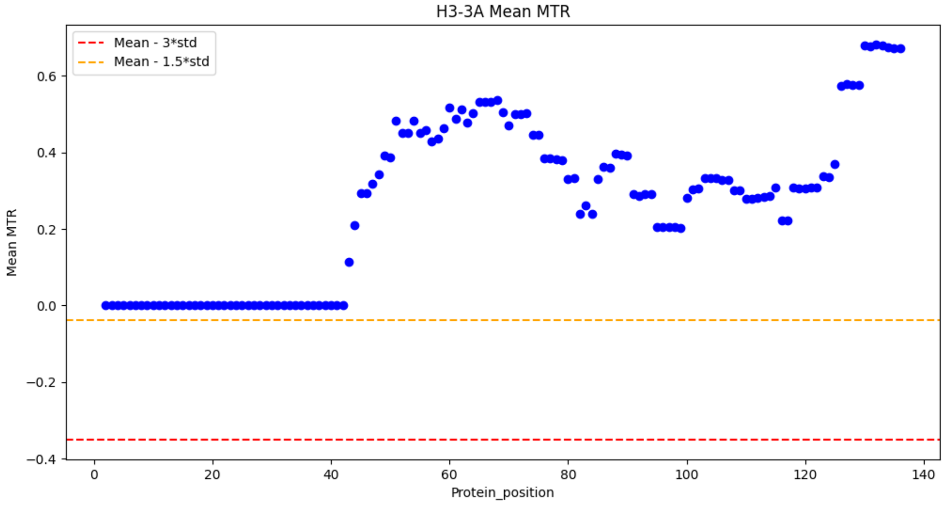

Figure 3.11: The scatter chart illustrates the deviation of MTR score through the protein length.

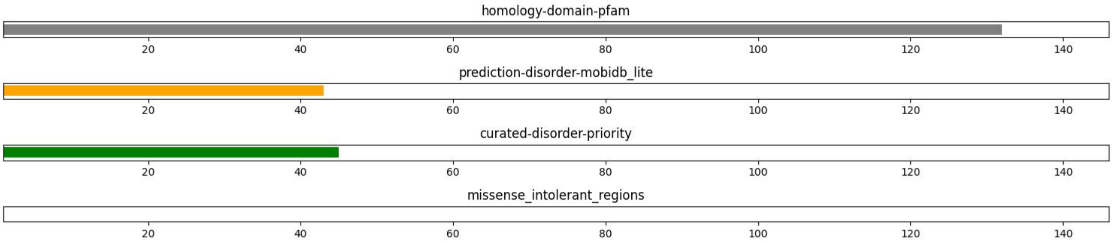

Figure 3.12: The plot illustrates the MobiDb features and the MIRs for the H3-3A.

As it is indicated with big red dots in figure 3.13 even at very high MTR score 1.08 there is an Identified Pathogenic Missense Variant (IPMV). The interesting part of this plot is the positions 321 to 324 which the algorithm identifies as MIR which according to the figures 3.13 and 3.14 happen to be inside the second curated IDR of this protein which has no intersection with the domain of this protein.

- **MYCN** This gene also plays a role in many diseases like Malignant neoplasm of the body of the uterus, Medulloblastoma, Glioblastoma, Pancreatic adenocarcinoma and Neuroblastoma which make this gene an important case for study.

  According to the figures 3.15 and 3.16 almost all of the IPMV of this gene happen within the MIRs that were identified by this study which provides good validation for the robustness of the used algorithm.

  According to the figures 3.15 and 3.16 the interesting identified MIR that had 1 IPMV at

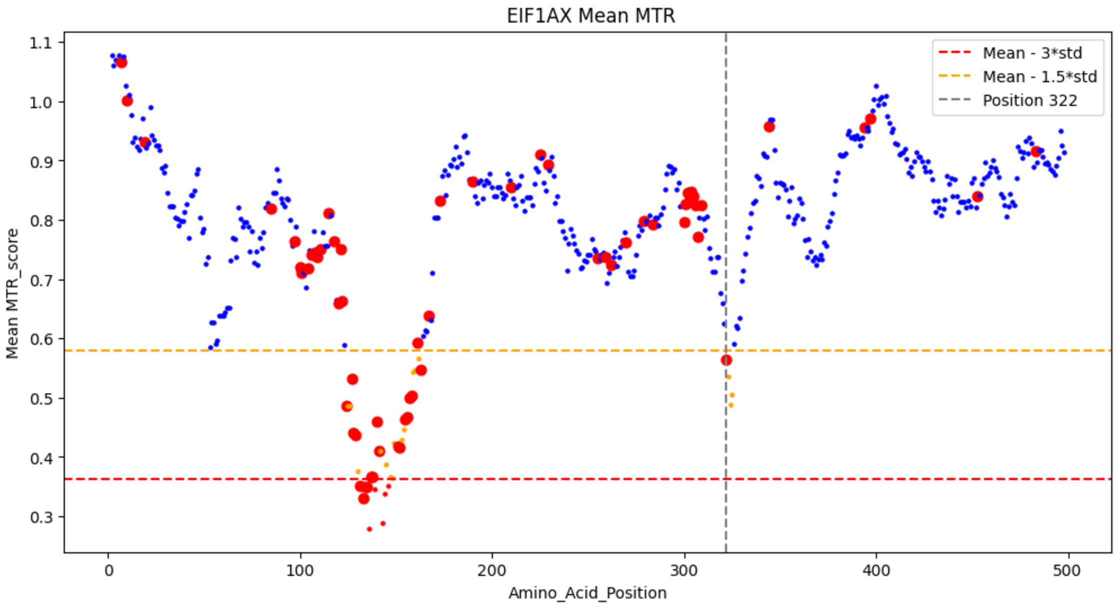

Figure 3.13: The scatter chart illustrates the deviation of MTR score through the protein length.

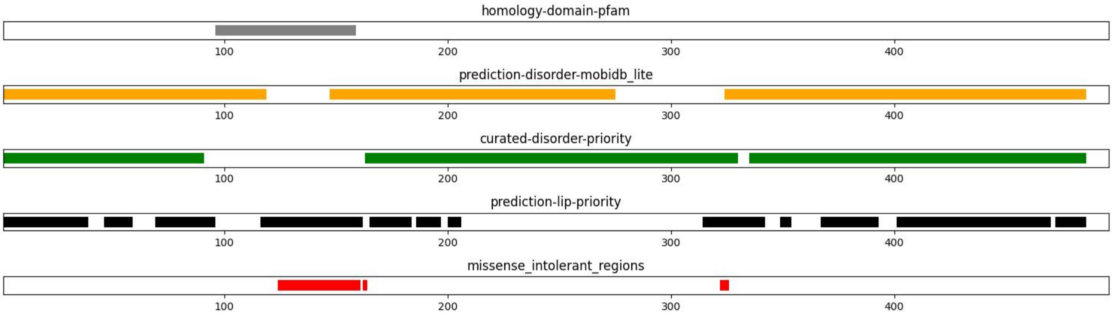

Figure 3.14: The plot illustrates the MobiDb features and the MIRs for the MECP2.

position 44 and has overlap with the domain of the protein, curated IDR, drive binding mode disorder to disorder and lip which specify the importance of the identified MIR within this gene.

- **PAK1** This gene happens to play a part in Intellectual developmental disorder with macrocephaly, seizures, and speech delay.

  According to the figure 3.17 and 3.18 there are many MIRs that are identified by the algorithm and some of them contain IPMV like the positions 110, 470, 474 and 476. These positions validate more the quality of identified MIRs but it is worth mentioning that there are also some identified IPMV that do not fall into the MIRs due the fact that

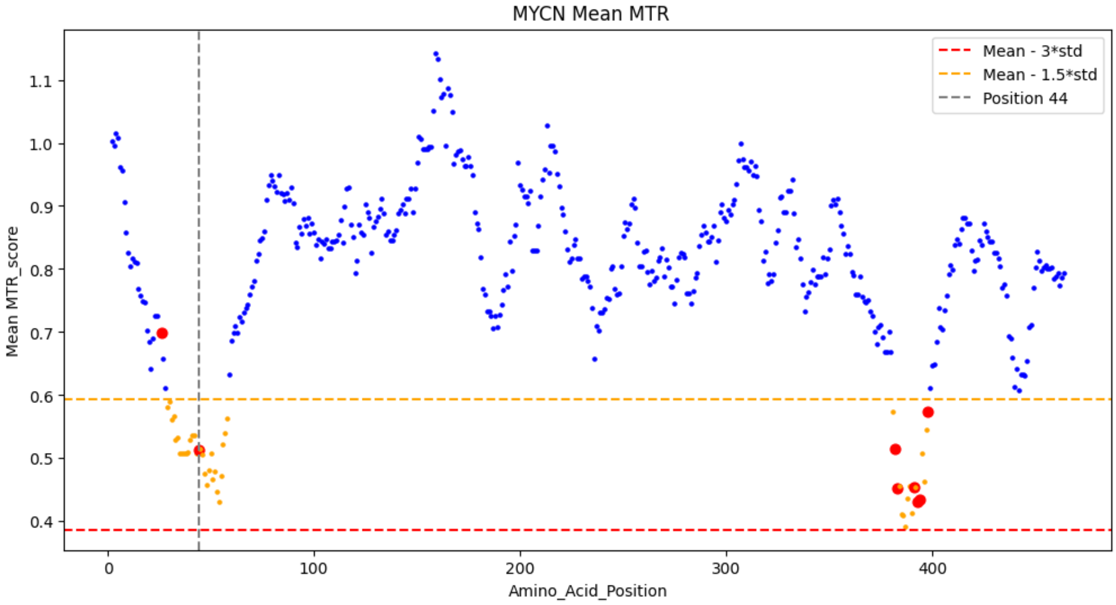

Figure 3.15: The scatter chart illustrates the deviation of MTR score through the protein length.

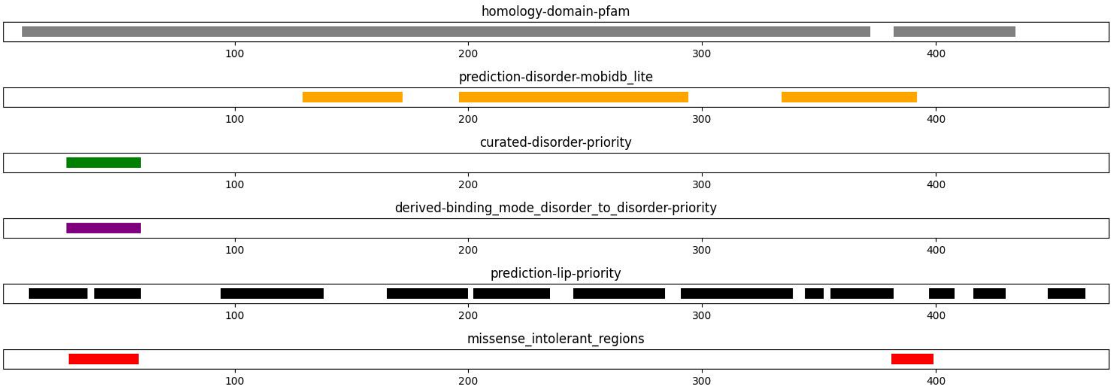

Figure 3.16: The plot illustrates the MobiDb features and the MIRs for the MYCN.

these positions have high MTR score.

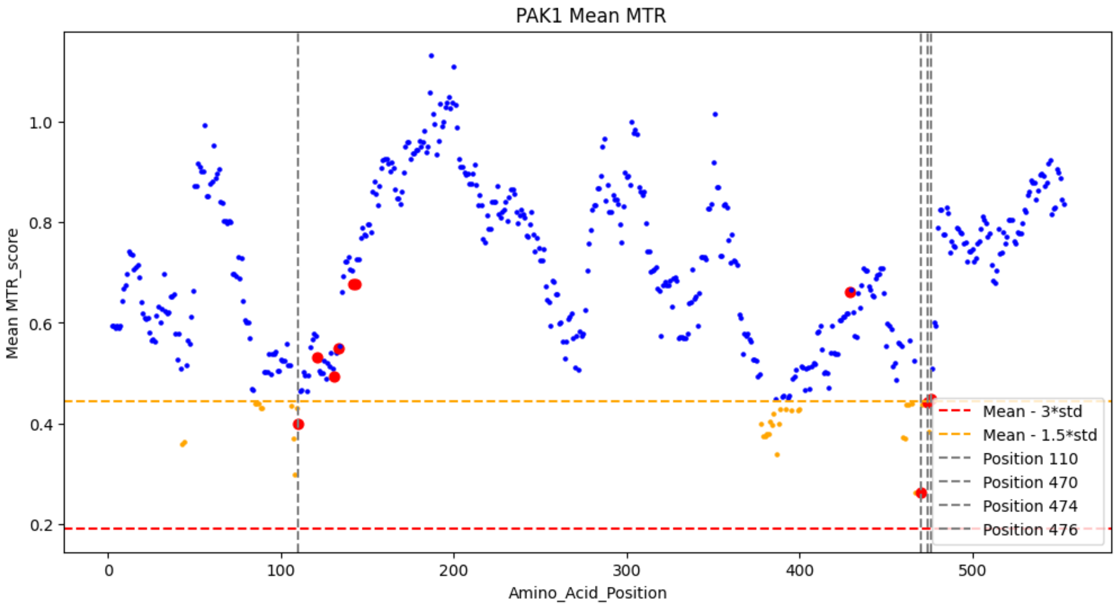

Figure 3.17: The scatter chart illustrates the deviation of MTR score through the protein length.

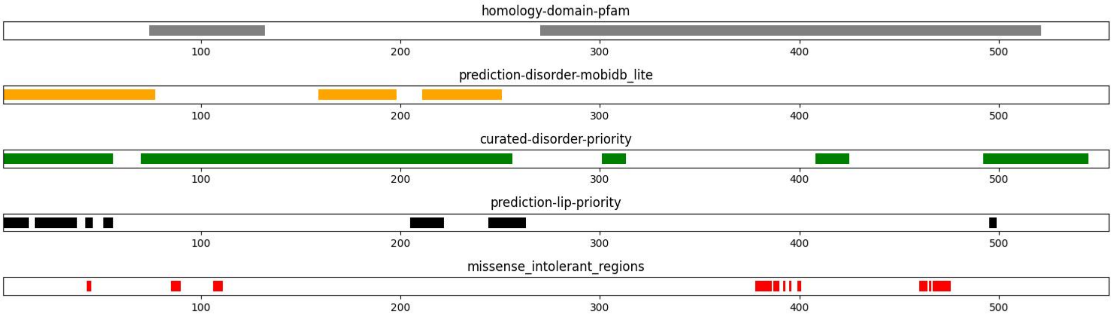

Figure 3.18: The plot illustrates the MobiDb features and the MIRs for the PAK1.

4
Discussion and Conclusion

All things considered the MTR dataset besides the algorithm that this study developed provides a good MIR dataset for almost 80% (854 IDP) of the Disprot Human IDPs which were the Target Proteins that this study tends to observe. As illustrated in the previous chapter, there is no evidence of correlation between MIRs with IDRs or any other selected features from Mobidb Dataset. However, the probability of having MIRs within the Domains of the protein is much higher than the same probability for other selected features which was not unexpected due the fact that other studies in the literature also achieved the same result about the domains. Furthermore, the probability of having MIRs only within the Curated IDRs or the Predicted ones using the MobiDB-lite software with no intersection with other selected features is almost 6%. The MIRs contain many IPMV according to the Clinvar dataset but there is no evidence that indicates the IPMV happens only within the identified MIRs for this study due the fact that the original idea behind the MTR dataset. So if the position is identified with high tolerance to mutation, the algorithm can not classify the position within MIRs.

Most of the IPMV was located inside the MIRs that had intersection with the domain of that IDP but this study indicates that some IPMV also happen within the MIRs that had intersection with only the IDRs which magnify the importance of the IDRs within IDPs.

All of these results have been based on some assumptions and limitations it is intended to be discuss here:

- The dataset has been used to extract the necessary score for finding the MIRs although was more reach from other available datasets from the normal population point of view, that was engaged for MTR calculation, it was based on build version hg19 and that made us using the lift-over tool which leads us to loss in a part of the original dataset.

- The MTR dataset was built for only 85000 transcripts from ensemble version 95 which leads to discarding 180 proteins from the Disprot Human IDPs.
- The Disprot dataset, though gold standard and reliable, contains only 1071 genes that were not ambiguous or obsolete so this study can suggest the use of a more populated set of IDPs to reach more valid results.
- Using The Lift-over tool caused rejection in some positions inside the genes so 30 genes were discarded because they did not preserve their sequence after the lifting over process. So from the 1071 IDPs that were considered for the study, only 861 IDPs met the requirement criteria.
- Considering the MTR score for each position an iid random variable might not be correct but in order to perform the analysis and simplify the problem was taken into consideration so this study suggests the use of a more sophisticated method for this matter if applicable.
- Considering 1.5 standard deviation was driven from trying the 2 and 3 standard deviation which lead to fewer MIRs which can suggest a more appropriate approach of finding this hyperparameter.

At the end it is suggested to recalculate the MTR score using The gnomAD v4.1.0 dataset which contains data from 730,947 exomes and 76,215 whole genomes so that the scores be more precise and also consider more populated IDP dataset to have more reliable results. This suggestion was not performed by this group as the latest version of gnomaAD was introduced publicly at the final day of this team. Developing a more realistic hypothesis for finding the MIRs is highly recommended.

References

[1] M. Choi, U. Scholl, W. Ji, T. Liu, I. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, and et al., “Genetic diagnosis by whole exome capture and massively parallel dna sequencing.” *PNAS*, vol. 106, no. 45, pp. 19 096-–19 101, 2009.

[2] Y. Yang, D. Muzny, J. Reid, M. Bainbridge, A. Willis, P. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, and et al., “Clinical whole-exome sequencing for the diagnosis of mendelian disorders.” *N Engl J Med*, vol. 369, no. 16, pp. 1502–1511, 2013.

[3] H. Beltran, K. Eng, J. Mosquera, A. Sigaras, A. Romanel, H. Rennert, M. Kossai, C. Pauli, B. Faltas, J. Fontugne, and et al, “Whole-exome sequencing of metastatic cancer and biomarkers of treatment response.” *JAMA Oncol*, vol. 1, no. a, pp. 466-–474, 2015.

[4] S. Sherry, M. Ward, M. Kholodov, J. Baker, L. Phan, E. Smigielski, and K. Sirotkin, “dbsnp: the ncbi database of genetic variation.” *Nucleic Acids Research*, vol. 29, no. 1, pp. 308-–311, 2001.

[5] M. Landrum, J. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, and et al., “Clinvar: improving access to variant interpretations and supporting evidence.” *Nucleic Acids Research*, vol. 46, no. 1, p. D1062–D1067, 2017.

[6] P. Stenson, E. Ball, M. Mort, A. Phillips, J. Shiel, N. Thomas, S. Abeysinghe, M. Krawczak, and D. Cooper, “Human gene mutation database (hgmd®): 2003 update.” *HGV*, vol. 21, no. 6, pp. 577–581, 2003.

[7] D. MacArthur, T. Manolio, D. Dimmock, H. Rehm, and et al., “dbsnp: the ncbi database of genetic variation.” *Nature*, vol. 508, no. 1, pp. 469-–476, 2014.

[8] M. Lek, K. Karczewski, E. Minikel, K. Samocha, E. Banks, T. Fennell, A. O’Donnell-Luria, J. Ware, A. Hill, B. Cummings, and et al., “Analysis of protein-coding genetic variation in 60,706 humans.” *Nature*, vol. 536, no. 1, pp. 285-–291, 2016.

[9] G. Project, C., A. Auton, L. Brooks, R. Durbin, E. Garrison, H. Kang, J. Korbel, J. Marchini, S. McCarthy, G. McVean, and et al., "A global reference for human genetic variation." *Nature*, vol. 526, no. 1, pp. 469––476, 2015.

[10] S. Petrovski, Q. Wang, E. Heinzen, A. Allen, and D. Goldstein, "Genic intolerance to functional variation and the interpretation of personal genomes." *PLos Genet*, vol. 9, no. 8, 2013.

[11] K. Samocha, J. Kosmicki, K. Karczewski, A. O'Donnell-Luria, E. Pierce-Hoffman, D. MacArthur, B. Neale, and M. Daly, "Regional missense constraint improves variant deleteriousness prediction." *bioRxiv*, vol. 148353, no. http://dx.doi.org/10.1101/148353, 12 June 2017, preprint: not peer reviewed, 2017.

[12] M. Silk, S. Petrovski, and D. B. Ascher, "Mtr-viewer: identifying regions within genes under purifying selection." *Nucleic Acids Research*, vol. 47, no. 1, 2019.

[13] J. Traynelis, M. Silk, Q. Wang, S. Berkovic, L. Liu, D. Ascher, D. Balding, and S. Petrovski, "Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation." *Genome Res.*, vol. 27, pp. 1715––1729, 2017.

[14] C. Tokheim, R. Bhattacharya, N. Niknafs, D. Gygax, R. Kim, M. Ryan, and R. Masica, D.L. andandKarchin, "Exome-scale discovery of hotspot mutation regions in human cancer using 3d protein structure." *Genome Res.*, vol. 76, pp. 3719––3731, 2016.

[15] E. Medina-Carmona, I. Betancor-Ferna ́ndez, J. Santos, N. Mesa Torres, S. Grottelli, C. Batlle, A. Naganathan, E. Oppici, B. Cellini, S. Ventura, and et al., "Insight into the specificity and severity of pathogenic mechanisms associated with missense mutations through experimental and structural perturbation analyses." *Hum. Mol. Genet.*, vol. 28, pp. 1––15, 2019.

[16] S. Iqbal, E. Pe ́rez Palma, J. Jespersen, P. May, D. Hoksza, H. Heyne, S. Ahmed, Z. Rifat, M. Rahman, K. Lage, and et al., "Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants." *Proc. Natl. Acad. Sci. USA*, vol. 117, pp. 28 201––28 211, 2020.

[17] S. Althari, L. Najmi, A. Bennett, I. Aukrust, J. Rundle, K. Colclough, J. Molnes, A. Kaci, S. Nawaz, T. van der Lugt, and et al., "Unsupervised clustering of missense variants in

hnf1a using multidimensional functional data aids clinical interpretation." *Am. J. Hum. Genet.*, vol. 107, pp. 670––682, 2020.

[18] H. Dietz, J. Saraiva, R. Pyeritz, G. Cutting, and C. Francomano, "Clustering of fibrillin (fbn1) missense mutations in marfan syndrome patients at cysteine residues in egf-like domains." *Proc. Natl. Acad. Sci. USA*, vol. 1, pp. 366––374, 1992.

[19] A. Kamburov, M. Lawrence, P. Polak, I. Leshchiner, K. Lage, T. Golub, E. Lander, , and G. Getz, "Comprehensive assessment of cancer missense mutation clustering in protein structures." *Proc. Natl. Acad. Sci. USA*, vol. 112, pp. E5486––E5495, 2015.

[20] K. Talbot, C. Ponting, A. Theodosiou, N. Rodrigues, R. Surtees, R. Mountford, , and K. Davies, "Missense mutation clustering in the survival motor neuron gene: a role for a conserved tyrosine and glycine rich region of the protein in rna metabolism?" *Hum. Mol. Genet.*, vol. 6, pp. 497––500, 1997.

[21] C. Wang, P. Dixon, S. Decordova, M. Hodges, N. Sebire, S. Ozalp, M. Fallahian, A. Sensi, F. Ashrafi, V. Repiska, and et al., "Identification of 13 novel nlrp7 mutations in 20 families with recurrent hydatidiform mole; missense mutations cluster in the leucine-rich region." *J. Med. Genet.*, vol. 46, pp. 569––575, 2009.

[22] M. Geisheker, G. Heymann, T. Wang, B. Coe, T. Turner, H. Stessman, K. Hoekzema, M. Kvarnung, M. Shaw, K. Friend, and et al., "Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains." *Nat. Neurosci.*, vol. 20, pp. 1043––1051, 2017.

[23] S. Lelieveld, L. Wiel, H. Venselaar, R. Pfundt, G. Vriend, J. Veltman, H. Brunner, L. Vissers, , and C. Gilissen, "Spatial clustering of de novo missense mutations identifies candidate neurodevelopmental disorder-associated genes." *Am. J. Hum. Genet.*, vol. 101, pp. 478––484, 2017.

[24] ——, "Exploration of rare-missense variant clustering in mendelian disease-genes." *PhD thesis, University of Oxford.*, 2020.

[25] M. Buljan, P. Blattmann, R. Aebersold, , and M. Boutros, "Systematic characterization of pan-cancer mutation clusters." *Mol. Syst. Biol.*, vol. 14, p. e7974, 2018.

[26] T. Hayeck, N. Stong, C. Wolock, B. Copeland, S. Kamala karan, D. Goldstein, , and A. Allen, "Improved pathogenic variant localization via a hierarchical model of subregional intolerance." *Am. J. Hum. Genet.*, vol. 104, pp. 299––309, 2019.

[27] E. Pe ́rez Palma, P. May, S. Iqbal, L. Niestroj, J. Du, H. Heyne, J. Castrillon, A. O'Donnell-Luria, P. Nurnberg, A. Palotie, and et al., "Identification of pathogenic variant enriched regions across genes and gene-families." *GenomeRes.*, vol. 30, pp. 62––71, 2020.

[28] L. Wiel, C. Baakman, D. Gilissen, J. Veltman, G. Vriend, and C. Gilissen, "Metadome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains." *Hum. Mutat.*, vol. 40, p. 1030–1038, 2019.

[29] Z. Coban-Akdemir, J. White, X. Song, S. Jhangiani, J. Fatih, T. Gambin, Y. Bayram, I. Chinn, E. Karaca, J. Punetha, and et al. Baylor-Hopkins Center for Mendelian Genomics, "Identifying genes whose mutant transcripts cause dominant disease traits by potential gain-of-function alleles." *Am. J. Hum. Genet.*, vol. 103, p. 171–187, 2018.

[30] M. Aspromonte, M. Nugnes, F. Quaglia, A. Bouharoua, C. DisProt, S. Tosatto, and D. Piovesan, "Disprot in 2024: improving function annotation of intrinsically disordered proteins," *Nucleic Acids Research*, 2023.

[31] N. H. G. R. I. (NHGRI). Deoxyribonucleic acid (dna). [Online]. Available: https://www.genome.gov/

[32] pressbooks. Deoxyribonucleic acid structure (dna). [Online]. Available: https://pressbooks.calstate.edu/explorationsbioanth2/chapter/3/

[33] D. J. Futuyma, "Evolution (4th ed.)," *Sinauer Associates*, 2017.

[34] T. A. Brown, "Genomes (4th ed.)," *Garland Science.*, 2016.

[35] P. J. Keeling and W. F. Doolittle, "Universal horizontal gene transfer and the origin of cyanobacteria." *Nature*, vol. 395, p. 621–626, 1998.

[36] C. Darwin, "On the origin of species by means of natural selection, or the preservation of favored races in the struggle for life." *John Murray*, 1859.

[37] D. N. Cooper, "Human gene mutation," *Bios Scientific Publishers.*, 2000.

[38] alevelbiology. types-of-mutations. [Online]. Available: https://alevelbiology.co.uk/notes/types-of-mutations/

[39] M. J. Landrum, J. M. Lee, M. Benson, D. Robinson, G. Riley, W. Jang, D. R. Maglott, and et al., "Clinvar: public archive of interpretations of clinically relevant variants." *Nucleic Acids Research*, vol. 46, pp. D1065–D1069, 2018.

[40] A. Burian, W. Zhao, T. Lo, and T.-S. D.M., "Genome sequencing guide: An introductory toolbox to whole-genome analysis methods." *Nucleic Acids Research*, vol. 49, p. 815–825, 2021.

[41] ncbi. hg19 vs. hg38. [Online]. Available: https://www.ncbi.nlm.nih.gov/grc/report-an-issue

[42] P. D. Stenson, D. G. Ballinger, M. Mort, A. D. Phillips, V. C. Sheffield, and A. T. Aragaki, "Human gene mutation database (hgmd®): 2017 update." *Human Mutation*, vol. 38, pp. 665–677, 2017.

[43] S. Chen, L. C. Francioli, J. K. Goodrich, R. L. Collins, Q. Wang, J. Alföldi, N. A. Watts, C. Vittal, and et al. Baylor-Hopkins Center for Mendelian Genomics, "A genome-wide mutational constraint map quantified from variation in 76,156 human genomes." *BioRxiv*, 2022.

[44] E. S. Kaitlin, J. A. Kosmicki, K. J. Karczewski, A. H. O'Donnell-Luria, E. Pierce-Hoffman, D. G. MacArthur, B. M. Neale, and M. J. Daly, "Regional missense constraint improves variant deleteriousness prediction," *BioRxiv*, 2017.

[45] F. Dewey, M. Murray, J. Overton, L. Habegger, J. Leader, S. Fetterolf, C. O'Dushlaine, C. Van Hout, J. Staples, and C. e. a. Gonzaga-Jauregui, "Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the discovehr study." *Science*, vol. 354, p. 1549, 2016.

[46] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, and M. e. a. Landray, "Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." *PLoS Med.*, vol. 12, p. e1001779, 2015.

[47] W. McLaren, L. Gil, S. Hunt, H. Riat, G. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The ensembl variant effect predictor." *Genome Biol.*, vol. 17, p. 122, 2016.

[48] B. Yates, B. Braschi, K. Gray, R. Seal, S. Tweedie, and E. Bruford, "Genenames.org: the hgnc and vgnc resources in 2017." *Nucleic Acids Res.*, vol. 45, pp. D619––D625, 2017.

[49] P. Romero, M. Oñate, and F. J. Moreno, "Intrinsically disordered proteins in plant pathology." *Plant Physiology*, vol. 152, pp. 479–489, 2010.

[50] K. S. Dunker, C. J. Lawson, P. Brown, R. M. Williams, P. Romero, C. M. Sykes, and �. Obradović, "Intrinsically disordered protein regions in lcds and other functionally significant sequences." *Journal of molecular biology*, vol. 315, pp. 657–682, 2001.

[51] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins and their roles in maladies." *The Journal of biological chemistry*, vol. 290, pp. 1–11, 2015.

[52] P. Nair, U. Kühlmann, and R. Jahn, "Snap receptors: Snare-protein receptors that orchestrate membrane fusion." *Cell and Tissue Research*, vol. 317, pp. 21–31, 2004.

[53] C. Niranjan, S. Mohan, and A. Prakash, "Comprehensive analysis of short linear motifs in human proteome and their interactions with ptms." *BMC Genomics*, vol. 7, pp. 1–11, 2006.

[54] D. Piovesan, A. Del Conte, D. Clementel, A. Monzon, M. Bevilacqua, M. Aspromonte, J. Iserte, F. Orti, C. Marino-Buslje, and S. Tosatto, "Mobidb: 10 years of intrinsically disordered proteins." *Nucleic Acids Res.*, vol. 51, pp. D438–D444, 2023.

[55] Silk, M. and Petrovski, S. and Ascher,D.B. Mtr-viewer. [Online]. Available: https://biosig.lab.uq.edu.au/mtr-viewer/downloads

[56] gatk. Picard version 4.0.5.2. [Online]. Available: https://gatk.broadinstitute.org/hc/en-us/articles/360036831351-LiftoverVcf-Picard

[57] Uniprot. id-mapping. [Online]. Available: https://www.uniprot.org/id-mapping

[58] biopython. Bio.blast.ncbixml module. [Online]. Available: https://github.com/biopython/biopython/blob/master/Bio/Blast/NCBIXML.py

[59] MobiDB. Mobidb rest api. [Online]. Available: https://mobidb.org/api/download

Acknowledgments

I would like to express my sincere gratitude to all those who have supported me throughout this project.

First and foremost, I am deeply indebted to my supervisor, Professor Emanuela Leonardi. Her invaluable guidance and mentorship were instrumental in shaping this project from its inception to completion. Professor Leonardi's willingness to provide step-by-step guidance was crucial in navigating the research process and overcoming challenges.

I am also grateful to Ivan Micetic for his dedication to teaching me about various programming languages. His expertise significantly enhanced my technical skills and allowed me to tackle the programming aspects of this project with confidence.

Finally, I would like to extend my thanks to all my colleagues at the BioComputing UP Laboratory. Their collaborative spirit, helpfulness, and stimulating discussions fostered a supportive environment that greatly facilitated my progress. Completing this project would not have been possible without their camaraderie and shared knowledge.

Thank you all for your invaluable contributions.